MACHINE LEARNING-ADAPTED RAPID VISUAL SCREENING METHOD FOR PRIORITIZING SEISMIC RISK STATES OF MASONRY STRUCTURES

YIĞMA YAPILARIN SİSMİK RİSK DURUMLARININ ÖNCELİKLENDİRİLMESİNE YÖNELİK MAKİNE ÖĞRENMESİNE ADAPTE EDİLMİŞ HIZLI GÖRSEL TARAMA YÖNTEMİ

Onur COŞKUN

Assoc. Prof. Alper ALDEMİR

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfilment to the Requirements

for be Award of the Degree of Doctor of Philosophy

in Civil Engineering

2023

ABSTRACT

MACHINE LEARNING-ADAPTED RAPID VISUAL SCREENING METHOD FOR PRIORITIZING SEISMIC RISK STATES OF MASONRY STRUCTURES

Onur COŞKUN

Doctor of Philosophy, Department of Civil Engineering

Supervisor: Assoc. Prof. Dr. Alper ALDEMİR

November 2023, 178 pages

The majority of earthquake-related losses are associated with fully collapsed buildings. So, the determination of the seismic risk of buildings is essential for building occupants located in active earthquake zones. Unfortunately, the existing techniques employed to assess the risk status of extensive building inventories lack the requisite speed and precision for dependable decision-making. Furthermore, post-catastrophe risk categorizations of structures heavily rely on the expertise of engineering teams. Consequently, the decision on risk distributions of building stocks before and after hazards requires more sustainable and precise methodologies that include other means of technological advancement. In this study, using a database consisting of 4,356 masonry buildings provided by the Ministry of Environment, Urbanization and Climate Change (general directorate of geographical information systems), Engineering Firms and Gazi University, the building properties were determined, and detailed static analyzes were made. Then, for the first time in the literature, a new, fast and accurate seismic evaluation method has been proposed, which is associated with detailed evaluation results of structures with the help of machine learning algorithms. Within the scope of the study, the data set was subjected to data preprocessing methods (Synthetic Minority Oversampling Technique (SMOTE), Backward Feature Elimination and Forward Feature Selection, Feature Importance, and Feature Correlation methods). First, fifteen parameters obtained from detailed seismic analysis results, building drawings and building photographs were selected by applying data preprocessing and reduced to six parameters with the highest success impact. To achieve this, size reduction methods were used and considering some selected parameters from the street walking. In addition, the minority data classes were reproduced synthetically with the Synthetic Minority Oversampling Technique Method (SMOTE) during the training phase, and the success rate for test data was maximized. In this study, nine machine learning algorithms, namely; Logistic Regression, Decision Tree, Random Forest, Multivariate Adaptive Regression Spline, Support Vector Machine, K-Nearest Neighbor, Gradient Boosting Algorithm, Extreme Gradient Boosting Algorithm, LightGBM Algorithm and where all these algorithms work together with Voting Classifier Method are used. The risk layers of the buildings were estimated by creating risk classes according to the ratio of the floor shear force of the risky walls to the total floor shear force (V_e/V_r = RVS) or the damage detection level.. At the end of the study, this vulnerability assessment method that creates the risk layers of existing buildings in the literature and can determine the most dangerous or non-risk buildings class has been proposed. This is important for deciding the starting point of urban transformation and assessing the seismic vulnerability of buildings in different regions. As a result of the analysis of the algorithms in the study, the correct prediction rates of the three-tier risk class (RVS values) for the learning database (i.e., 3,484 buildings) and the test database (i.e., 872 buildings) of the proposed method were determined as approximately 99.19% and 86.58%, respectively. High success rates were also obtained in the estimation of RVS values with two and four layers. The parameter selections of the proposed method in the study were determined in a way that can be obtained from the photographs of the buildings with the Convolutional Neural Network structures. Therefore, without the need for technical personnel and without entering the building, with the automation methods of the structures, after the parameter selection, the estimations of the RVS values using machine learning methods can be made with high accuracy. This process is employed to identify, catalog, and prioritize the buildings at highest risk of sustaining damage in designated regions during an upcoming earthquake.. For this reason, this method is of great importance in order to determine and strengthen Turkey's weak structures and minimize the loss of life and property.

Keywords: Seismic risk estimations, masonry structures, machine learning, seismic risk classification, Deep Learning, Pre-Trained Convolutional Neural Networks, dimension Reduction, SMOTE

ÖZET

YIĞMA YAPILARIN SİSMİK RİSK DURUMLARININ ÖNCELİKLENDİRİLMESİNE YÖNELİK MAKİNE ÖĞRENMESİNE ADAPTE EDİLMİŞ HIZLI GÖRSEL TARAMA YÖNTEMİ

Onur COŞKUN

Doktora, İnşaat Mühendisliği Bölümü

Tez Danışmanı: Supervisor: Doç. Dr. Alper ALDEMİR

Kasım 2023, 177 sayfa

Depremlerden kaynaklanan kayıpların çoğu, tamamen çökmüş binalarla ilişkilidir. Bu nedenle, aktif deprem bölgelerinde bulunan bina kullanıcıları için binaların sismik riskinin belirlenmesi büyük önem taşımaktadır. Ne yazık ki, büyük bina stoklarının risk durumunu tahmin etmek için kullanılan mevcut yöntemler, güvenilir, hızlı ve doğru karar vermek için yetersizdir. Ayrıca binaların büyük doğal afetler sonrası risk sınıflandırmaları tamamen mühendislerden oluşan teknik ekibin tecrübesine bağlıdır. Bu nedenle, bina stoklarının tehlikelerden önceki ve sonraki risk dağılımlarına ilişkin karar, diğer teknolojik ilerleme araçlarını içeren daha sürdürülebilir ve doğru yöntemler gerektirir. Bu çalışmada Çevre ve Şehircilik Bakanlığı, Mühendislik Firmaları ve Gazi Üniversitesi tarafından sağlanan 4356 adet yığma binadan oluşan veri tabanı kullanılarak bina özellikleri belirlenmiş ve detaylı statik analizleri yapılmıştır. Daha sonra literatürde ilk kez, yapıların ayrıntılı değerlendirme sonuçlarının makine öğrenmesi algoritmaları yardımıyla ilişkilendirildiği yeni, hızlı ve doğru bir sismik değerlendirme yöntemi önerilmiştir. Çalışma kapsamında veri seti veri ön işleme yöntemlerine (Sentetik Azınlık Yüksek Örnekleme Tekniği (SMOTE), Geriye Özellik Eleme ve İleriye Özellik Seçimi, Özellik Önemi ve Özellik Korelasyon yöntemleri) tabi tutulmuştur. Öncelikle ayrıntılı sismik analiz sonuçları, bina çizimleri ve bina fotoğraflarından elde edilen on beş parametre veri ön işleme uygulanarak seçilmiş ve sonunda başarı etkisi en yüksek altı parametreye indirgenmiştir. Bunu başarmak için boyut küçültme yöntemleri kullanılmış ve sokaktan elde edilebilecek verilerin olması göz önünde bulundurulmuştur. Ayrıca eğitim aşamasında Sentetik Azınlık Yüksek

Örnekleme Tekniği Yöntemi (SMOTE) ile azınlık veri sınıfları sentetik olarak yeniden üretilmiş ve test verilerinin başarı oranı en üst düzeye çıkarılmıştır. Bu çalışmada dokuz makine öğrenmesi algoritması; Lojistik Regresyon, Karar Ağacı, Rastgele Orman, Çok Değişkenli Uyarlanabilir Regresyon Spline, Destek Vektör Makinesi, K-En Yakın Komşu, Gradyan Artırma, Aşırı Gradyan Artırma, Hafif Gradyan Artırma Algoritmaları ve tüm bu algoritmaların birlikte çalıştığı Oylama Sınıflandırma Yöntemi kullanıldı. Çalışmada riskli duvarların kat kesme kuvvetinin toplam kat kesme kuvvetine (V_e / V_r =RVS) oranına veya hasar tespit seviyesine göre risk sınıfları oluşturularak binaların risk katmanları tahmin edilmiştir. Çalışma sonunda literatürde mevcut binaların risk katmanlarını oluşturan ve en tehlikeli veya risksiz bina sınıfını tespit edebilen bir yöntem önerilmiştir. Bu, kentsel dönüşümün başlangıç noktasının belirlenmesi açısından önemlidir. Çalışmadaki algoritmaların analizi sonucunda, önerilen yöntemin öğrenme veri tabanı (yani 3.484 bina) ve test veri tabanı (yani 872 bina) için üç katmanlı risk sınıfının RVS değerlerini doğru tahmin oranları sırasıyla yaklaşık %99,19 ve %86,58 olarak belirlenmiştir. İki ve dört katmanlı RVS değerlerinin tahmininde de yüksek başarı oranları elde edilmiştir. Çalışmada önerilen yöntemin parametre seçimleri Evrişimli Sinir Ağı yapıları ile binaların fotoğraflarından elde edilebilecek yöntemlere uygun olması da sağlanmıştır. Sonuç olarak, tez kapsamında en çok başarı yüzdesi elde edilen birleşik öğrenme ve tahmin etme yöntemi, teknik personele ihtiyaç duymadan ve binaya girmeden, yapıların otomasyon yöntemleri ile entegre olabilen, sokaktan yapılacak parametre seçimine uygun olarak RVS değerlerinin tahminleri yüksek doğrulukta yapılabilmektedir. Prosedür, belirli bir bölgede olabilecek bir deprem sırasında hasar görebilecek en savunmasız binaları tespit etmek, envanterini çıkarmak ve risk sıralamasını yapmak için uygulanır. Türkiye'nin zayıf yapılarının tespiti ve güçlendirilmesi bu nedenle can ve mal kaybının en aza indirilmesi için bu yöntem büyük önem taşımaktadır.

Anahtar Kelimeler: Sismik risk tahminleri, yığma yapılar, makine öğrenimi, sismik risk sınıflandırması, Derin Öğrenme, Ön Eğitimli Evrişimli Sinir Ağları

ACKNOWLEDGEMENT

Foremost, I would like to express my deep gratitude to my supervisor Doc. Dr. Alper ALDEMİR for supporting and guiding me with his vast knowledge and experience throughout my thesis study. It was a great honour and pleasure to work with him.

Mustafa ŞAHMARAN, the creator of my master's thesis, In my thesis jury, Prof. Dr. Hakan ERDOGAN, Prof. Dr. Murat Altuğ ERBERİK , Assoc. Prof. Dr. Burcu GÜLDÜR ERKAL

I would also like to thank the project numbered 221M186 within the scope of the 1001 program of the Scientific and Technical Research Council of Turkey (TÜBİTAK) for the support it has provided, part of which is my thesis work.

Finally, I would like to thank my mother, my father, my brother İlker COŞKUN, Ali Erhan YILMAZ, Dr. İsmail Raci Bayer and my friend Yunus Emin BACIK for their endless support, encouragement, and precious love. Also, I am eternally grateful to my dear wife Gizem and my daughter Elif Duru, who supported and motivated me throughout this process.

The financial support from the Scientific and Technical Research Council of Turkey (TUBITAK – 221M186) is gratefully acknowledged.

Onur COŞKUN

November 2023, Ankara

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SYMBOLS AND ABBREVIATIONS

Abbreviations

BA	Building Age
BM	Building Mass
EZ	Earth Zone
CR	Compressive Strength
CM	Confusion Matrix
DT	Decision Tree
FA	First Floor Area
GS	Grid Search
GB	Gradient Boosting
KNN	K-Nearest Neighbor
LightGBM	Light Gradient Boosting
LR	Logistic Regression
MARS	Multivariate Adaptive Regression Spline
ML	Machine Learning
NF	Number of Floors
RF	Random Forest
RVS	The ratio of floor shear force of the risky walls to the total floor shear force
RS	Classified ratio of the floor shear force on the risk walls to the total shear force
SC	The short period (0.2s) is the design spectral acceleration coefficient
SD	Design spectral acceleration coefficient of 1 s
SVM	Support Vector Machine
SMOTE	The Synthetic Minority Oversampling Technique Method

ADASYN	Adaptive synthetic sampling
SS	Shear Strength
SW	Specific Weight
TS	Diagonal Tensile Strength
TN	True Negatives
TP	True Positives
URM	Unreinforced Masonry
VQ	Visual Crack and Mortar Quality
XGB	Extreme Gradient Boosting

1. INTRODUCTION

Developing nations encounter various challenges related to urban planning and the quality of construction. This issue is mainly due to the lack of expertise, legislation and funds in developing countries. Also, the increase in population generally results in unplanned urbanization. The bad aspects of unplanned urbanization can only be seen by society when it becomes a problem for living. In other words, The haphazard development witnessed in many developing regions is unsustainable when it comes to effective city planning and urbanization. The other aspect of this issue is that the low quality of construction causes loss of lives during natural hazards, including earthquakes, tsunamis, etc. Therefore, these countries are obliged to ameliorate their previous mistakes by implementing some precautions or re-urbanization methods. To this end, the non-vulnerable buildings should be filtered from the whole building stocks. This process requires a different approach. In the literature, there existed some attempts to determine the seismic vulnerability of individual buildings. But studies show that the results of these methods are far from reality when compared with detailed analyzes (Coskun et al. 2020).

One of the most critical problems in our country is the earthquake risk. Our nation is located in the Mediterranean-Alpine-Himalayan seismic belt, which is one of the most active and seismically active areas in the world. According to the studies carried out as of 2010, it has been reported that one-fifth of the total earthquakes in the world occurred in the Mediterranean-Alpine-Himalayan seismic belt. Approximately 93% of our country's lands and approximately 98% of its population are in danger of earthquakes. This fact has been proven many times in recent years, both by the seismic hazard studies conducted in the literature on our country and by earthquakes of magnitude below and above that we encounter in our daily lives. As obtained from the annual reports published by the Kandilli Observatory, the number of earthquakes with a magnitude of five or more that have occurred in our country in the last twenty years is 73. Although it is challenging to determine the material and moral losses of our country during these earthquakes, according to the report published by the General Directorate of Disaster Affairs, Earthquake Research Department, Seismology Branch Directorate for the years 1900-2009, a total of 554,365 buildings were destroyed or severely damaged and out of use. And approximately 92,463 of our citizens lost their lives. The Disaster and Emergency Management Presidency Earthquake

Department's Van, İzmir, and Elazığ Earthquakes Report states that in the October 23, 2011, Mw = 7.2 Van Earthquake, 48666 buildings were destroyed or severely damaged, 604 people died, and 1301 people were injured. In the aftermath of the earthquake, significant damage was observed not only in structures designed in adherence to earthquake regulations predating 2007 but also in buildings designed according to the 2007 Turkish Earthquake Code. In the report, 117 dead, 1,034 injured and 15,000 homeless were recorded for Turkey, while in Greece; it was stated that there were 2 dead and 19 injured people. In addition, the number of destroyed buildings was determined as 71, and the number of buildings to be demolished urgently and heavily damaged was 653. In the Elazığ earthquake, with a magnitude of M_w = 6.8, 42 people lost their lives, 137 people were injured, and more than 4,000 structures were severely damaged. According to the published report, only after the Elazığ earthquake the money that came out of the state's coffers exceeded 50 million dollars. The most destructive earthquake experienced recently is the 2023 Kahramanmaraş earthquakes. On February 6, 2023, two earthquakes with magnitudes of 7.7 Mw and 7.6 Mw struck, occurring nine hours apart. These seismic events had their epicenters in the Pazarcık and Ekinözü districts of Kahramanmaraş. Official statistics indicate that the combined impact of these earthquakes resulted in significant casualties. In Turkey, a minimum of 50,783 individuals lost their lives, while in Syria, official records indicate that at least 8,476 people perished. Furthermore, the total number of injuries exceeded 122,000 people. . According to the 2023 Parliamentary Earthquake Investigation Commission Report, the total cost of the earthquake was 148.8 billion dollars.

In order to reduce the destructive effects of these earthquakes, many studies have been carried out to identify existing buildings in Turkey. Law No. 6306 was published on 16/05/2012 for the identification of buildings in Turkey that are at risk of severe damage or collapse under the risk of earthquakes and the demolition and construction of new ones. When the initial publication of the law occurred, the approach used to assess the earthquake susceptibility of pre-existing structures relied upon the recommendations detailed in the 7th Chapter of the Regulations for Constructing Buildings in Seismic Areas. This chapter, titled 'Assessment and Reinforcement of Existing Buildings,' was introduced in 2007." Subsequently, the damage level of the existing structures was tried to be estimated with the calculation methods titled Principles Regarding the Detection of Risky Buildings in the annex of the Implementation Regulation of the Law No. 6306 published in 2013 and 2019.

In this context, a method was published in the annex of the Implementing Regulation of the same Law, which calculates the rapid risky structure detection of existing structures. Although technical experts are required when using this method, the accuracy rate is very low compared to the detailed analysis of the structure (Coskun, 2019).

When the activity reports of the Ministry of Environment, Urbanization and Climate Change are examined, the number of buildings demolished by the end of 2022 within the scope of this law is 242212 buildings. About 80% of these have been replaced with safer and more robust structures. However, considering the millions of risky buildings in Turkey, the conversion of such a small number of buildings shows that more destruction and loss of life await us in the upcoming disasters. To address and resolve this issue, it is essential that the buildings in Turkey are prepared for possible earthquakes, starting with the structures that require urgent intervention, by determining the priority order of the risky ones quickly and effectively before being exposed to an earthquake. But, It's a difficult task to continuously examine and monitor a building's seismic safety and vulnerability., Especially when extensive site assessments are required during the methods used with the first detailed data entry. The methodologies designed to determine the seismic risk by examining individual buildings in detail were quite mature, but these include material testing, plan drawings, material quality, etc. The main reason for this is; inputs and procedures require significant time, manpower, and computational power of experienced engineers. This required the use of sophisticated analysis tools with detailed inputs, such as these detailed seismic assessment techniques becoming useless if the number of buildings in question reaches thousands or more. In other words, These methodologies are not suitable for assessing the seismic risk of a substantial inventory of buildings (Coskun and Aldemir, 2022).

Also, most losses from earthquakes are associated with the local or total collapse of buildings. Especially examining structural cracks or determining the wall type in a masonry structure damaged by an earthquake is a very dangerous task for civil engineers. To overcome these difficulties, with the advancement of technology, it has become essential to determine the seismic risk of buildings by observations that can be made from outside the building. (Coskun et al., 2020) This motivates the search for technological solutions for the safety assessment of existing structures. Machine learning and artificial intelligence studies are at the forefront of these technological solutions. With the advancement of machine learning, studies to assess the seismic risk of buildings, especially from the outside, have started to increase recently. Yet, none of these studies have put forth a versatile seismic risk assessment method for masonry structures that integrates machine learning algorithms.

However, Masonry buildings make up a substantial portion of the urban infrastructure, constituting the majority of the existing building stock across European territories. While more recent structures often utilize reinforced concrete or steel frameworks, the historical and established urban landscape predominantly consists of masonry buildings.In this context, it is impossible to convert so many masonry stocks through detailed analysis, both in terms of technical staff and cost. For the stated reasons, current studies have focused on rapid seismic risk assessment methods. This is the key to salvation for fast and cheap transformation, especially before the Istanbul earthquake, which is expected to happen in the near future(Cardinali et al., 2022).

Nonetheless, none of the studies within the existing literature have introduced a flexible seismic risk assessment approach for masonry structures that incorporates machine learning algorithms. Therefore, this thesis focused on developing a simple, rapid visual scanning method to estimate the damage level of masonry buildings using machine learning algorithms. It is aimed that the parameters required for the Procedure used in the proposed method can be collected quickly and safely without the need for technical personnel to enter the building.

2. LITERATURE REVIEW

2.1. Introduction

Over the past few years, there has been a notable increase in the attention given to evaluating seismic risk for masonry buildings situated in areas susceptible to earthquakes. The way all kinds of structures are evaluated has changed due to increased computing capacity, which makes it possible to represent more degrees of freedom.At the same time, the utilization of artificial intelligence in the construction discipline allowed the calculation methods to be made faster. It is unrealistic to try to make a comprehensive review of all the methods applied in this section. Instead, studies on rapid risky structure detection of structures, their success rates and their contributions to the literature will be mentioned. This section will be divided into two main sections. Methods in rapid building safety assessments using artificial intelligence in risk analysis and rapid building safety assessments made with more detailed or statistical approaches will be reviewed. While these methods differ significantly in their assumptions and complexity, the question is "Which method is better?" should not be. Instead, realizing that all these methods are suitable for specific applications, their contributions to the literature will be mentioned.

Vulnerability is defined by taking into account other factors such as the material of the structure, ground effect, condition of the buildings, construction quality, irregularity in the structure's design, earthquake resistant design level (Jeddah, 2016). In general, the studies were carried out on the basis of these features of the structures.

The field of machine learning (ML), which includes algorithmic methods for learning from data, is one of the most important areas of science today. Intelligent systems that are driven by data have the capacity to transform human knowledge and experience into timely and well-informed decision-making. When attempting to forecast the future accurately, they mix concepts from the disciplines of statistics and probabilities with mathematical optimization approaches. Modern scientific research is leading the way in building seismic assessment. A number of scholars have presented forth techniques for calculating the damage response of structures subjected to seismic vibrations without extensive analysis. Modern fast computing techniques based on the usage of ML have been developed as a result of the increase in computing capacity. However, a lack of experience with using complex ML architectures may have an impact on the AI model's performance, which ultimately lowers the algorithm's generalization and reliability, which should be characteristics of these systems. The most recent studies provide an interpretable, completely validated ML technique for forecasting building damage from earthquakes (Demertzis et al., 2023).

In the literature, studies have also been carried out with methods that can be detected by walking on the street in rapid scanning. Rapid assessment methods started with (Federal Emergency Management Agency, 1998) in the US and (Standard for Seismic Evaluation of Existing Reinforced Concrete Buildings, 2001)in Japan in the early 2000s. At the beginning of the first studies on this subject, (Hassan and Sozen, 1997) used the data from 46 buildings damaged in the earthquake in Erzincan in 1992. There are different studies proposed by many researchers, especially in Turkey, for the rapid assessment of buildings before earthquakes.

There are also fast and automated methods in the literature that perform a quick scan of data from the building's photographs. The studies cited in the literature, such as those conducted by (C. Wang et al., 2021, Cooner et al., 2016, Li et al., 2018, Z. Xu et al., 2018,Sublime and Kalinicheva, 2019, Kerle et al., 2019, Stepinac and Gašparović, 2020, etc.) have achieved notable success in generating damage maps for post-earthquake damage assessment by employing unmanned aerial vehicles. In an attempt to develop new techniques that would shorten the time it takes to process data, these studies have also evolved to include post-earthquake permanent displacement calculations. (LI et al., 2011, X. Wang et al., 2020), as well as research that categorizes post-earthquake damages. Last but not least, research is done using images taken by unmanned aerial vehicles to identify the physical characteristics of buildings and their structural flaws, such as soft floors (Yu et al., 2020). However, it's important to note that none of these studies were originally designed for proactive use before disasters to enhance preparedness efforts. Furthermore, ML techniques have been advanced for the prediction of structural systems based on image data, as exemplified by the research conducted by (Geiß et al., 2015). in their 2015 research. Their research revealed that machine learning techniques, particularly the random forest and support vector machine algorithms, have the capability to forecast various structural systems.

(e.g., masonry, confined masonry, reinforced concrete frame, steel frame) with a high degree of accuracy. Nevertheless, it's worth emphasizing that none of these studies were explicitly developed for proactive, pre-disaster applications aimed at improving preparedness. The use of random forest and support vector machine algorithms constitutes a pivotal component of their methodologies.

The process of assessing a building's potential for earthquake damage is known as the evaluation of seismic vulnerability in masonry buildings. This evaluation usually consists of a visual survey of the building as well as an examination of several structural features, such as the kind and caliber of the walls, the mortar's strength, and the presence of reinforcement. In addition to these data, a number of strategies and procedures have been established to assess the seismic vulnerability of masonry constructions. Some of these methods include:

- Empirical methods: These methods rely on statistical data and historical earthquake damage records to estimate a building's potential for damage. Penalty points are assigned based on the significance of various parameters, as determined from statistical data and informed by past performance assessments of similar buildings during seismic events.
- Analytical methods: These techniques employ mathematical models to replicate a building's response during an earthquake and predict potential damage. They encompass a range of methodologies, including finite element analysis, impulse analysis, and response spectrum analysis. These methods occupy a middle ground between swift structural assessment and comprehensive, in-depth evaluation.
- Artificial intelligence integrated methods: These methods are methods that are performed by teaching the data of structural damages that occurred as a result of old earthquake data or existing detailed analysis of buildings using machine learning algorithms or artificial neural networks to these algorithms. (C. Wang et al. 2021, -Cooner et al. 2016, Li et al. 2018, Xu et al., 2018, Sublime and Kalinicheva 2019, Kerle et al. 2019, Stepinac and Gašparović 2020, Coskun and Aldemir, 2022, etc.)
- Detailed methods based on assumptions: These are capacity-based quick assessment methods that evaluate the building capacity based on the exterior visual of the building, as well as the location of the structural elements on the critical floor

(considered mostly as the ground floor), dimensions and types of the structural elements. Methods such as P25 (2005), Yakut (2005), DURTES (2003) and PERA (2014) are capacity-based second-stage methods in which both facade visual and plan information are used. These methods are not very suitable for Turkey's building stock due to both cost and technical staff demand. (Ekici, 2022)

- Visual rapid scanning methods: These are rapid assessment methods that generally use the exterior image of the building (street photo or street view) and other seismicity parameters of the region. (FEMA- 154), Sucuoğlu et al. 2007, RYTEİE-2019 ,(Aldemir et al., 2020, Coskun et al., 2020)

Studies on this subject have emphasized the importance of considering both structural and non-structural components of masonry structures, also factors such as material and construction quality, age of the building, wall type, floor height, and presence of reinforcement. Reinforcement Many studies have taken into account the seismicity of the area where each structures is located, as well as the use of multiple methods and techniques in the assessment process.

The swift assessment of seismic vulnerability in reinforced concrete and masonry buildings is a highly significant and widely discussed subject in Turkey. In addition, artificial intelligence methods/smart software based on field studies and analytical data have also been used in recent years regarding rapid pre-earthquake assessments. In the present day, sophisticated software solutions such as artificial intelligence, genetic algorithms, and fuzzy logic play a pivotal role in addressing intricate engineering challenges. These software tools are highly favored due to their ability to deliver remarkably accurate predictions and offer substantial benefits in terms of cost-effectiveness and time efficiency. (Ekici, 2022). The use AI started in the 1950s, thanks to the development of computer science, Machine Learning in the 1980s, Deep Learning in the 2010s, and especially in the last five years. This has led to the very popular use of Convolutional Neural Networks (CNN) methods. For the stated reason, rapid scanning methods based on visuals are also divided into two those made with the help of artificial intelligence and those made with statistical methods (Beyhan 2023).

2.2. With No Artificial Intelligence

2.2.1. Detailed rapid assessment

The methods in this class are evaluation methods made after the capacity or detailed survey calculations are made. The methods in this class fall between a quick assessment and a detailed assessment. Although these methods cannot be used in the damage assessment of the building, they are the methods made by the capacity calculation of the undamaged state of the building or calculated by the dimensions of the columns and beams of its survey.

This process demands a different approach compared to previous practices. Existing literature has seen some efforts to evaluate the seismic vulnerability of individual structures.. Nonetheless, these approaches usually require the utilization of intricate analysis tools and the meticulous collection of extensive geometric and material data from the building under scrutiny. This data acquisition procedure necessitates a substantial investment of time and expertise to thoroughly assess the building's materials and structural characteristics. Unfortunately, due to constraints such as limited time, inadequate funds, and a shortage of personnel, These intricate procedures are ill-suited for concurrently evaluating a large quantity of buildings.

The study proposed by (Johnson and Fick, 2018) examined a comprehensive dataset representing 752 buildings from seven different earthquakes, overall serious damage trends, and its correlation with the calculated priority index. In the study, recent building inspections Following earthquakes, a substantial dataset containing information about damage levels and building performance was generated. A high correlation was found between the priority index method proposed in the study and the damage detection rates. The priority index proposed in the study consists of the wall index and the column index. In the study, it is mentioned that this index can be used as a common screening criterion for low-rise and mid-rise buildings. These two indices represent the weighted ratio of the cross-sectional areas of the columns and walls to the total floor area above the building's foundation.In the study, the Priority index was calibrated using the data obtained from 49 buildings examined after the 1992 Erzincan Earthquake. In addition, after the 1999 Düzce, Turkey, 2008 Wenchuan, China and 2010 Haiti earthquakes, the results were evaluated in tabular form using similar studies. At the end of the study, it was determined that the severe damage trends after the Bingöl, Turkey and Pisco Peru earthquakes did not follow the increasing vulnerability trend for the smaller priority indices observed after the other earthquakes. In addition, a decreasing trend was observed in the priority index, which decreases as the number of floors increases in the data set, indicating that the risk of serious damage is higher in buildings with three floors and above.

Yücemen et al. (2004) implement a method called discriminant analysis to analyze earthquake damage data collected from the 12th November 1999 Düzce earthquake. The study was performed using Fisher's linear discriminant function and the SPSS statistical analysis program. Initially, six prediction variables were identified, but this was later reduced to 3. In the study, the damage status was classified into either two or three categories, and the success rates were compared. The six main forecast variables are as follows:

- Number of Floors (N): The total count of floor systems located above ground.
- Soft Story Index (SSI): This measure is characterized as the proportion of the ground floor's height in relation to that of the first floor.Overhang Ratio (OHR): Overhang area is defined as the floor area beyond the outermost frame lines on all sides in a typical floor plan. The overhang ratio is defined as the sum of the overhang areas on each floor divided by the area of the ground floor.
- Normalized Redundancy Score (NRS): A reinforced concrete frame building's Normalized Redundancy Ratio is computed.
- Minimum Normalized Lateral Stiffness Index (MNLSTFI): Calculating this index involves examining the columns and structural walls situated on the ground floor. Columns are defined as reinforced concrete elements with a "maximum cross-sectional dimension/minimum cross-sectional dimension ratio" below seven, whereas all other reinforced concrete elements are categorized as structural walls.
- Minimum Normalized Lateral Strength Index (MNLSI): This index is established by considering the columns, structural walls, and partition walls (typically constructed using clay bricks) on the ground floor. To calculate it, reinforced concrete elements with a "maximum cross-sectional dimension/minimum cross-sectional dimension ratio" of less than seven are designated as columns, whereas all other reinforced concrete elements are classified as structural walls..

The study found that the number of floors was the most effective prediction parameter, while the yield rate, minimum normalized lateral stiffness and minimum normalized lateral strength indices were less effective and could be statistically excluded from the discriminant analysis. However, from an engineering perspective, the study suggests that these other parameters should be included in damage estimation evaluations. When the damage class was estimated with a two-class statistical model, approximately 69% of the damaged buildings were correctly classified, with a correct classification rate of around 71% for heavily damaged or collapsed buildings. The study concludes that these correct classification rates were similar for the three- and six-parameter estimates. When the data was divided into three damage classes, the success rate dropped to 54%, with a success rate of 58% for heavily damaged and collapsed buildings. The damage from the 1992 Erzincan, 1999 Bolu, Düzce, and Kaynasli earthquakes is given in Table 2.1.

Table 2.1. Correct classification rates of damage states for different earthquakes based on the proposed statistical model

Damage database	**Number of buildings**	**Correct classification rates (%)**			
		Two-damage states		Three-damage states	
		Six parameters	Three parameters	Six parameters	Three parameters
1992-Erzincan	43	95.3	88.4	65.1	62.7
1999-Bolu,Düzce, Kaynasli	152	81.6	82.2	66.4	67.1

"A method developed by (Yakut, 2004) for rapid seismic safety assessment of reinforced concrete buildings. In order to implement this method, which was developed considering the building conditions in Turkey, the dimensions of the building support system on the ground floor and the results of the bearing pressure test must be known. Firstly, by assuming that there is no lateral load, the shear resistance of each column and partition element is calculated. By summing the values calculated for each element on the ground floor, the total base shear resistance (V_c) of the building is obtained. The method also calculates the building resistance value (V_{yw}), which includes the contribution of the fill walls to the horizontal load-carrying capacity. The method defines the building yield resistance from this value. To determine the safety degree of the building, the Base Capacity Index (BCPI), which can be called the capacity-effect ratio, is calculated. Although the method has been successful, the process of taking cores and conducting surveys can be time-consuming and costly."

The FEMA 310 method, created by the Federal Emergency Management Agency (FEMA) in 1998, is designed for evaluating the seismic performance of pre-existing structures.The method is used to evaluate buildings classified according to various construction materials and load-bearing systems, such as reinforced concrete, masonry, and wood. The assessment of a building's seismic performance is based on non-structural and foundation-ground characteristics. This approach comprises three tiers of assessment within each seismic zone. Throughout these evaluation phases, buildings are appraised according to their performance in terms of life safety or immediate occupancy. The seismic behavior of the building before an earthquake is decided based on a control list. In the first stage of the control list, the structural safety level of the building is determined based on the load-bearing system and ground parameters. The process is detailed and time-consuming, particularly in the calculation phase, where values such as the base shear force, floor shear force, floor displacement values, and natural vibration period are calculated. Equations and tables in accordance with the building type are available for use in the calculation of these values. Comparing the obtained values with limit values yields the building's safety level. During the second stage of the evaluation, the unfavorable aspects of the building are juxtaposed with the obtained results. Obtained through the method for linear static, linear dynamic, special and non-structural elements. In the third stage of evaluation, a detailed analysis is required for buildings that exceed a height of 30.5 m, have lateral irregularity, have vertical irregularity, or have a plan dimension ratio greater than 1.4.

The "Standard for Seismic Evaluation of Existing Reinforced Concrete Buildings" is a swift assessment technique utilized in Japan to appraise the vulnerability of pre-existing structures to seismic damage and to improve their earthquake resilience. This method relies on the examination and structural analysis of buildings at the actual site. It is intended for the use in buildings with a maximum of six floors. Before the method is applied, the type of structural system , the construction year , and the plan dimensions of the building should be determined. The method is not recommended for buildings with unusual structural systems, very poor material quality, more than 30 years of history, or fire risk. The results obtained using this method allow for an assessment of the probable building's performance in response to seismic activity during an earthquake. The method consists of three different stages that provide more realistic results and require more detailed inspections and calculations. The initial stage of the evaluation entails an assessment of the building's structural system, age, and physical condition. Depending on the outcomes of these assessments, an index (Is) is established to characterize the seismic performance of the building. The estimated damage to the building during an earthquake is calculated by comparing the Is index with the comparison index (Iso) that is deemed appropriate for the building. This comparison is carried out separately for all critical floors and for two principal earthquake directions. If Is > Iso, the building is considered safe against earthquakes, otherwise, the seismic reliability of the building is considered uncertain. The second level of examination involves calculating the stiffness capacity of columns and beams using the load-carrying capacity method. The assumption is made that the beams in the structural system are rigid. The third level of examination takes into account the behavior of the beams. The SD and T calculations are the same as in the second level. In conclusion, The Japanese procedure relies on the seismic index (SI), which is derived from the basic seismic index, incorporating resistance and stiffness indices, irregularity index, and time index (TI), to forecast the overall seismic resistance capacity of a floor. The evaluation process is contingent on the various parameters mentioned earlier, and it may lack clarity in terms of categorizing buildings using a distinct scoring or rating system. The stages of the assessment require lengthy calculations, and the fact that it is only intended for buildings with six floors or fewer raises questions about its suitability for the building stock in many countries including Turkey.

In the study of Cardinali et al. (2022), a hybrid approach was developed to assess the seismic sensitivity of modern masonry buildings already in existence. This method was designed to work at two distinct levels: the urban scale and the building scale. The integration of both scales led to the creation of a hybrid approach, which involves a systematic application of analytical techniques at the urban level.In this research, buildings were spatially defined, and information related to their vulnerability was categorized and stored in a Geographic Information System (GIS) database, which also contained essential architectural and structural details. The hybrid method allowed for the reversible application of analytical studies in the city of Florence, and it introduced an innovative, rapid risk analysis method based on GIS. This method considered various factors, including building height, plan length, floor height, plan area, and the materials used in construction.

The outcomes derived from this procedure were subsequently utilized to fine-tune a simplified methodology that relied on the geometric and structural attributes mentioned earlier for a large dataset of buildings. This dataset was then employed to categorize the buildings into typological groups, with a specific case study selected for further analysis. In the analytical phase, taking into account specific characteristics and feedback, fragility curves were derived, aiding in the assessment of seismic fragility for the chosen case study. The findings were presented in the shape of fragility curves and damage scenarios, which underscored the various response patterns and assessed the susceptibility of these structures within the city of Florence. Creating a spectral displacement/spectral acceleration curve that takes into consideration the masses and base motion involved in the analytical computations is the aim of this procedure. It can be characterized as "hybrid" since it provides a thorough assessment of seismic vulnerability by combining statistical and simplified data with numerical results.

2.2.2. Rsv using building image and earthquake effect parameters

The Maras earthquakes in 2023, which affected 10 provinces and caused unbearable pain to our country, shows that it is very important to quickly and effectively determine the priority order of the risky buildings and make them ready for the earthquake, starting with the structures that require urgent intervention. In the realm of literature, rapid seismic assessment methods have been devised to ascertain the risk conditions of existing buildings. The primary goal of rapid assessment methods is not to provide a precise classification of whether buildings are earthquake-resistant or not. Instead, the aim is to swiftly and accurately evaluate the current condition of buildings and their earthquake resistance status, categorizing them based on their priority, with a focus on identifying the most vulnerable structures.

In a study by Sözen (2014) related to this topic, it was found that the evaluation of seismic risk of building stocks can only be achieved through detailed analysis tests and by shifting from an approach of searching for individual building safety to filtering out buildings with high vulnerability from the larger building stock. The efficacy of this approach hinges on the capacity to consistently identify high-risk buildings within a substantial building inventory through appropriate techniques. This filtering approach is very important in terms of providing information on where to start the solution to the problem with a rational approach. In the literature, many efforts have been made to propose rapid assessment methods for determining the seismic risk distribution of building inventories. In general, in these studies, quick building performance evaluations are conducted using methods that rely on observations made from the outside of the building and certain assumptions, considering both the soil class and the seismic impact.

A seismic risk assessment technique known as rapid visual screening (RVS) is employed to evaluate the earthquake resistance of typical reinforced concrete buildings with fewer than eight stories. This method, developed by Halûk Sucuoğlu in 2007, is widely utilized for assessing buildings in earthquake-prone regions. The assessment process begins by identifying several factors that influence the seismic performance of the building, such as the number of floors, the visual quality of the building, the presence of soft-story conditions, heavy loads, short column effects, collision risks, topographical effects, seismic hazards, and local soil conditions. Based on these factors, a base score is assigned to the building. Subsequently, the performance score is determined by taking into account the building's location within a seismic risk zone (earthquake zone) and referencing the relevant values from the Base Scores (BS) and Safety Scores (VS) table. The performance score is calculated by taking into account the scores of the observed safety effects, which are determined by the equation: $PS = BS \times VSMi \times Vsi$, where PS represents the performance score, BS represents the base score, VSMi represents the safety score multipliers, and Vsi represents the safety score. The performance score of a building in the range of 0 to 30, the building is considered to require the highest priority in terms of seismic assessment and further detailed analysis is necessary. Conversely, if the performance score exceeds 100, the building is deemed to necessitate the lowest-priority assessment. In summary, the RVS Method proves to be a valuable tool for appraising the seismic resilience of typical reinforced concrete buildings and determining the requisite measures for improving their seismic performance.

Fema 155-ATC-21 (1998) is a rapid assessment method developed by FEMA 155-ATC-21, for buildings located in earthquake-prone areas. It is based on visual observations and allows the identification of structures that may experience severe damage during an earthquake through street views. The method is easy to apply, inexpensive, and does not require any static calculations. There are three different zones available based on the seismicity of the area where the building is located (low, medium, and high-risk earthquake zones). The data collection form appropriate for the earthquake zone where the building is located is selected, and the main score of the building is obtained based on the type of building. Then, factors that may negatively affect the earthquake behavior of the building are identified on site. If the year of construction of the building is known, information about material properties, construction techniques, and regulations of that time can be obtained. Information about the researcher should be included in the form if necessary to obtain the researcher's observations during a more detailed investigation if required. Information affecting the performance of the building, such as the supporting system, the number of floors, floor area, and floor type, is used in building rating. When rating the building, the damage status of elements such as parapets, chimneys, exterior cladding, and roof openings that are not part of the supporting system and may pose a danger during an earthquake should be determined. Finally, a building sketch is drawn for the relevant area, and a photo of the building is taken and added to the form. The assessment stage consists of determining the supporting system and identifying the materials used, based on all the information collected, the information is rated. The final score (S) of the building, which is necessary for decision-making, is then obtained by subtracting the values of the factors that may change the earthquake performance of the building from this score. A higher final score means that the building has a higher earthquake resistance. If the score is less than 2, a detailed investigation is required. However, the opinions of the expert examining the building are more important. Based on this, ATC 21 categorizes existing structures into two groups. The first group consists of buildings with adequate earthquake performance (S>2), and the second group consists of buildings with insufficient earthquake resistance and requiring detailed investigation (S<2). FEMA 154, which has both a detailed and fast scanning method, is one of the cornerstones of rapid visual screening.

Guidelines for assessing a building's seismic performance can be found in the "Handbook for the Seismic Evaluation of Buildings" (FEMA-301) published by the Federal Emergency Management Agency (FEMA). This handbook is intended to assist engineers, building officials, and others who are responsible for evaluating buildings in seismic hazard areas. It offers suggested practices for developing seismic retrofitting plans and evaluating the seismic stability of older structures. The handbook covers the entire process of seismic evaluation, including the evaluation of seismic hazard, the determination of seismic demands, the evaluation of structural and non-structural components, and the development of seismic rehabilitation plans. The handbook also provides information on seismic rehabilitation techniques, codes and standards, and the development of seismic hazard maps. The aim of FEMA-301 is to provide practical information to help building owners, designers, and engineers make informed decisions about the seismic safety of their buildings.

Grünthal (1998) employed the European Macroseismic Scale to assess the risk level of both masonry and reinforced concrete structures. Vulnerability classes (A to F) were assigned based on different levels of invulnerability. In the study, masonry building types were categorized into various classes, including:Rubble stone/fieldstone

- Rubble stone/fieldstone
- Simple stone
- Massive stone
- Unreinforced brick/conrete blocks
- Unreinforced brick with RC floors
- Reinforced brick and confined masonry
- Reinforced concrete buildings

Vulnerability within the scope of operation factors affecting the structure, material, quality and workmanship, ductility, location, the condition of the buildings, their design, irregularity of the building shape, earthquake resistant design (ERD) level, etc. defined by taking into account other factors such as "Validity Table" within the scope of the study classifies the durability of structures in a manageable way, taking into account both the building type and other factors. The subdivisions of the structures, marked with letters from “A” to “F”, are roughly determined according to different levels of invulnerability, not from an architectural point of view. Different building behavior and failure types are shown for both masonry, wood, steel and reinforced concrete buildings. These safety classes, determined to determine masonry pictures risk measures, lacked the limit to the actual performance of the structures.

Aldemir et al. (2020) introduced a novel approach for evaluating the seismic risk of Unreinforced Masonry (URM) buildings. This method leverages binary logistic regression and draws from an extensive database comprising detailed seismic assessment analyses carried out on 543 URM buildings. The research presents an innovative and quick screening approach for URM buildings, which is based on the outcomes of comprehensive seismic assessments. The proposed method and its key estimation variables are outlined in the table below (Table 2.2.)

Table 2.2. URM subcategories and numeric displays of selected prediction variables of buildings

Estimation Variables	**Possible Values**			
Number of Stories, N	1 2	3 4	5 6	7 8
Seismic Zone, SZ	S_{DS} > 0.75g : 1	0.75g >S_{DS}≥0.50g : 2	0.50g>S_{DS}≥0.25g : 3	S_{DS} < 0.25g : 4
Soil Condition, SC	V_{s30} > 700m/s : 1	700>V_{s30}>400m/s : 2	400>V_{s30} >200m/s : 3	V_{s30} < 200m/s : 4
Age of Building, AG	Any integer value			
Structural System, SS	RC Frame : 0	RC Frame with Shearwall : 1		
Neighboring Structure Status, NS	Separate : 0	Adjacent : 1		
Short Column, SCol	No : 0	Yes : 1		
Vertical Irregularity, VI	No : 0	Yes : 1		
Overhang, OH	No : 0	Yes : 1		
Plan Irregularities, PI	No : 0	Yes : 1		
Soft Story, SoftS	No : 0	Yes : 1		
Position of Neighboring Slabs, SLoc	Levelled : 0	Non-leveled : 1		
Slope of the Soil, Sslop	Flat : 0	Sloped : 1		
Effect of Construction Date, CD	After 2007 : 1	Between 1997-2007 : 2	Between 1975-1997 : 3	Before 1975 : 4

A system based on penalty scores (PSi) was suggested in this study. Following the establishment of penalty scores for the various URM building weaknesses (i.e., seven separate estimating variables), each building was given a base score (BS) based on its seismic class (SC). When all penalty scores and base scores were added together, the building risk score (BRS) was determined (Equation 2.1). According to this method, URM structures were deemed "risky" or seismically sensitive if their BRS rating was less than zero. Other structures were labeled as "non-risky" or "seismally unvulnerable."

$$BRS = BS + \sum_{i=1}^{7} PS_i \qquad (2.1)$$

The database was filtered by seismic class after the data were transformed into a format compatible with the chosen statistical methodology. Binary logistic regression was then performed individually on each filtered database. SPSS statistical analysis program was used in all analyzes performed in the study. In summary, the working principle of the proposed method, the data set with categorical variables as seismic reliability of the buildings, the base score that can increase with the positive features of each building and the penalty points of the building negativities were tried to be estimated. The significance of the parameters was ascertained using logistic regression. The method obtained at the end of the study was tested with 100 buildings and a successful result was obtained with a 5% margin of error. This study, which is the disadvantage of other rapid scanning studies, does not require too many skilled personnel and does not require too much time, has contributed to the literature in the field of prediction of rapid earthquake safety class. Consequelty, this study concluded that this approach resulted in a promising method with some accuracy problems in predicting the test database.

In other study conducted by (Coskun et al. 2020), a new approach for fastly estimating seismic assessment results using statistical analysis methods was introduced. This method was specifically developed for assessing the seismic sensitivity of reinforced concrete buildings in Turkey and was based on the comprehensive evaluation outcomes of 545 such structures. To develop this new method, 400 of the detailed evaluation results were employed for training purposes. The variables used for estimating seismic assessment outcomes in this method are provided in the table below (Table 2.3).

Table 2.3. Numerical representation of the selected estimation variables

Estimation Variables	**Possible Values**							
Number of Stories, N	1	2	3	4	5	6	7	8
Seismic Zone, SZ	SDS > 0.75g : 1		0.75g >SDS≥0.50g : 2		0.50g>SDS≥0.25g : 3		SDS < 0.25g : 4	
Soil Condition, SC	Vs30 > 700m/s : 1		700>Vs30>400m/s : 2		400>Vs30 >200m/s : 3		Vs30 < 200m/s : 4	
Age of Building, AG	Any integer value							
Structural System, SS	RC Frame : 0		RC Frame with Shearwall : 1					
Neighboring Structure Status, NS	Separate : 0		Adjacent : 1					
Short Column, SCol	No : 0		Yes : 1					
Vertical Irregularity, VI	No : 0		Yes : 1					
Overhang, OH	No : 0		Yes : 1					
Plan Irregularities, PI	No : 0		Yes : 1					
Soft Story, SoftS	No : 0		Yes : 1					
Position of Neighboring Slabs, SLoc	Levelled : 0		Non-leveled : 1					
Slope of the Soil, Sslop	Flat : 0		Sloped : 1					
Effect of Construction Date, CD	After 2007 : 1		Between 1997-2007 : 2		Between 1975-1997 : 3		Before 1975 : 4	

In order to present a novel fast visual screening technique for reinforced concrete (RC) buildings, the risk characteristics of RC buildings as established by in-depth seismic assessment studies utilizing intricate numerical models were considered. The risk conditions found during these thorough evaluations were converted into binary values (RS) as part of this procedure, with 0 denoting "non-risky" buildings and 1 denoting "risky" ones. Then, using the database that was chosen, a statistical model was built to assess how the selected estimation variables affected the risk status (see Equation 2.2).

$$RS = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \cdots + \beta_n X_n \qquad (2.2)$$

The binary risk states of RC buildings (RS) and the chosen estimation variables (X1, X2,..., Xn) were integrated into the statistical model. Then, using STATA (2015), an ordinary least square (OLS) regression analysis was carried out. In the STATA (2015) software, the results of this analysis were referred to as marginal effects. In this statistical model, variables with marginal effect values close to 1 suggest that the presence of those variables would elevate the seismic risk level of RC buildings. On the other hand, marginal effect values close to zero suggest a decrease in seismic vulnerability and a positive influence on seismic risk.Additionally, multiple linear regression analysis was carried out in SPSS (2006) utilizing the linear statistical model as stated in Equation 2.2 and incorporating all of the chosen estimation variables (i.e., n= 14). As a result, according to the marginal effects of the parameters according to STATA (2015) and importance coefficients according to SPSS (2006), the new rapid scanning method was able to predict seismic risk analysis results with high accuracy (%80).

Kumar et al. (2017) made a prioritization study as a result of the scores made with the forms prepared for masonry and reinforced concrete buildings. As masonry building types, scores were collected for reinforced concrete beam, brick masonry, stone masonry, adobe and mixed masonry structures and statistical analyzes were made with gaussian distribution curves. The structures' damage estimates were categorized as having no damage, light, moderate, severe, and high risk of collapsing as a result of the scoring analysis. In the study conducted for 9099 buildings in the state of Himachal Pradesh in India, it was mentioned that since the rapid screening scores of the structures were close to each other, it was difficult in the classification stage, detailed analysis was made for the low-scored structures, but the results were out of the scope of this article.

Coskun (2019) proposed a method with a higher success rate as an alternative to the first-stage assessment method in accordance with the points specified in the Annex-A part of the

Regulation on the Principles of Determination of Risky Structures (RYTEİE-2019) within the scope of the Law on Transformation of Areas Under Disaster Risk No. 6306. In this study, Structure Score (SS) were developed by determining the importance coefficient of the structure determined from complicated statistical analyses performed in SPSS and STATA programs. Based on these importance coefficients, certain parameters had their scores modified, and new parameters like building age were introduced. The success rates were then compared with the method utilized under Law No. 6306. In the study, using 100 test data, the success rate of the method employed under Law No. 6306 was found to be 64%, while the success rate of the method referred to as "Structure Score (SS)" was 84%. It was concluded within the scope of the study that factors such as reinforcement class and concrete compressive strength had a significant impact. However, they were excluded from the method to streamline the process of generating a rapid seismic safety classification for buildings.

In another study, a database of 321 reinforced concrete school buildings in Istanbul was compiled (Mazılıgüney 2012). The construction year, average plan area, average floor height, average distance to fault, average concrete compressive strength, average steel tensile strength, and average duration were among the parameters that were identified in this study's initial assessment of school buildings. The World Bank-financed project involved an analysis of the buildings in question using Turkey's current seismic design code. The current preliminary seismic performance evaluation procedures used in the literature were compared. Mentioned methods, (FEMA 155, 1998, Halûk Sucuoğlu, 2007, (Yakut, 2004, Yücemen et al., 2004). Sucuoglu et al. and Ozcebe et al. procedures are not compatible for reinforced concrete school buildings in Turkey; Yakut procedures have a 50% success rate. It was concluded that the most successful method was ATC21 (FEMA 155) with a success rate of 78.68%.

The seismic assessment of historic brick-masonry buildings in Vienna was investigated in the study by Adam and Achs (2012), using the RVS (Rapid Visual Scanning) method as a foundation or reference. The study presented a fast visual scanning technique that uses penalty points for a number of structural characteristics, such as the ground condition, foundation, secondary structures, seismic hazard class, plan regularity, height regularity, horizontal stiffness, and local failure. Due to the consistent typology of these specific building types, the RVS methodology was adopted, enhancing the validity and quality of the seismic assessment. In this context, parameters associated with the structure, such as regularity, conservation status, and geometry of the examined buildings, were assessed. Additionally, the study considered the human and economic impact of earthquake-induced damage to the structures, the number of people exposed, and the significance of the buildings. Scores were calculated for each set of parameters, and based on these scores, the buildings were categorized into one of four vulnerability classes. The research, involving 375 buildings, led to the identification of damage maps and the marking of potentially at-risk structures on these maps.

For the determination of the seismic safety of buildings with a coefficient of 5 or more in the city of Chennai, FEMA (2002) developed a methodology by adapting its proposed RVS format technique to the buildings in Chennai City (Rajarathnam and Santhakumar, 2015). Aerial photographs were used to detect irregularities in buildings for the first time for a rapid data collection. It was noted in the study's scope that evaluating a big number of buildings one at a time takes a lot of labor and time, and that using the aerial photography approach on the GIS platform is a great way to get around this issue. Aerial photos were utilized in the study to pinpoint a few significant irregularities that contribute to the building's seismic sensitivity. Within the scope of the study, it can be determined from an aerial photograph that a building has plan irregularities based on the plan profile; It is also mentioned that the height of each building can be obtained by photogrammetric approach on photographic images of different elevations.

Figure 2.1. Detection of the irregularity in the plan and the floor height of the building from satellite images. (Rajarathnam and Santhakumar, 2015)

In the context of the study, measures to mitigate the vulnerability of structures with different irregularities were also addressed. The study concluded that these recommendations could be valuable in assessing the overall cost of upgrading structures at a site, serving as a preliminary study before conducting a more detailed assessment. In the proposed method, it is assumed that the combined influence of both plan and vertical irregularities contributes 100% to the damage effect. The relative weights assigned to these irregularities are 30% for plan irregularities and 70% for vertical irregularities. The methodology of this approach is outlined below.

- Zooming in on a building reveals its layer.

- Initially, the building ID is recorded, and the address/descriptions are entered into their respective columns.
- The visual inspection of the building begins with an examination of its floor plan, focusing on the lateral dimensions. Subsequent observations and explanations are provided based on the floor plan.
- The elevation assessment involves considering the height of each floor and documenting this information.
- While certain vertical irregularities, such as floor standing and floating columns, were not identifiable through aerial imagery, they were confirmed during the ground truth assessment and were evaluated alongside the irregularities detected through aerial imagery for the scoring analysis. The influence of an expansion joint in enhancing the building's smoothness was also taken into account during the ground truth verification and was integrated into the score calculation, especially when the expansion joint wasn't visible in the aerial photograph. In the end, a score is determined by considering both plan and vertical irregularities.
- A higher score signifies that a building is less susceptible to earthquake damage, whereas a lower score implies a higher vulnerability to earthquakes.
- All of these specific details are recorded in a Microsoft Excel spreadsheet and subsequently integrated into the Geographic Information System (GIS) attribute data.
- Building attributes can be extracted from the attribute table using aerial photographs of buildings in conjunction with the digital vector map.

Finally, the effect of each parameter and the combination of irregularities on the damage status is mentioned (Figure 2.2). The graphs indicate that the most adverse irregularities are associated with the combination of soft story and non-parallel systems. It was further noted that this is followed by the combination of diaphragm discontinuity and soft story irregularities. Additionally, the presence of a soft floor has been identified as one of the primary factors contributing to elevated damage levels in buildings in Chennai.

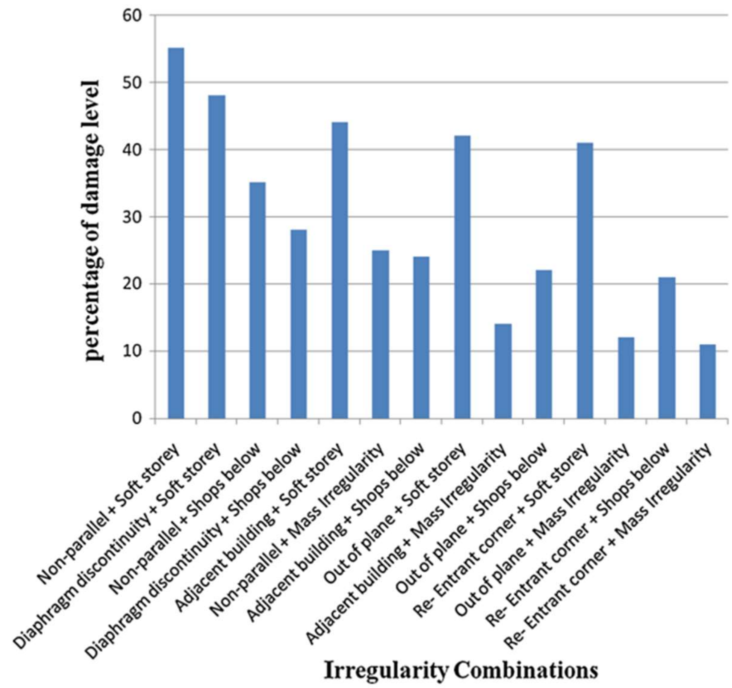


Figure 2.2. Effect of building properties on percentage of damage

The study suggests a fast and low-cost method that can be used for risk reduction in general, to determine the vulnerability of GIS-based buildings by classifying them between A-D, and to develop the effect coefficients by evaluating the building parameters.

2.2.3. Artificial intelligence-assisted rapid assessment

Studies on artificial intelligence-supported rapid evaluation in the field of construction engineering are becoming increasingly significant as AI technologies are leveraged to enhance various aspects of civil engineering projects and infrastructure management. Within the construction industry, numerous research endeavors have been undertaken, encompassing areas such as Risk Assessment and Management, Structural Health Monitoring, and Construction Quality Control, with the utilization of artificial intelligence. These studies in the realm of civil engineering are predominantly centered on the development and implementation of artificial intelligence models, the seamless integration of sensor technologies, the harnessing of remote sensing techniques and data analytics, and the creation of decision support systems. The overarching objective of these research initiatives is to fortify the security, resilience, efficiency, and sustainability of civil infrastructure and projects. In this pursuit, researchers consistently seek novel and inventive ways to leverage image processing and machine learning to effectively address the distinct and intricate challenges encountered within the realm of civil engineering.

In their study, Achs and Adam (2012) introduced a method for the rapid seismic assessment of historical brick-masonry buildings in Vienna, Austria, employing a visual screening approach. The aim of the study was to develop a fast and efficient approach for evaluating the seismic safety of these buildings and to identify any potential structural weaknesses that may need to be addressed in order to improve their resistance to earthquakes. The study used visual inspections and limited non-destructive tests to gather data on the condition of the buildings and their structural components. The study's results revealed that the visual screening method proved to be a valuable tool for rapidly and accurately evaluating the seismic performance of historic brick-masonry buildings in Vienna.

The study conducted by Milosevic et al. (2020) provides a methodology for defining fragility curves, which are graphical representations of a building's probability of exceeding various levels of damage over a given period of time. Ground motion density the study applies the methodology to a mixed masonry and reinforced concrete (RC) building stock and aims to measure their seismic fragility. The methodology involves performing nonlinear static analyzes on representative building models to derive fragility curves. The results of the analyzes are used to measure the seismic performance of the building stock and to identify key factors contributing to its seismic vulnerability. The study demonstrates the utility of the fragility curve approach for assessing the seismic performance of mixed masonry reinforced concrete stocks and provides valuable insight into the considerations that influence their seismic behavior. The results of the research can be applied to enhance the seismic design and reinforcement of mixed masonry-reinforced concrete structures, ultimately mitigating the risk of damage and collapse in the event of earthquakes. The fragility curve approach can also be applied to other building stocks to assess their seismic susceptibility and inform decision-making regarding seismic hazard mitigation measures

In a study proposed by Rashidi et al.2016, the detection of the building material type is aimed at using Multi-Layer Perceptor (MLP), Radial Basis Function (RBF) and Support Vector Machine (SVM) algorithms in machine learning. The study aims to differentiate between three distinct building material types within a dataset of building images. This is accomplished through a two-step approach that includes feature extraction and classification. In the feature extraction stage, color and texture attributes of the image are extracted either for individual pixels or for clusters of neighboring pixels, which are often referred to as image blocks. Subsequently, these feature vectors are subjected to classification by a specific algorithm to accomplish the task of material detection. The primary goal of the research is to assess the outcomes of building material detection algorithms using various machine learning techniques. Instead of analyzing individual pixels, the researchers adopted a block-based approach for material identification. The reason for adopting this block-based approach is that images of construction materials typically display consistent and interconnected pixel patterns. It is more efficient to process groups of pixels together rather than analyzing each pixel in isolation. In this suggested block-based method, each block in the target image consists of a set of m x m pixels, and the pertinent features are extracted from these blocks before being fed into a classifier for the identification of building materials. At the end of the study, visual estimation was made in three categories: concrete, brick, and others, and these estimates were evaluated with machine learning algorithms, resulting in successful outcomes (Figures2.3-2.5).

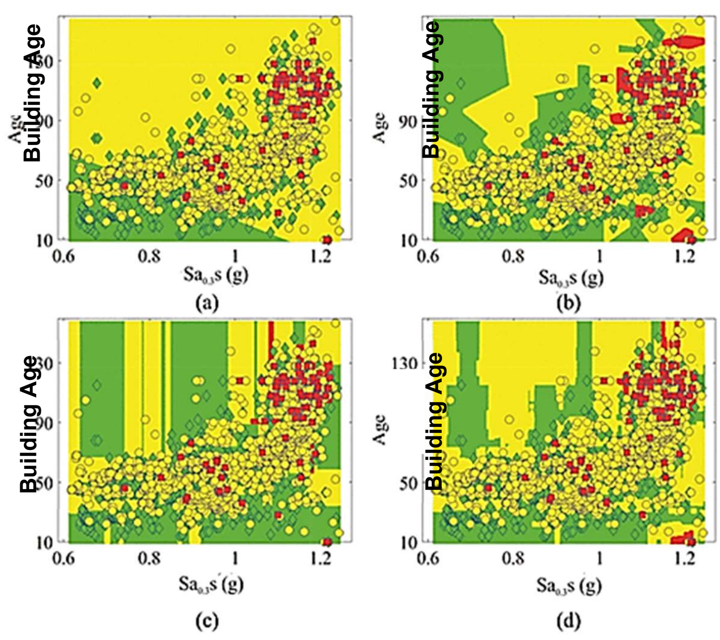


Figure 2.3. Analysis of age and ground acceleration with machine learning algorithms (Mangalathu vd., 2020)

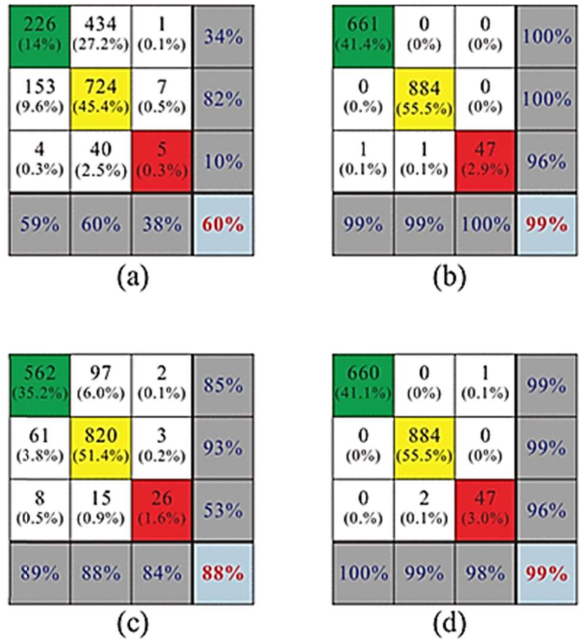


Figure 2.4. Performances of various machine learning techniques for the training set: (a) LDA, (b) KNN, (c) DT ve (d) RF (Mangalathu vd., 2020)

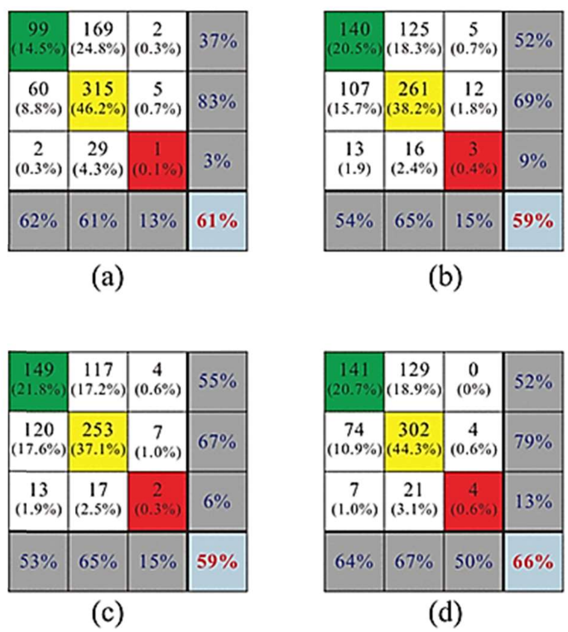


Figure 2.5. Performance of various machine learning techniques for test set: (a) LDA, (b) KNN, (c) DT ve (d) RF (Mangalathu vd., 2020)

Zhang et al. (2018)suggests the application of ML federate algorithms, particularly top-performing decision trees and random forests, for predicting the damage that a building might endure after an earthquake. The study categorizes "unsafe conditions" as positive class and "safe conditions" as negative class, as unsafe conditions are considered as critical and negative. The proposed machine learning framework is applied to reinforced concrete buildings for assessing structural safety. The significance of selecting optimal parameters for the efficient running of the study is also emphasized. In the study, the most appropriate parameters are selected through grid search methods. The study first calculates the probability of exceeding certain limit states based on either a known earthquake event or the post-earthquake damage status of the building. Secondly, The probability of surpassing limit states is determined by considering the risk posed by both the primary earthquake event and any subsequent aftershocks. A Markov process model is utilized to estimate the probability of a building transitioning from different damaged states to another after an earthquake, in both assessment categories. The study concludes that while a detailed examination of the structure with the proposed method may take time, accurate results were obtained in estimating the damage situation by using the obtained data.

Yucemen et al. (2004) proposed a model combining engineering and Statistical techniques are employed to compute the potential damage of a building by considering seismic factors such as ground motion intensity and the structural characteristics of the building. The methodology uses a statistical approach and takes into account the uncertainty in the input data. The results of the methodology applied to earthquake damage data from the 1999 Düzce earthquake showed that the model can accurately predict the damage potential of buildings under different seismic scenarios. The proposed methodology can provide useful information for seismic assessment and retrofit planning, and serve as a tool for decision-making in earthquake engineering.

Xu et al. (2019) proposes a machine learning approach to detecting damage to buildings in satellite imagery. The authors use convolutional neural networks (CNNs), a type of deep learning algorithm, to classify images of buildings as damaged or undamaged. The study shows that CNNs can achieve high accuracy in detecting building damage, even with low-resolution and limited information satellite images. The methodology used in the study can be applied to post-disaster assessments and help first responders and other organizations quickly identify damaged buildings and prioritize response efforts. Valuable for its contributions to the literature, this method concludes that the use of CNNs only for building damage detection in satellite imagery besides rapid assessment is a promising area of research and suggests that further studies could explore the use of other machine learning algorithms and improve the accuracy of the damage detection system.

In a study by Kumari et al. (2022), the researchers explore the utilization of machine learning algorithms and web-based procedures for estimating the damage scores of pre-existing buildings. Utilizing the Django web framework in the Python programming language, this study evaluates the efficacy of different machine learning algorithms, including decision trees, random forests, and support vector machines, for predicting building damage scores. The performance of these algorithms is contrasted with the conventional visual inspection approach. Furthermore, the research investigates the potential of employing a web-based platform to enable building owners and stakeholders to access the estimated damage scores. The study's findings indicate that the use of machine learning algorithms can enhance the precision of damage score estimates, and a web-based platform offers a convenient and easily accessible means of obtaining these results. The authors concluded that the proposed approach has the potential to be a valuable tool for decision-making in the evaluation and management of existing buildings following natural disasters.

The work by Stepinac and Gašparović (2020) provides a comprehensive examination of the latest technological advancements related to the assessment of safety and seismic vulnerability in masonry structures. The authors analyze various novel technologies such as machine learning, computer vision, and remote sensing, among others, and their potential utilization in masonry assessment. The research offers a framework for the efficient and accurate gathering of extensive spatial data for damage detection and mapping, leveraging the capabilities of modern technologies. The study, in particular, concentrates on the application of unmanned aerial vehicles (UAVs) for assessing building damage. The authors emphasize the importance of performing damage assessments with drones due to the limitation of visibility from ground level when making on-site assessments. The authors also discuss the use of different cameras such as RGB, thermal, multispectral, and hyperspectral cameras, as they have the capability to detect damage and anomalies in a building that cannot be observed by the naked eye. The study makes a substantial contribution to the literature by presenting a damaged situation through the use of 3D modeling of the structure utilizing the data obtained from the cameras. The article also highlights the need for further research and development to integrate these technologies effectively for practical use in the evaluation of masonry structures. The authors assert that the development and application of these technologies can significantly enhance the safety and reliability of existing masonry structures in the event of earthquakes and other natural disasters.

Shah (2016) assessed the seismic susceptibility of pre-existing masonry structures in two Jeddah, Saudi Arabian districts. The Rapid Visual Scan (RVS) method, a streamlined process for evaluating a building's seismic susceptibility through visual inspection of important structural elements like walls and foundations, was used in the study. An analysis of the outcomes from the RVS assessments was conducted to ascertain the overall susceptibility of the buildings in the two study regions. The findings indicated that a considerable portion of the buildings in these areas were at risk of experiencing seismic damage. The efficacious outcomes of this study can function as a basis for subsequent evaluations of seismic hazards and retrofit endeavors in Jeddah, thereby augmenting the area's resilience to earthquakes. In addition to highlighting the potential benefits of using the RVS method as a rapid assessment tool for evaluating building vulnerability in the context of seismic hazard assessment, this study highlights the significance of taking into account the seismic vulnerability of Jeddah's existing masonry structures.

(Ruggieri et al., 2021) The article "Machine Learning-Based Vulnerability Analysis of Existing Buildings" investigates the application of machine learning methods for assessing the vulnerability of pre-existing buildings in the context of seismic events. The study's findings suggest that the machine learning-based approach is effective in accurately forecasting the susceptibility of buildings to seismic events, thereby offering valuable insights for seismic evaluation and retrofit planning efforts. One of the tools used in the study was VULMA, which involved the labeling and posting of photos obtained from Google Street View. The type of roofing material, the duration of intervention, the presence of negative parameters such as heavy overhang, vertical vision, and the number of floors were labeled by a team that held the photographs. In another tool called Bi-Vulma, the tagged controls were not created using an artificial neural network, and the validation success rate was over 97%. The results suggest that the system has the potential for further development, including the possibility of using artificial neural networks to create a Decision Support System. The study also addressed the issue of data collection and the importance of ensuring the availability of sufficient and high-quality data for efficient forecasting models. The authors explained the selection of the data set to avoid unbalanced data and memorization, which are common issues in machine learning studies. The proposed framework can contribute to the literature by providing a method that can be developed using artificial neural networks and includes information such as the software interface, photos of buildings, and the number of people living in them. The sequencing scheme shown in Figure 2.6 contains the logical methodology of the study outlined. (Figure 2.6)

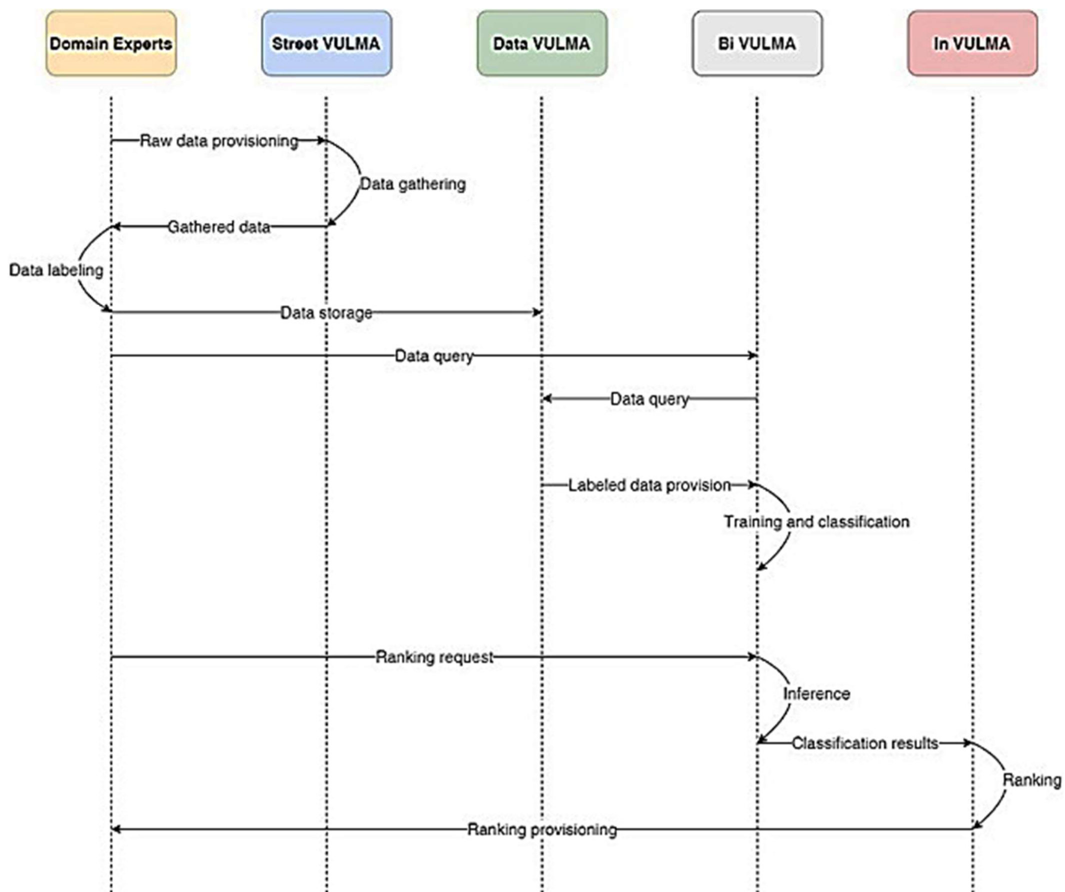


Figure 2.6. The working methodology and flowchart of the finding method

3. RESEARCH OF METHODOLOGY

In this research, a modern approach has been introduced to rapidly and precisely forecast the seismic performance of buildings using Machine Learning algorithms and artificial neural networks. This method obviates the necessity for technical personnel to physically enter the building, offering a convenient and efficient solution. As a result of the detailed analysis of a building, if the V_e/V_r value, which is the ratio of the shear force to the total floor shear force on the walls exceeding the risk limit, is above 0.5, it is said that the building is risky In Law No. 6306. Within the scope of this study, a method will be presented that will enable to give priority to structures with very high RS ratio by estimating the RS ratio and classifying the risky structures within themselves with both the detection of risky structures by machine learning and regression analysis and classification algorithms.

Within the scope of the study, a simple analysis of 12 parameters was made with the data, most of which were obtained from the protocols made by the Ministry of Environment, Urbanism and Climate Change and Gazi University, with detailed seismic analysis results and visual characteristics of the examination forms.

The data obtained from the Ministry of Environment, Urbanization and Climate Change and Gazi University became suitable for analysis first. First of all, since the scarcity of non-risky structures in more than four thousand data will cause the algorithms to memorize, these data have been reproduced synthetically in the training dataset with the SMOTE method. After that, the outliers in the data were either deleted or the average values were written and the operations were continued. Missing data were filled in parameters not included in the data set, such as short-term acceleration value, obtained from the Turkey Earthquake Zones Map. In order to decide how many results the data set should be divided into, the culuster selection in the Dendograms and Unsupervised methods was made. As a result, the more classes we could divide the risk into, the lower the cost of the detailed analysis to be done in Turkey as a result of the prioritization to be made. Then, under the title of feature engineering, the feature importance tables were prepared by considering the effect of the parameters on the result, and then 4 parameters were selected to facilitate effective and fast determination of the analysis results. In this context, Dimensionality Reduction methods are used in parameter selection. Among the selected parameters, it was concluded that it was not important with this data set, but considering the engineering knowledge, scientific and technical data, the required parameters were tried to be added to the detection method by multiplying the square of the parameters or other important parameters. Editing the data set, feature engineering and Dimensionality Reduction followed by modeling. In this section, success rates were compared under two sections, regression and classification, and all these algorithms were combined by selecting the parameters that would yield the most accurate results in the ensemble learning section. High success rates have been achieved by estimating algorithms with the voting method and Gradient Boosting. So as to obtain more accurate results, a hybrid method is obtained by using multivariate adaptive regression spline (MARS) used in regression analysis and classical machine learning algorithms, Gradient Boosting, Decision Tree Classifier, Logistic Regression, Random Forest, K-Mean Clustering, Support Vector Machine, LGBMClassifier, MLPClassifier.

3.1. Data Pre-processing

In this section, the procedure for addressing outliers will commence by initially scrutinizing the overall distribution and fundamental statistical outcomes pertaining to the dependent and independent variables within the dataset. Subsequently, an evaluation of data homogeneity and binary associations will be conducted through the application of parametric or nonparametric analysis tests. Following this, data pre-processing will culminate in the implementation of dimensionality reduction techniques and the optimal parameter selection, achieved through the utilization of Exploratory Data Analysis (EDA) methodologies.

3.1.1. Data description and basic data analysis

It is known that Turkey is an earthquake-prone country. Every major earthquake causes many casualties and economic losses by reason of the most of the buildings in Turkey have been built without engineering services. A rapid visual screening method was needed to filter the non-vulnerable buildings from the building stock as quickly as possible before or after the earthquake. Therefore, screening methods must finded by study solve this problem that rapid visual risk detech. This thesis examines the risk classification of structures built in Turkey between 1960 and 2021. The data used are 4,356 mansonry structure data obtained from the Turkish Ministry of Environment, Urbanization and Climate Change, engineering firms and Gazi University. In this context, twelve features were determined as inputs to predict the target feature, the risk situation. The data includes the structures of the masonry system. These data were labeled by obtaining detailed seismic analysis results, and risk classification was made in the labeling section by considering the ratio (V_e/V_r), which expresses the ratio of the shear force coming to the floor to the total shear force.

This study primarily aimed to create a straightforward and rapid visual screening method for predicting the damage level of masonry buildings using machine learning algorithms. The approach intended to collect the necessary parameters without requiring technical personnel to physically enter the potentially risky buildings. In essence, the study focused on developing a computational network based on external observations of structural properties. To achieve this, the network was trained using a substantial dataset of buildings for which detailed seismic risk assessment results were available. This dataset included 3,484 real buildings with seismic risk assessment analyses. Subsequently, the optimal ratio for data splitting, the performance of the algorithm in estimating seismic risk and damage levels was evaluated using 872 additional test real buildings with seismic risk assessment analyses (Joseph, V. R.2022).The machine learning algorithm created in this process is capable of estimating the risk and damage level that a structure may experience during a potential earthquake event. The earthquake risk analyses for the buildings in the database were conducted using the detailed seismic risk analysis method specified in the Urban Transformation Law No. 6306 (GABHR 2012) or the Turkish Earthquake Code (TEC 2007). In this context, detailed data such as architectural plans, material strengths, and other physical properties for all buildings were accessible. Technical analyses were carried out on these structures to train the machine learning algorithms. Before using the raw database, data preprocessing was conducted to eliminate irrelevant or misleading information, such as removing null values and categorizing selected parameters. Additionally, Table 3.1 provides information about the distribution of these parameters for reference.

Table 3.1. Brief summary of selected attributes

Abbreviation	Variable names	Data Type
EZ	Earthquake Zone	Categorical/Nominal
BA	Building Age	Numerical
NF	Number of Floors	Numerical
FA	Critical Floor Area	Numerical
BM	Building Mass	Numerical

VQ	Visual Crack and Mortar Quality	Categorical/Ordinal
CR	Compressive Strength	Categorical/Nominal
SW	Specific Weight	Categorical/Nominal
SS	Shear Strength	Categorical/Nominal
TS	Diagonal Tensile Strength	Categorical/Nominal
SD	Design spectral acceleration coefficient of 1 s	Numerical
SC	The short period (0.2s) is the design spectral acceleration coefficient	Numerical
RS	Classified ratio of the floor shear force on the risk walls to the total shear force	Numerical

In Turkey, masonry structures are typically constructed with fewer than four stories. However, in regions with low seismic activity, it is possible to use masonry structures for buildings of up to 8 stories. It's important to note that such cases are relatively rare. Consequently, the database used in this study was intentionally designed to include a limited number of masonry buildings with more than four stories, comprising only 25% of the total database.

The detailed seismic assessment of the buildings followed the guidelines outlined in GABHR (2012). In the numerical models used, all piers were represented by 2-node 3-D frame elements, and all slabs were modeled with 4-node thin shell elements. The modulus of elasticity in the numerical models was calculated using the formula (i.e., 200fm) specified in GABHR (2012).

Within each story, a rigid diaphragm was defined, particularly when a reinforced concrete slab was present. Subsequently, a response spectrum analysis was conducted, taking into account a reduced design spectrum (R=2), with separate analyses performed for two orthogonal directions. During the response spectrum analysis, the requirement for 95% mass participation was met in each orthogonal direction.

In the detailed seismic assessment analysis, the performance limits for slab elements were determined. The performance of each pier was categorized as Minimum Damage (MD) if the pier had the capacity to withstand the reduced design spectrum and gravity demands, or as Collapse Damage (CD) if it lacked the necessary capacity. The capacity of each pier was estimated based on the effective height, considering the full height, and considering modes of failure as specified in TEC 2018, such as diagonal tension and base sliding. Additionally, the axial load demands were compared with the axial load capacities of each pier, with a correction applied based on the slenderness ratio. The correction factor was set at 1 for slenderness ratios less than 6, 0.8 for slenderness ratios between 6 and 10, 0.7 for slenderness ratios between 10 and 15, and 0.5 for slenderness ratios exceeding 15. Piers with an axial load capacity lower than the demand were classified as Collapse Damage (CD).

The overall performance of each masonry building was considered satisfactory in terms of life safety if less than 50% of the total base shear at the first story was resisted by masonry piers categorized as having a Collapse Damage (CD) performance level.

From these two large databases (populations), three sampled databases (samples) were derived for use in machine learning algorithms. At this stage, the Latin Hypercube sampling method was used (MacKay et al. 1979). As it is known, in Latin hypercube sampling, it must first be determined how much data will be used in the sampling database to be created. Since it was determined by preliminary calculations that the resulting population databases did not require a great deal of time in the machine learning algorithm study with the computers used within the scope of the project, it was decided to use 80 percent of the entire population data in the training database and 20% in the test database.

The dataset used to determine the outcomes of this section is presented in Appendix A.

In this part of the study, frequency graphs will be shown to understand the reference range of the obtained verification and test data and the distribution of the parameter properties within the data set.(Table 3.2. and 3.3.) Thanks to these graphs, if the distribution of a feature of any parameter in the data is too much or too little, this parameter may not give accurate results. In addition, outliers within the parameter can be clearly seen in these graphs. (Table 0.7 and 0.8)

The tables contain statistical summaries of the data. When the statistical summaries were examined, parameters with large differences between minimum and maximum values (such as BA and BM) were detected. Additionally, it has been determined from these tables that there are many values with high standard deviations. This strengthens the possibility that there may be outliers in the data set. For this reason, data preprocessing must be applied to the data set.

Outliers that could significantly affect the analysis results were handled on a feature-by-feature basis. Each feature was examined individually and unrealistic values were removed based on predetermined thresholds. For example, as an upper limit, the building mass is limited to 500000, the coefficient is 7, and the building area is limited to 1000. Rows with lower values of 0 have been removed.

Table 3.2. Number of Variables and Statistical Summaries for the Masonry Buildings Training Database

Variable	Abbreviation	Data Type	Distribution in the Database	
Number of Floors	**NF**	**Numerical**	**mean**	**2.358783**
			std	**1.149407**
			min	**1.000000**
			25%	**1.000000**
			50%	**2.000000**
			75%	**3.000000**
			max	**6.000000**
Critical Floor Area	**FA**	**Numerical**	**mean**	**107.402758**
			std	**58.476722**
			min	**12.650000**
			25%	**72.087500**
			50%	**95.550000**
			75%	**126.000000**
			max	**953.500000**

Building Age	BA	Numerical	mean	43.068312
			std	13.335290
			min	3.000000
			25%	32.000000
			50%	42.000000
			75%	52.000000
			max	82.000000
Building Mass	BM	Numerical	mean	336.093201
			std	303.439074
			min	6.000000
			25%	111.557525
			50%	263.535000
			75%	444.812575
			max	2003.330000
Design Spectral Acceleration Coefficient of 1 s	SD	Numerical	mean	0.431811
			std	0.175160
			min	0.068000
			25%	0.326000
			50%	0.416010
			75%	0.548258
			max	1.042000
The short period (0.2s) is the design spectral acceleration coefficient	SC	Numerical	mean	0.991995
			std	0.325233
			min	0.170010
			25%	0.821000
			50%	1.011000
			75%	1.165000
			max	2.335000

Classified ratio of the floor shear force on the risk walls to the total shear force	**RSV**	**Numerical**	**mean** **std** **min** **25%** **50%** **75%** **max**	**0.803622** **0.249803** **0.000000** **0.620000** **0.930000** **1.000000** **1.000000**
Earthquake Zone	**EZ**	**Categorical/Nominal**	**1** **2** **3** **4** **5**	**1228** **556** **1520** **60** **116**
Compressive Strength	**CR**	**Categorical/Nominal**	**1.2** **1.4** **1** **0.5**	**1293** **1984** **12** **195**
Shear Strength	**SS**	**Categorical/Nominal**	**0.15** **0.18** **0.1**	**2862** **427** **195**
Diagonal Tensile Strength	**TS**	**Categorical/Nominal**	**0.25** **0.18** **0.1** **0.2**	**1291** **1988** **195** **10**

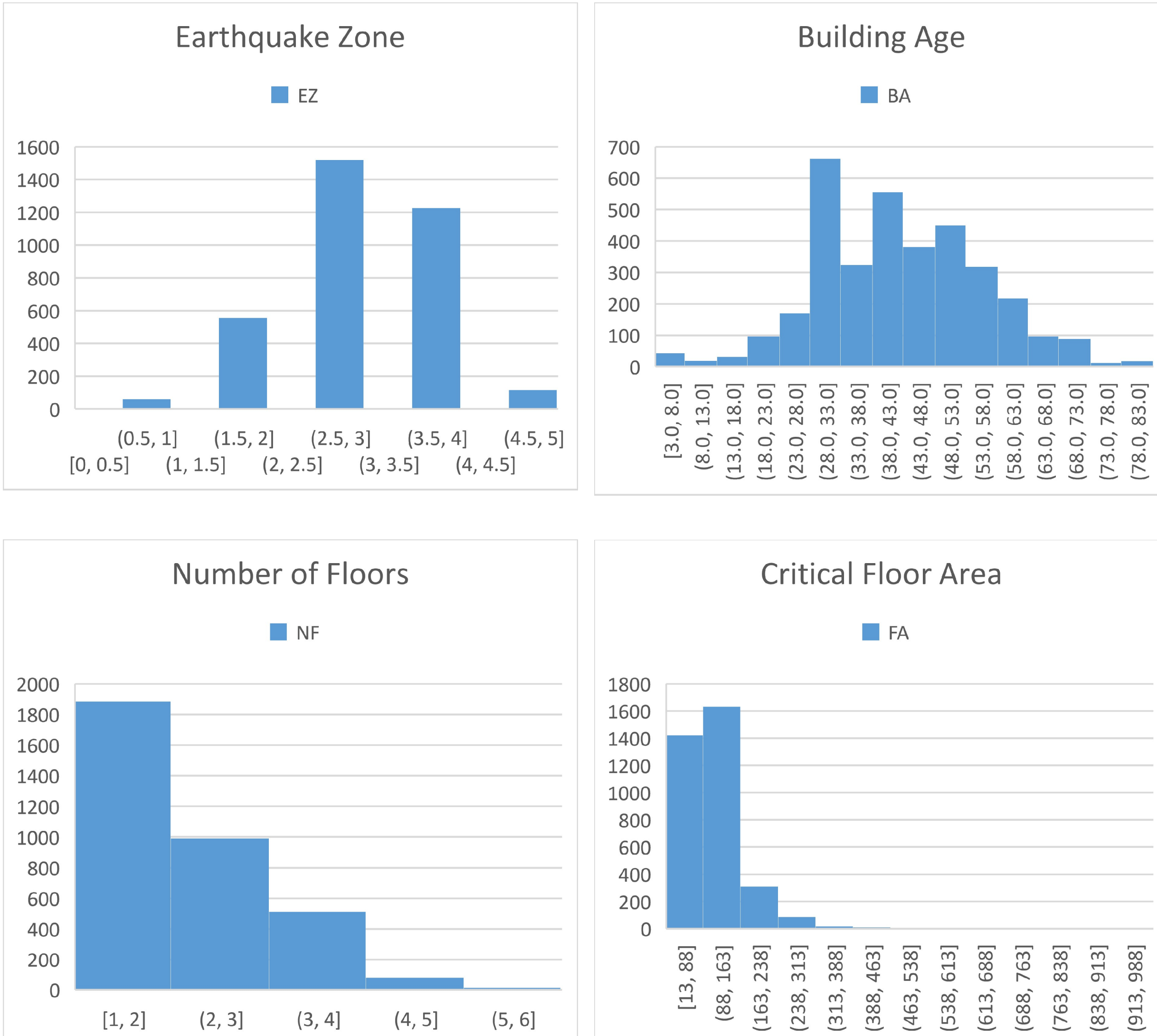
Table 3.3. Number of Variables and Statistical Summaries for the Masonry Buildings Test Database

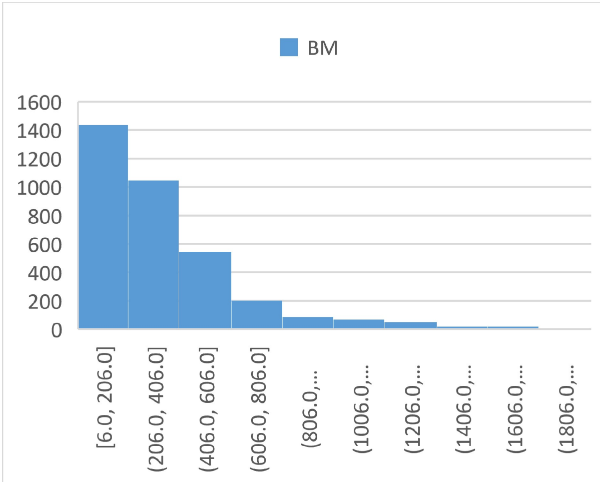
Variable	Abbreviation	Data Type	Distribution in the Database	
Number of Floors	NF	Numerical	mean	2.361239
			std	1.140633
			min	1.000000
			25%	1.000000
			50%	2.000000
			75%	3.000000
			max	7.000000
Critical Floor Area	FA	Numerical	mean	106.834908
			std	53.750313
			min	10.540000
			25%	73.045000
			50%	96.580000
			75%	127.000000
			max	500.000000
Building Age	BA	Numerical	mean	43.036697
			std	14.213269
			min	4.000000
			25%	32.000000
			50%	42.000000
			75%	52.000000
			max	82.000000
Building Mass	BM	Numerical	mean	339.708337
			std	310.105220
			min	12.350000
			25%	113.050000

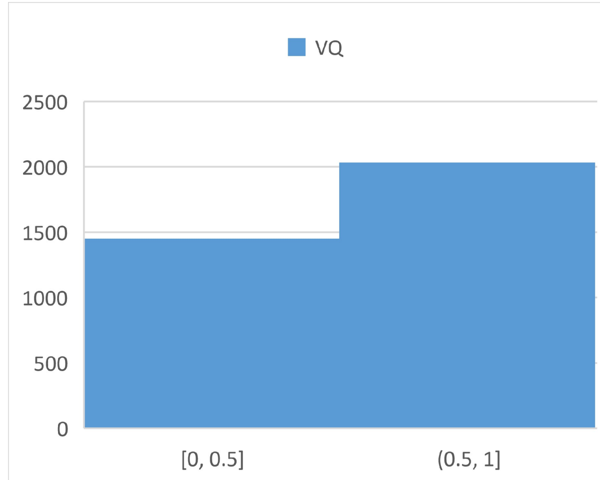
			50%	268.425000
			75%	439.577575
			max	1978.420100
Design Spectral Acceleration Coefficient of 1 s	**SD**	**Numerical**	mean	0.437660
			std	0.178701
			min	0.074000
			25%	0.331008
			50%	0.431005
			75%	0.553010
			max	1.502000
The short period (0.2s) is the design spectral acceleration coefficient	**SC**	**Numerical**	ean	0.996670
			std	0.327448
			min	0.176000
			25%	0.817750
			50%	1.015000
			75%	1.176000
			max	2.006000
Classified ratio of the floor shear force on the risk walls to the total shear force	**RSV**	**Numerical**	mean	0.812729
			std	0.250238
			min	0.000000
			25%	0.640000
			50%	0.930000
			75%	1.000000
			max	1.000000
Earthquake Zone	**EZ**	**Categorical/Nominal**	1	13
			2	142
			3	372
			4	320
			5	25

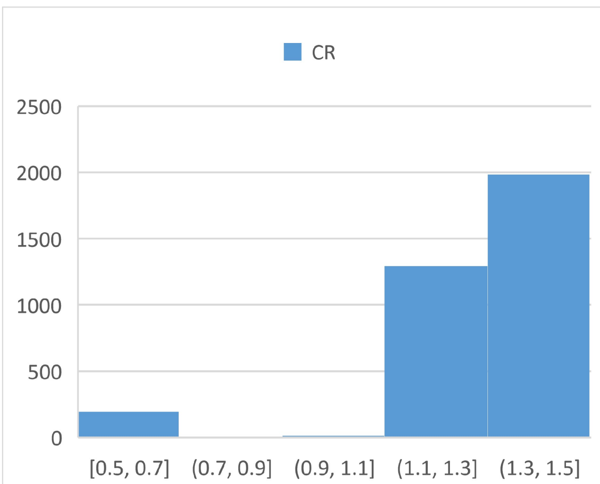
Compressive Strength	CR	Categorical/Nominal	1.2 1.4 1 0.5	348 483 37 4
Shear Strength	SS	Categorical/Nominal	0.15 0.18 0.1	732 103 37
Diagonal Tensile Strength	TS	Categorical/Nominal	0.25 0.18 0.1 0.2	346 485 38 3

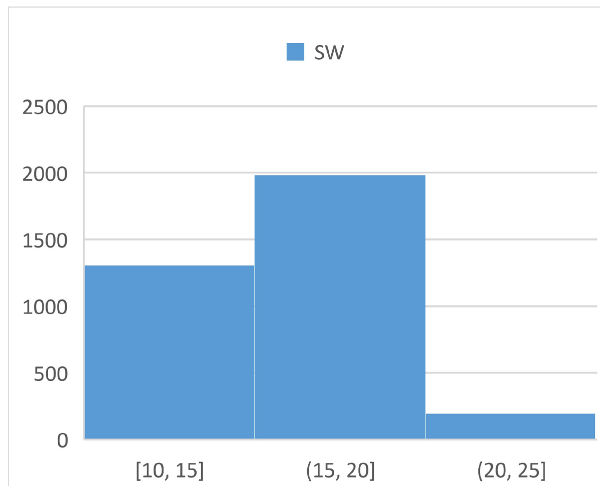
Table 3.4. Training dataset distribution

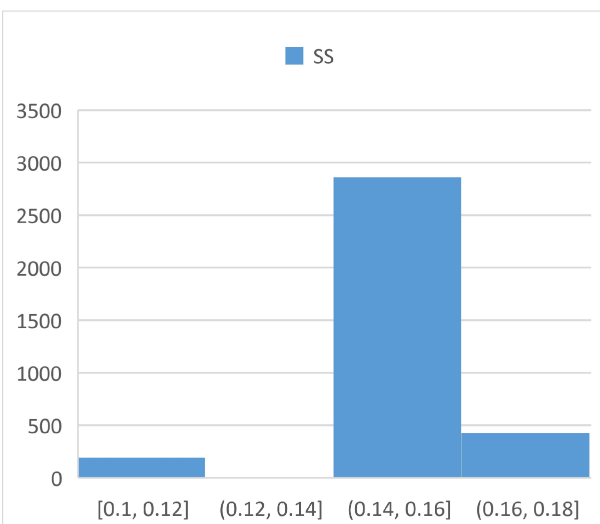


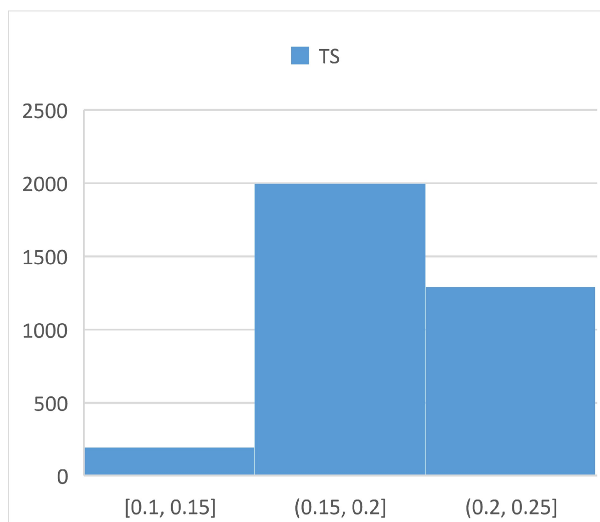

BM
1600
1400
1200
1000
800
600
400
200
0
[6.0, 206.0]
(206.0, 406.0]
(406.0, 606.0]
(606.0, 806.0]
(806.0,...
(1006.0,...
(1206.0,...
(1406.0,...
(1606.0,...
(1806.0,...

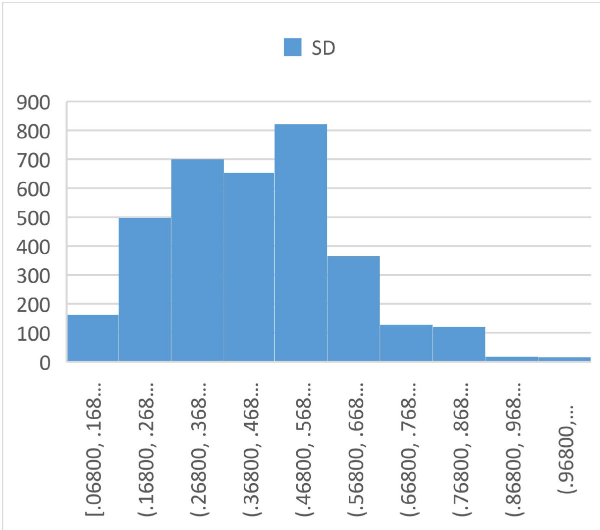

VQ
2500
2000
1500
1000
500
0
[0, 0.5]
(0.5, 1]

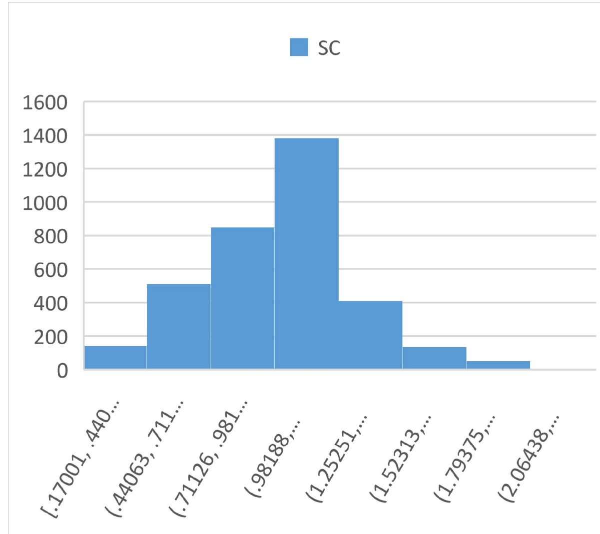

CR
2500
2000
1500
1000
500
0
[0.5, 0.7]
(0.7, 0.9]
(0.9, 1.1]
(1.1, 1.3]
(1.3, 1.5]

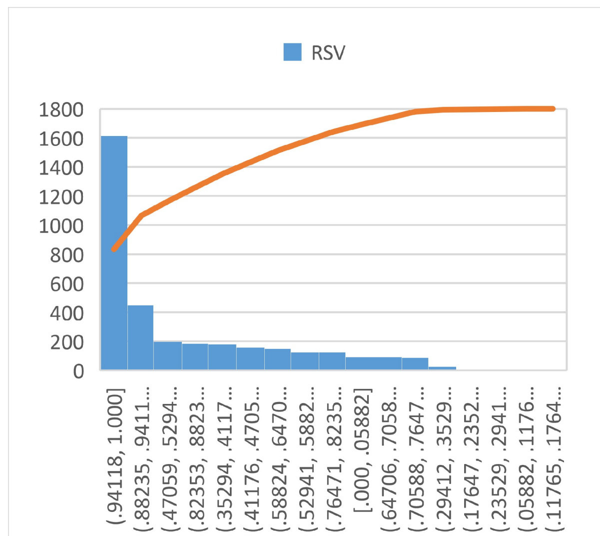

SW
2500
2000
1500
1000
500
0
[10, 15]
(15, 20]
(20, 25]


SS
3500
3000
2500
2000
1500
1000
500
0
[0.1, 0.12]
(0.12, 0.14]
(0.14, 0.16]
(0.16, 0.18]


TS
2500
2000
1500
1000
500
0
[0.1, 0.15]
(0.15, 0.2]
(0.2, 0.25]

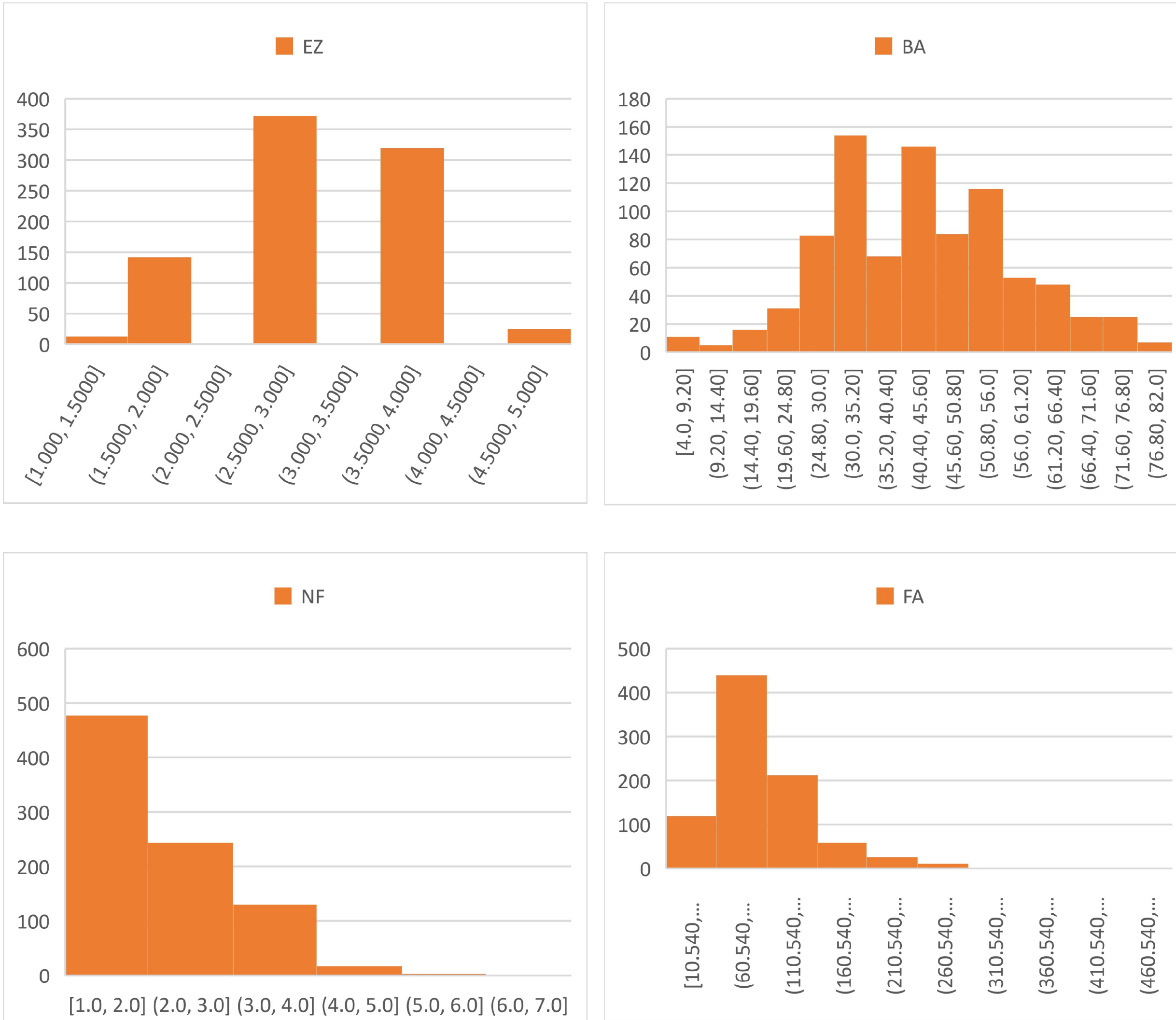


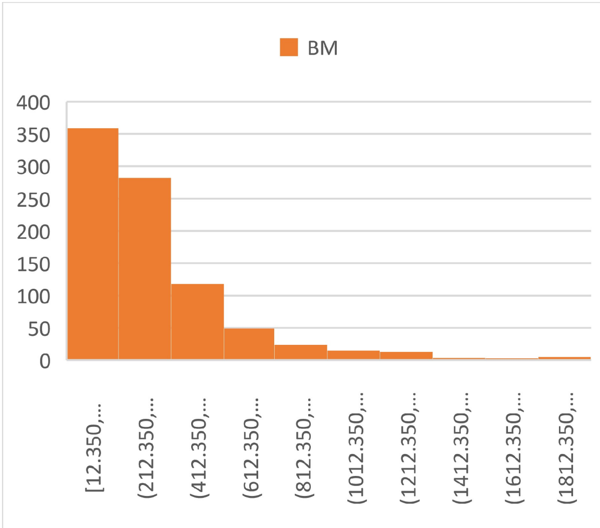


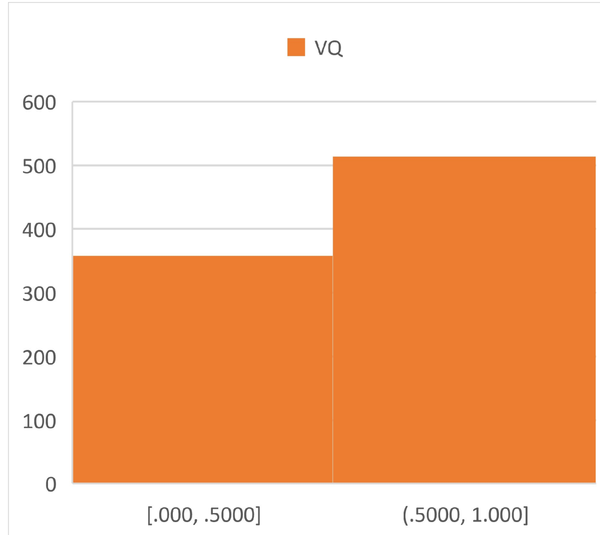


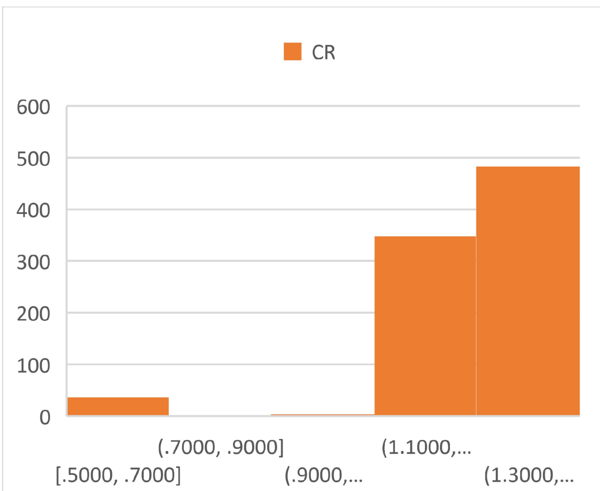
The left skewness in RVS and FA values is caused by outlier data and unbalanced data. A more balanced data set was obtained by classifying the RVS values in the following sections and deleting the outliers in the fa values.

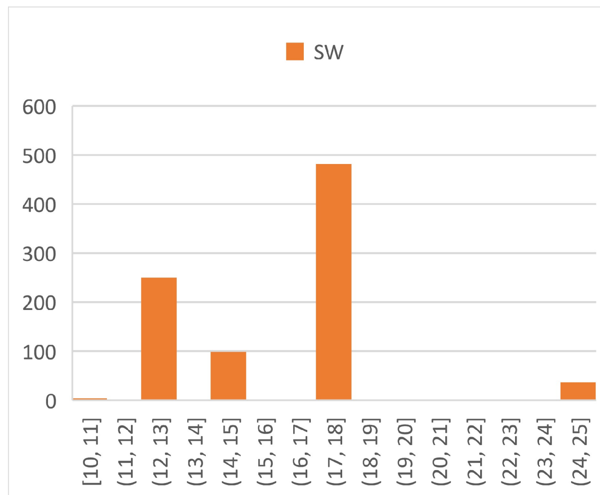
Table 3.5. Test dataset distribution

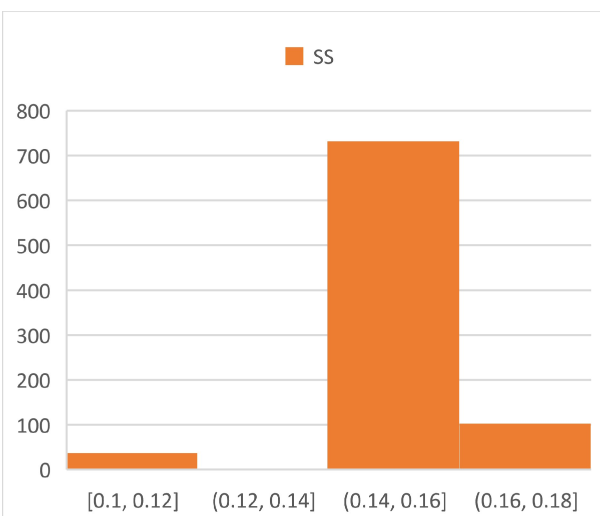


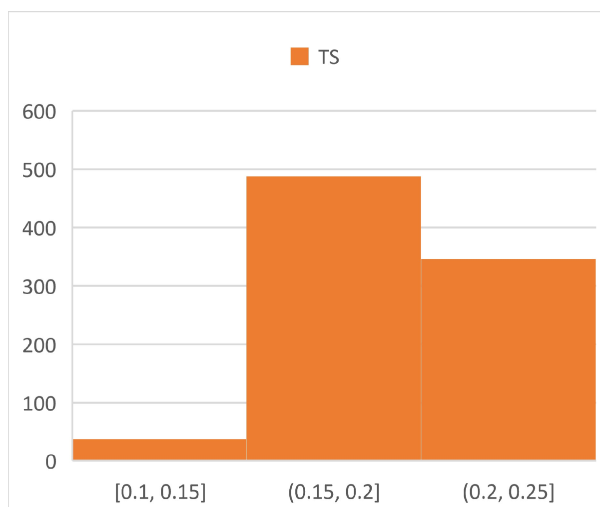

BM
400
350
300
250
200
150
100
50
0
[12.350,...
(212.350,...
(412.350,...
(612.350,...
(812.350,...
(1012.350,...
(1212.350,...
(1412.350,...
(1612.350,...
(1812.350,...

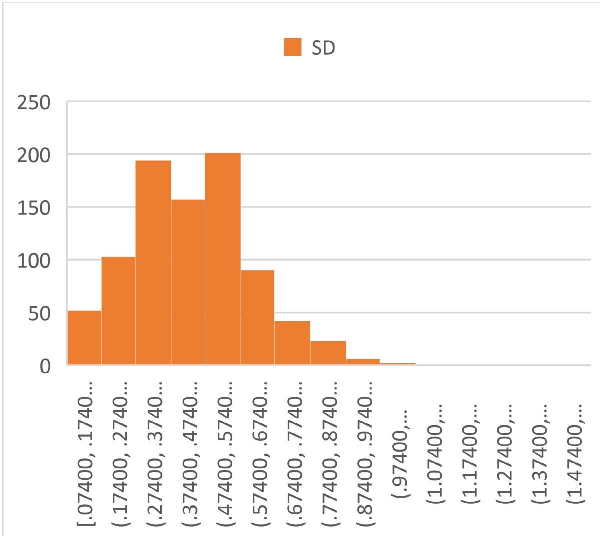

VQ
600
500
400
300
200
100
0
[.000, .5000]
(.5000, 1.000]

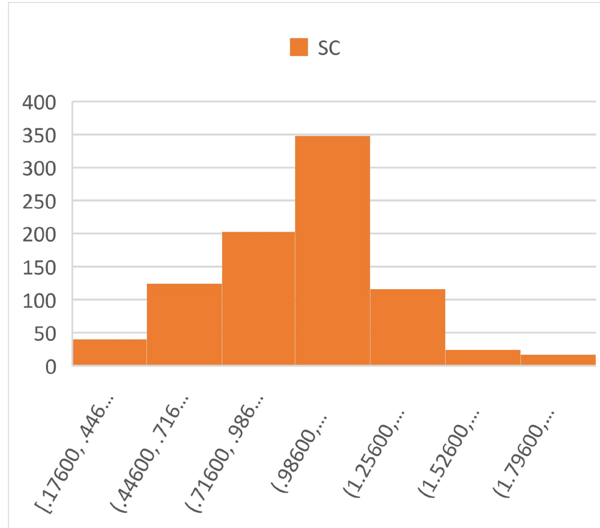

CR
600
500
400
300
200
100
0
[.5000, .7000]
(.7000, .9000]
(.9000,...
(1.1000,...
(1.3000,...

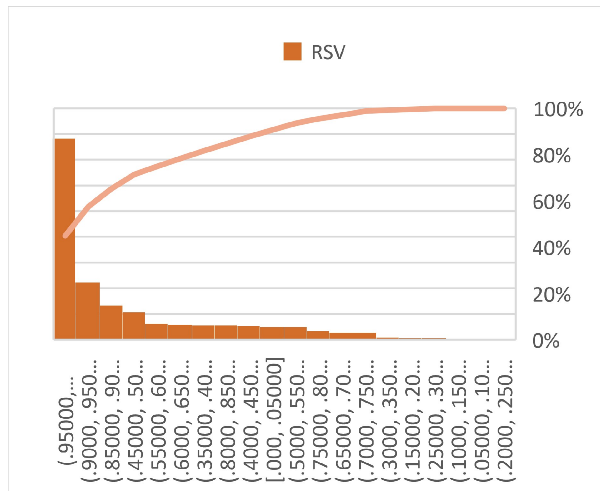

SW
600
500
400
300
200
100
0
[10, 11]
(11, 12]
(12, 13]
(13, 14]
(14, 15]
(15, 16]
(16, 17]
(17, 18]
(18, 19]
(19, 20]
(20, 21]
(21, 22]
(22, 23]
(23, 24]
(24, 25]


SS
800
700
600
500
400
300
200
100
0
[0.1, 0.12]
(0.12, 0.14]
(0.14, 0.16]
(0.16, 0.18]


TS
600
500
400
300
200
100
0
[0.1, 0.15]
(0.15, 0.2]
(0.2, 0.25]







Basic statistical techniques are essential tools for data summarization, analysis, and interpretation. These techniques provide a basis for decrypting data and making intelligent decisions. Mean, Standard Deviation, mode, Variance, and range are some of the basic statistical techniques most commonly applied. This study involved an initial examination of the overall data distribution using basic statistical analysis tests Friedman, J. H. (2001).

When the standard deviations, kurtosis, and diameter of the data were evaluated, it was observed that there were outliers. Therefore, these values have been changed to average values if possible. Moreover, comparisons of the basic statistical values of the test data and training data are similar, thus concluding that the test data is representative of the overall dataset. The data is a unique and valuable data set as it includes damage assessment data and detailed static analysis information (RVS). The sharing of data is kept confidential in accordance with Law No. 6698. If requested, it can only be requested from the relevant public institutions. Since categorical data is nominal and numerical data is discrete, it is suitable for regression and correlation analysis methods. For the stated reason, the homogeneity of the data and their relationship with each other will be evaluated with these methods in the following sections. After examining the distribution of the datasets, the number of possible outcomes of the RS variable leading to the most accurate predictions was investigated. For this purpose, Euclidean distances were plotted against data points (i.e., dendrogram). The least possible number of outcomes can be determined by counting the fewest points that intersect with any possible horizontal line drawn on the dendrogram. For the risk status variable, this number was equal to two or three (Figure 3.1). Therefore, this cross-check also confirmed the validity of two or three selected possible outcomes of the RS variable (i.e., risky or non-risky or medium risky). In other words, the created machine learning network has high success rates that can distinguish not only risky and non-risky buildings, but also moderate risk buildings.

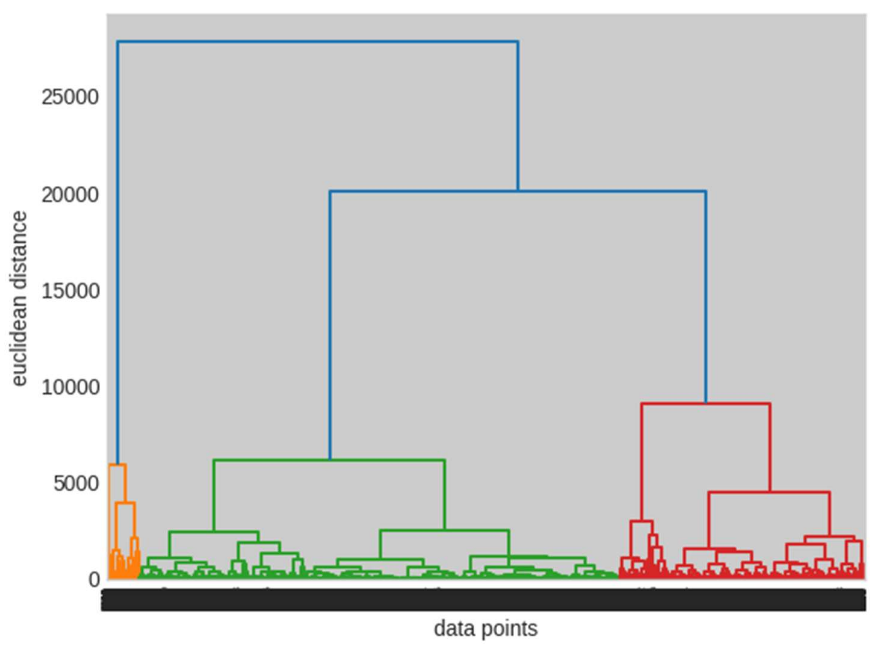


Figure 3.1. Dendrogram representing the least number of outcomes for the variable RS

After the dendogram method, another method was used to confirm the most optimal layer of the risk class. With the proposed Elbow method, the optimum “k” value is found and clustering is done with the k-means algorithm, algorithms such as this dendogram method can give us preliminary information about how many different results the data set can produce. Using this information, we can figure out how many outcomes the algorithms should divide our dataset into (such as risk-free or risky, medium risk, no risk).

The elbow technique is used to analyze how much variation is explained based on the number of clusters. It works on the tenet that the optimal number of clusters should be chosen so as to avoid having more clusters add noticeably to the data modeling or to the explanation of extra variation. The technique helps in determining an optimal number of clusters for clustering algorithms. Plotted against the number of clusters is the proportion of variation explained by the clusters. There will be a lot of information in the early clusters, but eventually the marginal gain will drop noticeably and the chart will show a different aspect (Bholowalia & Kumar, 2014). It was determined that the data set would be accurate using the bracket method based on 3 clusters, as shown in Figure 3.2.

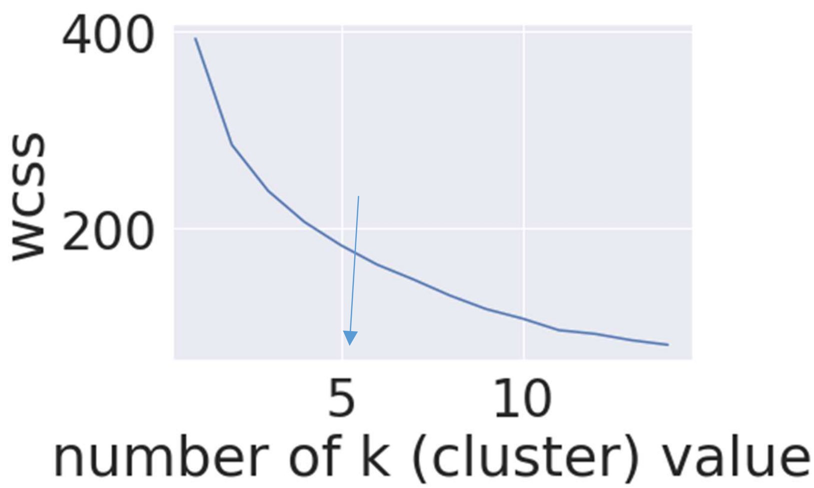


Figure 3.2. Elbow rule for classification of buildings

Finally, a second validation was done with K-Means Clustering, which is another method of clustering the risk class. Geon recommended K-Means, one of the most popular partitioning-based clustering techniques (Park et al., 2013). The cluster head selection process is relatively straightforward and quick. Initially, k out of n nodes are randomly categorized as either Risky, No Risk, or Moderate Risk. Subsequently, each of the remaining nodes determines its classification by selecting the nearest one among them based on the Euclidean distance. Consequently, the most suitable number of clusters for the data set is determined. As a result of the study with the data set, it is concluded that the separation into two clusters, namely risk-free/risky, is more manageable and a high success rate can be achieved (Figure 3.3).

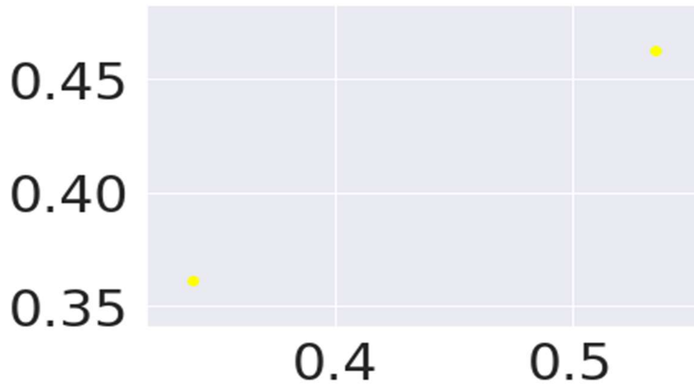


Figure 3.3. K-Means clustering method for classification of buildings

In summary, the methods concluded different outcome options. For instance, the dendrogram method, the Elbow Rule and the K-Means Cluster Method suggested 3 classes, 3 classes and 2 classes for the optimum classification of the Ve/Vr outcome parameter in the data set, respectively (Figure 3.2 - 3.4). Therefore, it has been determined that the V_e/V_r ratio for the algorithms to be used in the analysis should be arranged by taking the most optimum (for the most successful results) 3 classes or 2 classes (Figure 3.4).

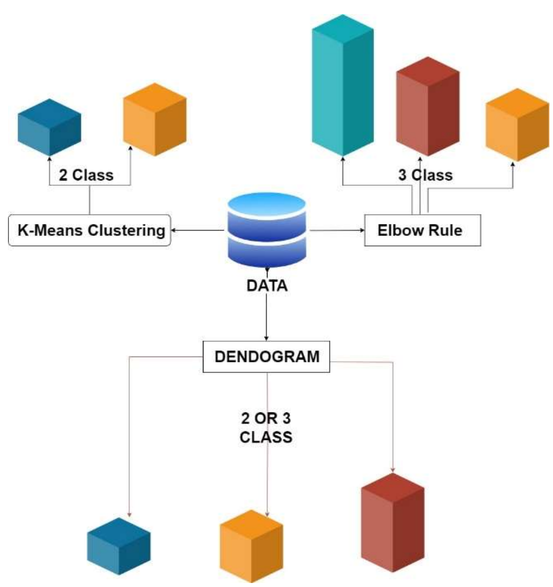


Figure 3.4. Optimum classification of V_e/V_r variable by means of Dendogram, Elbow rule and K-means clustering methods

The code used to determine the outcomes of this section is presented in Appendix 1.

3.1.2. Synthetic ninority oversampling technique (SMOTE)

In machine learning, class imbalance is a common problem, especially in classification tasks when one class significantly outnumbers the other.

In the context of this study, the dataset exhibited a significant imbalance, with 4,000 "Risky" structures and only 356 "Not Risky" buildings. This skewed distribution is visualized in Figure 3.5. Addressing this class imbalance is crucial to ensure that the model can effectively identify and predict the minority class (e.g., "Not Risky" buildings) and not be biased toward the majority class (e.g., "Risky" structures). Various techniques, such as oversampling, undersampling, or using appropriate evaluation metrics, are often employed to tackle this issue and improve the model's performance on imbalanced datasets.

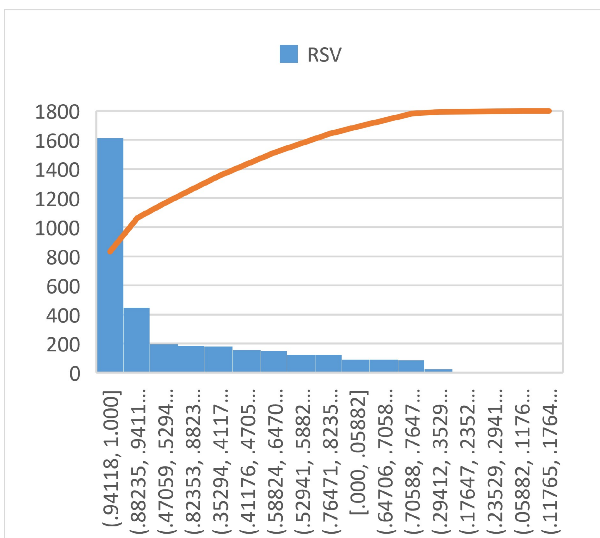


Figure 3.5. Showing balanced distribution of RVS values

By producing synthetic samples, SMOTE is a commonly used technique for oversampling a minority class and producing a more balanced dataset for model training. Choosing comparable instances in the feature space, drawing a line between them, and then drawing a new example along this line is how the technique operates. This technique enhanced the model's performance and helped balance the classes, particularly in correctly forecasting the minority class. It is crucial to remember that, although though SMOTE worked well for this study, there are a lot of different methods that can be used to correct class imbalance (Beyhan, 2023).

Alternatively, addressing imbalanced datasets involves oversampling the minority class. A basic approach is to duplicate instances from the minority class, but these copies don't offer any new insights to the model. Instead, a more advanced technique is to generate new examples by synthesizing from the existing ones, as outlined in (Brownlee 2020). To tackle this issue, an alternative to traditional random oversampling was introduced by Chawla et al. (2002).It's known as the Synthetic Minority Oversampling Technique (SMOTE), and it serves as a form of data augmentation for the minority class. The core idea behind SMOTE is to perform interpolation among neighboring instances in the minority class. This allows for the creation of new minority class examples within the proximity of existing ones, thus aiding classifiers in enhancing their generalization capabilities.

Upon a closer examination of the detailed analysis results within the dataset, it has been observed that there are imbalances in the ratio of risky wall shear force to the total wall shear force (RVS) across the entire floor. In other words, it is seen that the number of buildings with low-risk levels is unbalanced (Figure 3.4). For this reason, by synthetically increasing the number of low sismic risk buildings in the data set with the SMOTE method, the unbalanced data set was improved and the problems such as overfitting-low accuracy were eliminated.

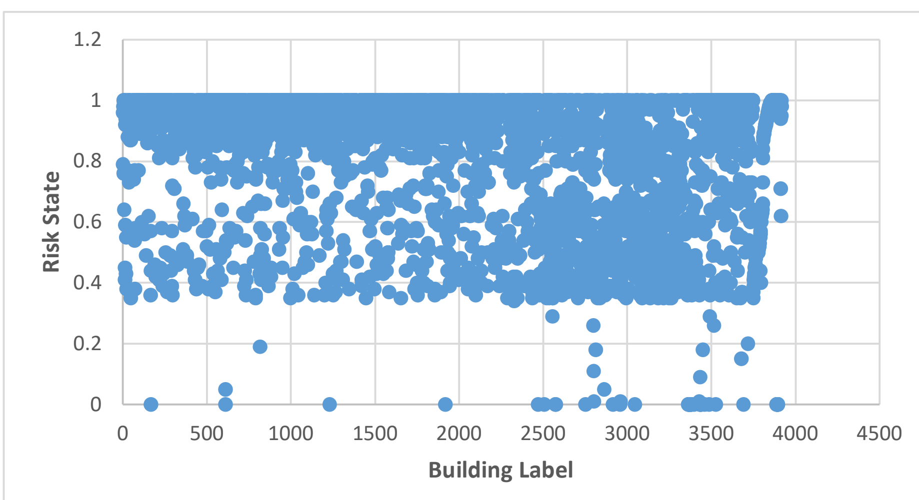


Figure 3.6. RVS scatter plot current dataset

With the SMOTE method, the number of buildings with low seismic risk in the training dataset has increased. As can be seen from the graph (Figure 3.6), this increase has resulted in a more balanced view of the building risk distribution. As mentioned before, this method prevented overfitting in the data set and increased success rates.

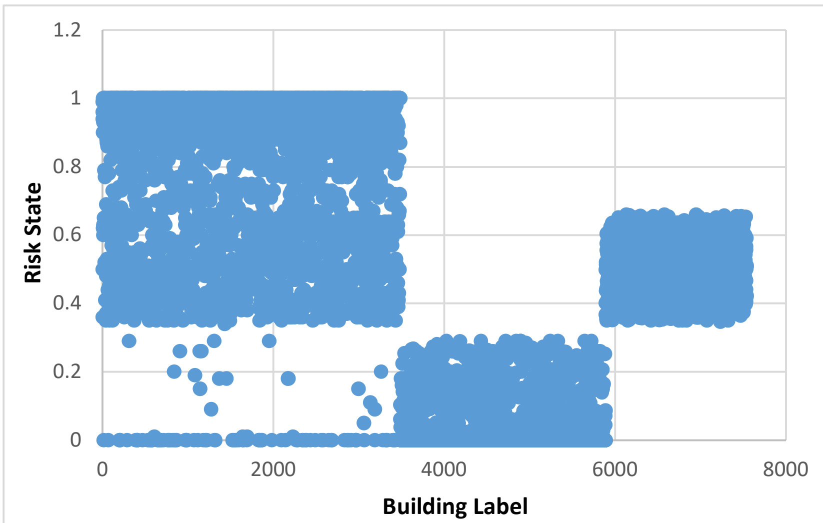


Figure 3.7. RVS scatterplots reproduced low-risk structures with the SMOTE method.

This synthetic replication process replicates the less risky structures in the training dataset after the dataset is separated as train/test. Therefore, there is no change in the number of data in the test data set. Following are the results of the distribution of risk-free, medium-risk, risky, and high-risk structures as a result of the four classification of the Ve/Vr variable in the training dataset and their synthetic reproduction with the SMOTE method.

- Before Undersampling, counts of label '1': 505
- Before Undersampling, counts of label '0': 103
- Before Undersampling, counts of label '2': 484
- Before Undersampling, counts of label '3': 2392
- ✓ After Undersampling, counts of label '0': 2392
- ✓ After Undersampling, counts of label '1': 2392
- ✓ After Undersampling, counts of label '2': 2392
- ✓ After Undersampling, counts of label '2': 2392

The code used to determine the outcomes of this section is presented in Appendix 1.

3.1.3. Data scaling

Normalization is utilized when algorithms operate independently of specific assumptions about the distribution of the data, whereas standardization is employed when algorithms make specific assumptions about the characteristics of the data distribution. The majority of Machine learning algorithms tend to underperform when dealing with numerical features that exhibit inconsistent scaling (Xue et al., 2019). The Min-Max scaler and the Standard scaler are two often used methods to address this problem (Kaur, 2020). The Min-Max scaler creates a range (typically 0-1) by zooming in or out, then scales numerical features into the specified range. The Standard-scaler, on the other hand, transforms numerical information in a way that produces a distribution with a mean value and a standard deviation of 0 and 1, respectively (Luo et al. 2020). In this study, considering normalization, the parameters in the data set were drawn to the same measure level by using the Standard-scaler. The standard-scaler normalization method ends in Equation 3.1 below. Where μ is the mean value and σ is the standard deviation.

$$z = \frac{(x - \mu)}{\sigma} \qquad (3.1)$$

In this part of the study, all parameters in the data were brought to the same scale and analyzes were performed in order to obtain accurate results.

3.1.4. Variable selection with dimension reduction

Dimensionality reduction involves the task of decreasing the quantity of attributes in a dataset while retaining as much of the variability present in the original dataset as feasible. In essence, it's a method for converting a large set of variables into a smaller set without sacrificing critical information. (Swarna et al. 2020). The section's major goal is to examine which dimensional reduction techniques are used in fast seismic risk estimation analysis. Here are a few advantages of using dimensionality reduction techniques on a dataset (Velliangiri et al. 2019).

1. Data storage can be cut down when the number of dimensions decreases.
2. It merely requires a shorter computation time.
3. It is possible to eliminate noisy, redundant, and irrelevant data.

4. It is possible to enhance data quality.
5. On more dimensions considered, some algorithms do not function effectively. Therefore, lowering these dimensions makes an algorithm more accurate and in creases its efficiency.
6. It streamlines the classification process and boosts productivity.

If the parameters in the dataset have little or an adverse effect on the risk outcome (for example, the building becomes more non-risky as the age of the building increases), these parameters may prevent the algorithm from making the correct estimation of the outcome. In addition, the presence of highly correlated parameters or the removal of parameters with the same effects (eg soil class and spectral acceleration coefficient) before analysis can increase the computation speed of the algorithm. In the context of feature selection, a high-dimensional dataset often contains a substantial number of attributes, some of which may be inaccurate, outliers, or redundant. This circumstance increases the search space's dimensionality and may make the dataset less suitable for learning. As a result, from the original dataset, a subset of the most pertinent features must be extracted. Computational principles guide the feature selection techniques that are used to select the most relevant features from the original set. Filtering, wrapping, and embedding are the three methods that these actions can be carried out. The use of feature rank in filter methods is the norm for feature collecting by arrangement (Velliangiri et al. 2019). It chooses only pertinent features, increasing the correlation between them in the original feature collection Figure 3.8.

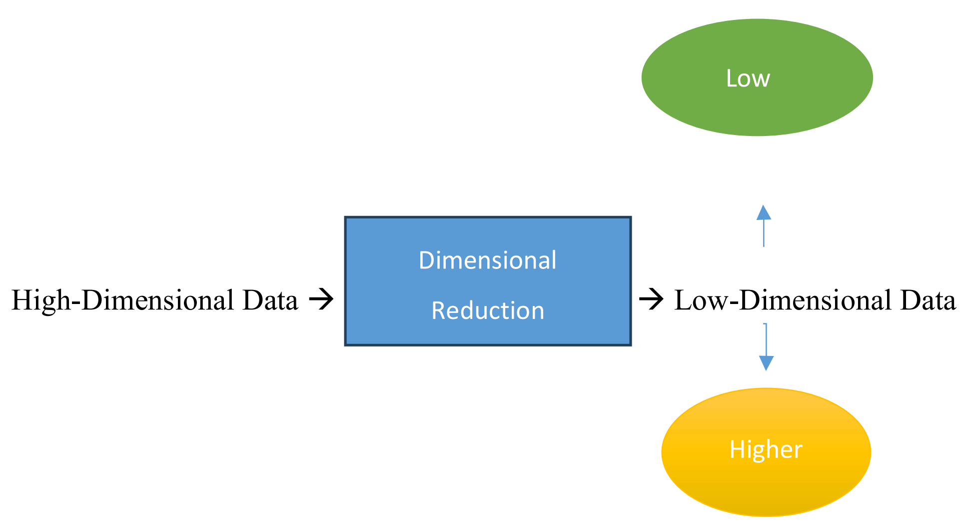


Figure 3.8. Dimension reduction advantages

In this study, the initial dataset consisted of twelve parameters. To enhance the analysis efficiency, five size reduction methods were employed to identify the six most influential parameters. Initially, a random forest regression algorithm was used to sort the parameters based on their feature importance, leading to the removal of the least effective ones.

Next, in the second and third steps, advanced feature selection and retrospective feature elimination techniques were utilized to eliminate structural crack and mortar quality from the parameter list. This decision was made because the shear strength and diagonal tensile strength parameters had weak effects on the seismic risk analysis results.

In the fourth step, the design spectral acceleration coefficient, which exhibited high correlation and similar effects on the results, was removed at 1 second. However, the short-term (0.2 seconds) design spectral acceleration coefficient was retained for evaluation in the machine learning process.

Finally, the building mass parameter was excluded from consideration due to its challenging detection for both technical personnel and automated rapid detection processes. This approach aimed to dissociate unnecessary parameters from the network and minimize bias (as depicted in Figure 3.9).

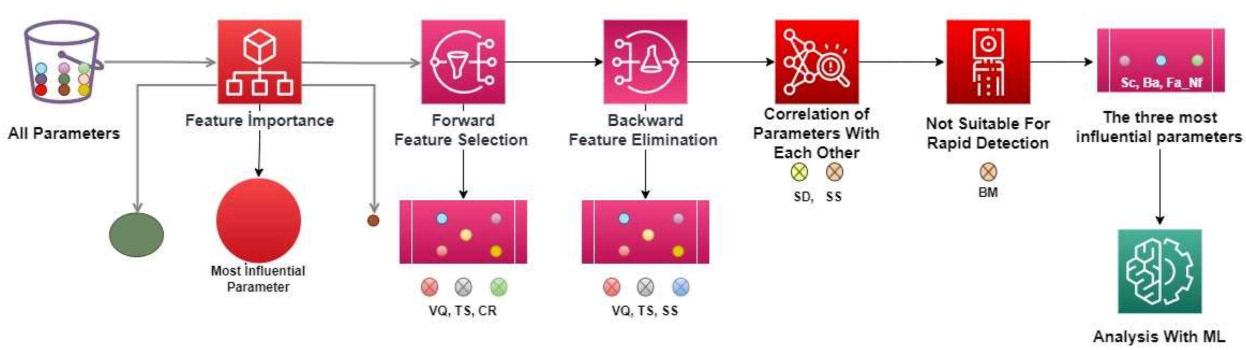


Figure 3.9. Dimension reduction methods used in the study

3.1.5. Feature importance

Many variable selection procedures rely on collaborative sequencing and model estimation of variable importance to create, evaluate, and compare a family of models. Generally, three types of variable selection methods are distinguished: the “filter”, in which the variable importance score is not dependent on a particular model design method; “wrapper,” that calculates scores using projected performance; and lastly, "embedded", which more closely integrates variable selection and model estimation. In the study, the "filter" model based on variable importance will be used (Kohavi and John 1997). The importance coefficients of the data generated in this step were determined according to the seismic risk analysis result with a random forest regressor (Breiman 2001) (Figure 3.10). These importance coefficients are used in algorithm analysis, and parameter selections are decided for variable selection.

The features' relative importance that affect the RVS value with the Random Forest Classifier algorithm is given in Figure 3.10. According to this graph, the design earthquake acceleration coming to the building, which is thought to be the most effective for seismic detailed analysis, the number of building floors that affect the belief period, and the age of the building, which indicates the technology in which the building is made, were determined as the three most effective parameters (SC, NF and BA).

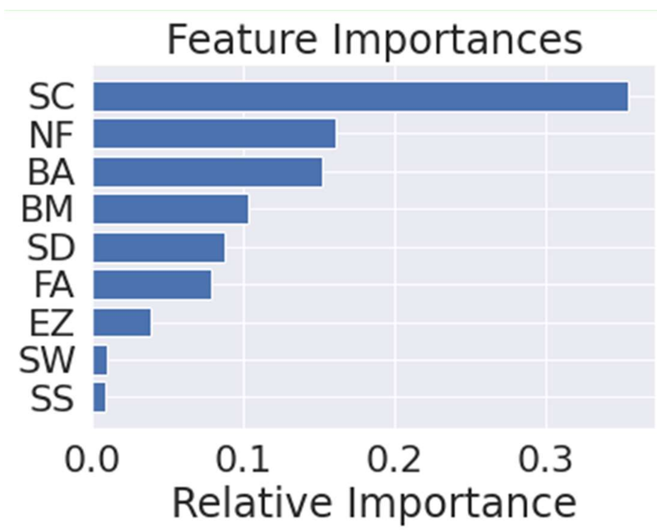


Figure 3.10. Relative importance method

3.1.6. The backward feature elimination and forward feature selection methods

In this section, unnecessary parameters were removed from the algorithms by using the Backward Feature Elimination and Forward Feature Selection methods, which are one of the Dimensionality Reduction techniques, and the parameters that affect the risk result.

To narrow down your focus and implement the 'Backward Feature Elimination' method, follow the steps outlined (Figure 3.11): (Santos et al., 2011)

- The dataset's current features had to be collected, then used for the testing model.
- Calculating the level of model performance was necessary.
- After calculating the model's output and removing one variable at a time—that is, one function would be removed in times the model would then be tested on the remaining n-1 variables.
- Choose the variable whose deletion resulted in the least (or no) difference in the output of the model, and then eliminate that feature.
- Keep repeating the previous technique until it becomes difficult for the variable to disappear.

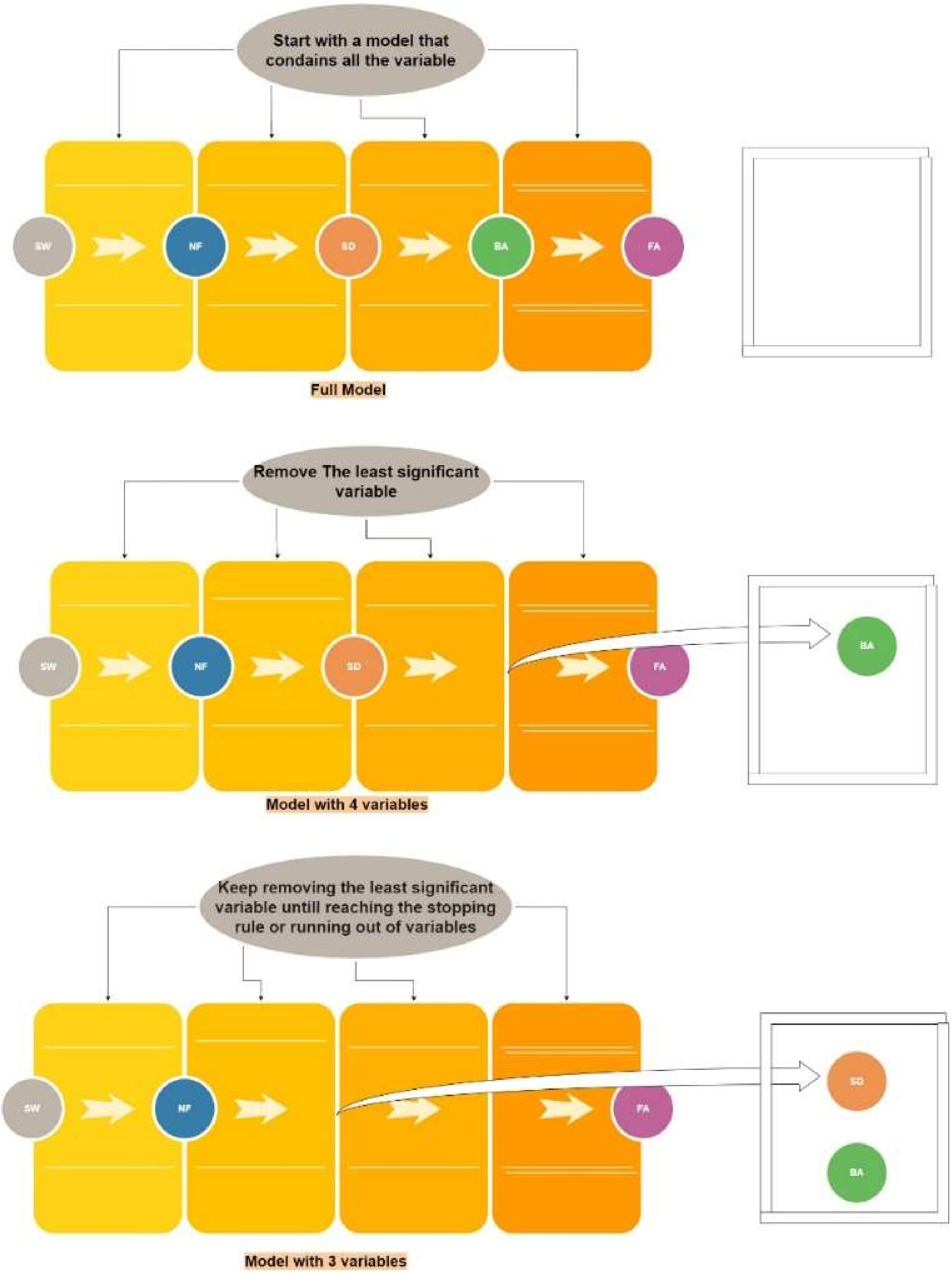


Figure 3.11. The backward feature elimination method

To focus more and follow the 'Backward Feature Elimination' method, follow the coming steps (Figure 3.12): (Macedo et al., 2019)

- Analyzes techniques that ignore complementarity against a baseline. Distribution environment Current research allows established techniques to interpret the objective function as approximations of a target objective function.
- How, from a theoretical point of view, the useful properties of the objective function are affected by different types of approaches and the drawbacks of the selected feature selection techniques are pointed out. Methods to avoid and methods that currently work best are outlined.
- After adding each function (n times) and calculating the output of the model, the model will be evaluated with all the additional variables, so each variable will be added one by one.
- Finally, the most effective parameters will be selected, and the ineffective ones will be eliminated.

Indeed, by applying the backward feature elimination method, not only does the algorithm's speed increase, but its success rate also improves. In this method, the initial step involves building a model using all the features simultaneously to assess its performance. Subsequently, the method systematically removes each variable one by one and evaluates the model's performance after each elimination. The variable that leads to the worst performance is eliminated from the final feature set.

On the other hand, the forward feature selection method operates in the opposite direction. It begins with an empty feature set and iteratively adds one parameter at a time, evaluating the model's performance after each addition. The parameter that contributes the most to enhancing the results is selected and retained in the final feature set.

Fundamentally, backward feature elimination focuses on removing the least important features, while forward feature selection systematically incorporates the most significant parameters to manage and enhance algorithm outcomes. Both of these techniques are pivotal in the context of feature selection and model optimization during data analysis.

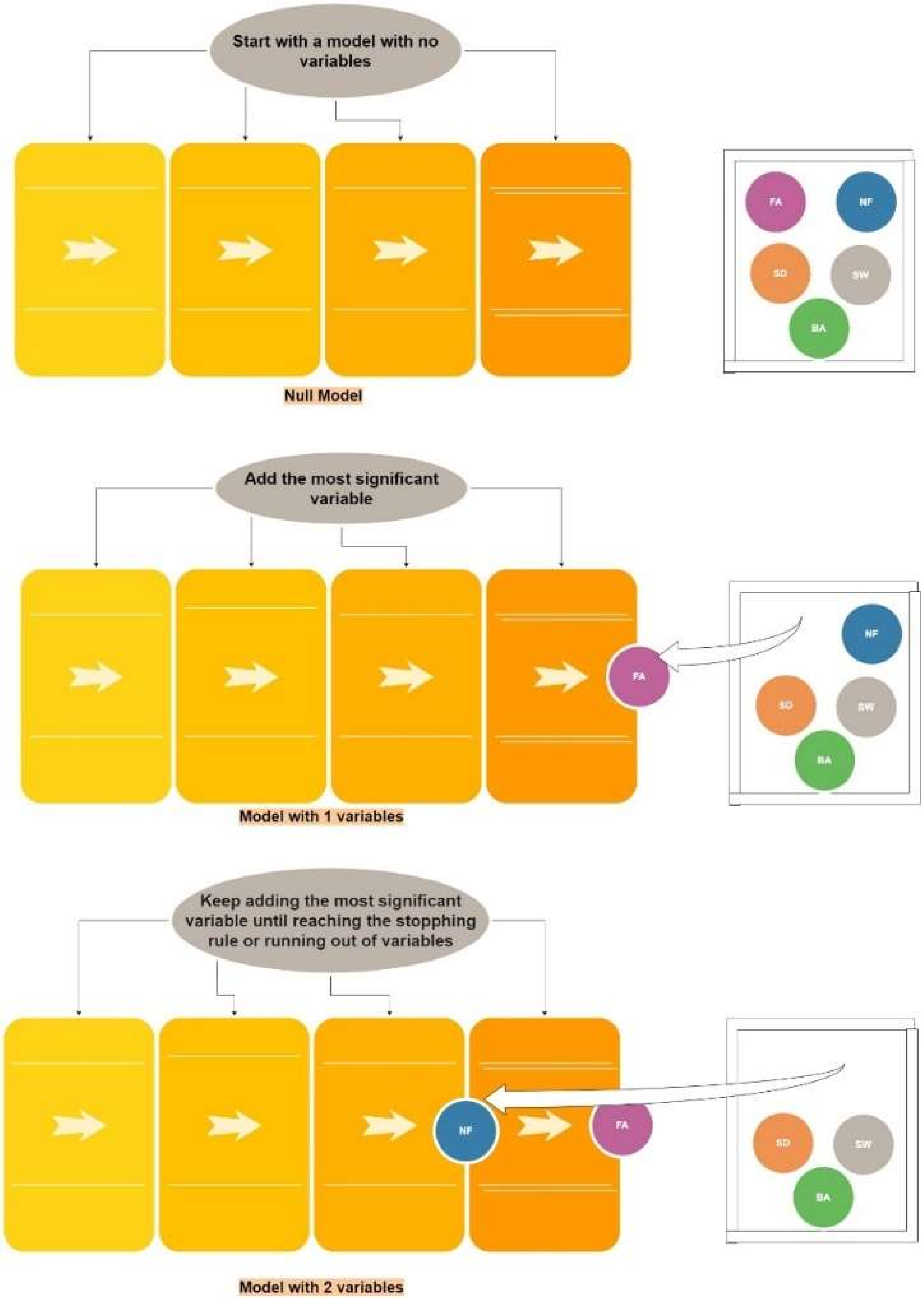


Figure 3.12. Forward feature selection methods

3.1.7. Correlation of parameters with detailed assessment analysis results

A comprehensive seismic assessment analysis of buildings is of utmost importance in understanding how a building behaves during seismic events. Typically, this detailed seismic assessment is mandated for the safety evaluation of all structures older than 20 years, particularly those situated in close proximity to earthquake-prone regions. Conversely, in countries with the objective of quickly identifying risky buildings for potential retrofitting, rapid screening methods should be employed. These methods are designed to swiftly filter and assess buildings, facilitating prompt action to address retrofitting needs. Therefore, this filtering operation is vital as structures that are at risk of collapse even in a small-scale earthquake (called a service earthquake) should be strengthened immediately, or new earthquake-resistant structures should be built instead of these structures. However, current rapid screening methods in the literature have a very limited correlation with the detailed assessment analysis results, as none of the methods was calibrated with the detailed analysis results. Thus, it is difficult to depend on the risk estimations of the rapid screening methods while taking actions at the seismic risk mitigation level. (Coskun & Aldemir, 2022) As a result, the primary aim of this study was to establish a network that can link seismic risk with building characteristics. Therefore, it was of great importance to determine the most effective variables to distinguish between risky and risk-free structures in line with the seismic risk assessments obtained from detailed analyses. This process can help eliminate unnecessary parameters from the network and reduce bias. Here, a very high correlation of one variable with another variable should also be excluded to improve algorithm performance. Within the scope of the study, the most effective parameter selection was made by examining the correlation maps and correlation graphs of each parameter with each other. The information provided shows the correlation between the variables in the dataset and the class label called “risk status” obtained from detailed analysis. (RS). (Table 3.6).

Table 3.6. Basic correlation analysis

TS	**RS**	**SS**	**RS**	**CR**	**RS**
0.18	0.867	0.15	0.837	0.15	0.837
0.1	0.787	0.10	0.787	0.10	0.786
0.25	0.748	0.18	0.70	0.189	0.700
0.20	0.70				
VQ	**RS**	**VD**	**RS**	**NF**	**RS**
1	0.865	1	1.000	7	1.000
0	0.787	0	0.727	4	0.936
				3	0.922
EZ	**RS**	**SW**	**RS**	5	0.912
4	0.864	18	0.868	6	0.818
2	0.837	25	0.765	2	0.750
3	0.811	13	0.775	1	0,680
1	0.788	10	0.729		
0	0.600	15	0.700		

In Table 3.6, values above 0.5 contribute to the building's risk, while values below 0.5 contribute to the building's risklessness. For example, if the Soil Zone Type is soil with loose sand, gravel and hard clay layers (EZ=0.864), the RS value increases to 0.864, indicating that it contributes to the risky nature of the building. If the Soil Zone Type is Solid Ground and hard rocky ground, the RS value is reduced by 0.600, contributing less to the building's risk. It can be said that the result obtained according to the structures in the data set is reasonable, since the soil with loose sand, gravel and hard clay layers affects the buildings negatively.

Furthermore, the relationships among various parameters, including the spectral acceleration coefficient (EDA Multivariate Analysis), building age, number of floors, typical plan area, and other factors, are illustrated in Figure 3.13. It's evident that the risk level exhibits a strong association with an increased likelihood, spectral acceleration coefficient, and building age. On the other hand, the correlation between floor area and the risk level, as determined by the detailed analysis, is relatively limited. Among these parameters, Short Period Spectral Acceleration (SC), Building Age (BA), and the number of floors (NF) exhibit the highest correlations with the results of the detailed seismic evaluation analysis. This correlation is theoretically sound. Older structures that have not undergone engineering upgrades are more prone to seismic risks. Similarly, as the number of floors and the spectral acceleration coefficient of a building increase, the horizontal displacement requirements and earthquake forces acting on the building naturally escalate, thereby increasing the seismic risk. (Coskun and Aldemir, 2022)

Figure 3.13. Correlation head maps with the risk situation (EDA Multivariate Analysis)

The correlation of the parameters with each other also has an important effect on the parameter elimination. Since parameters with high correlation with each other will affect the result at the same rate, removing one of the parameters will affect the algorithm in a good way. In this context, there is a graph showing the relations of the parameters in Figure 3.14. As it can be understood from Figure 3.16, a direct correlation of the SC parameter with the SD parameter and the BM parameter with FA and NF parameters has been determined. Graphs showing the individual relations of these parameters with each other will be shown in the following sections.

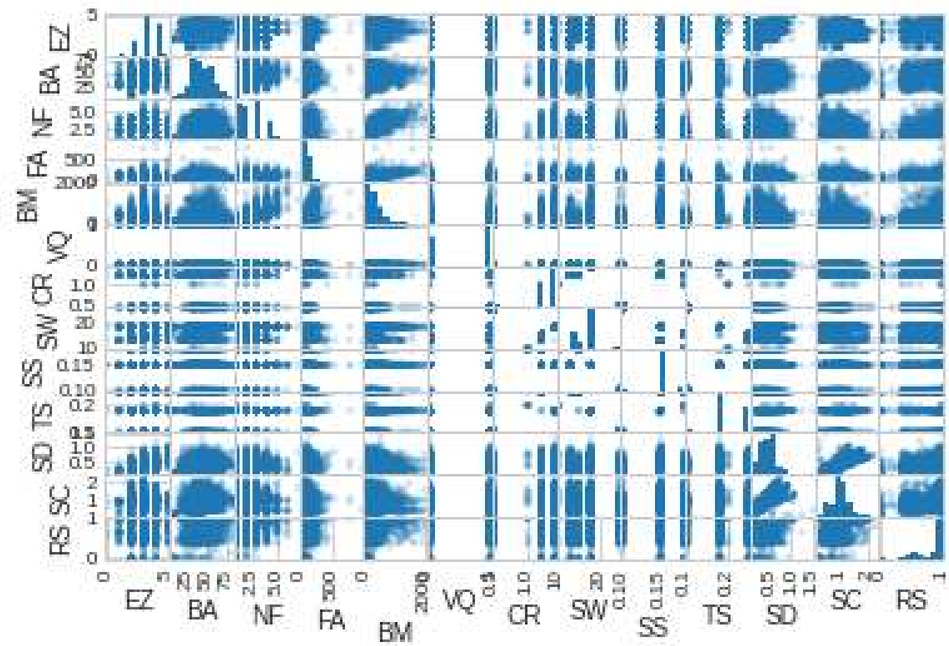


Figure 3.14. All parameters correlation of graphich

In addition, the distribution of some parameters to the risk result also gives information about the correlation. For example, in Figure 3.15, the effects of bad and good crack-and-mortar conditions on the risk outcome are the same. In reality, we conclude that the construction must be risky if the mortar quality is poor and cracks are present. However, since there is an equal distribution here, it is concluded that this parameter does not consist of real values.

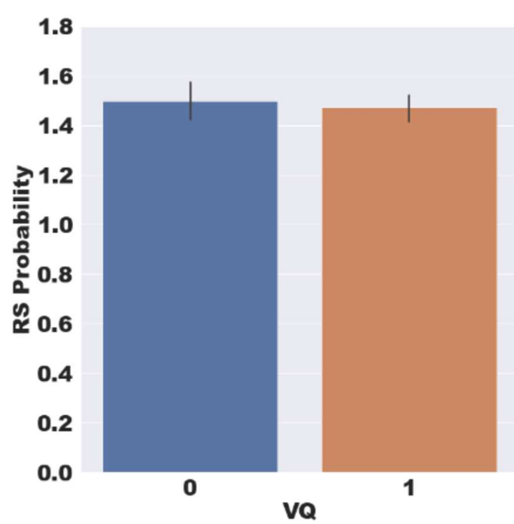


Figure 3.15. Effect of visual quality on risk outcome

In addition, parameters with high correlation with each other were chosen. For example, since the short spectral acceleration coefficient and the 2 s spectral acceleration coefficient have a very high correlation with each other, only one of these parameters was chosen. In another example, only one of these parameters was chosen because the linear correlation between BM parameter and the NF parameter was high. (Figure 3.16)

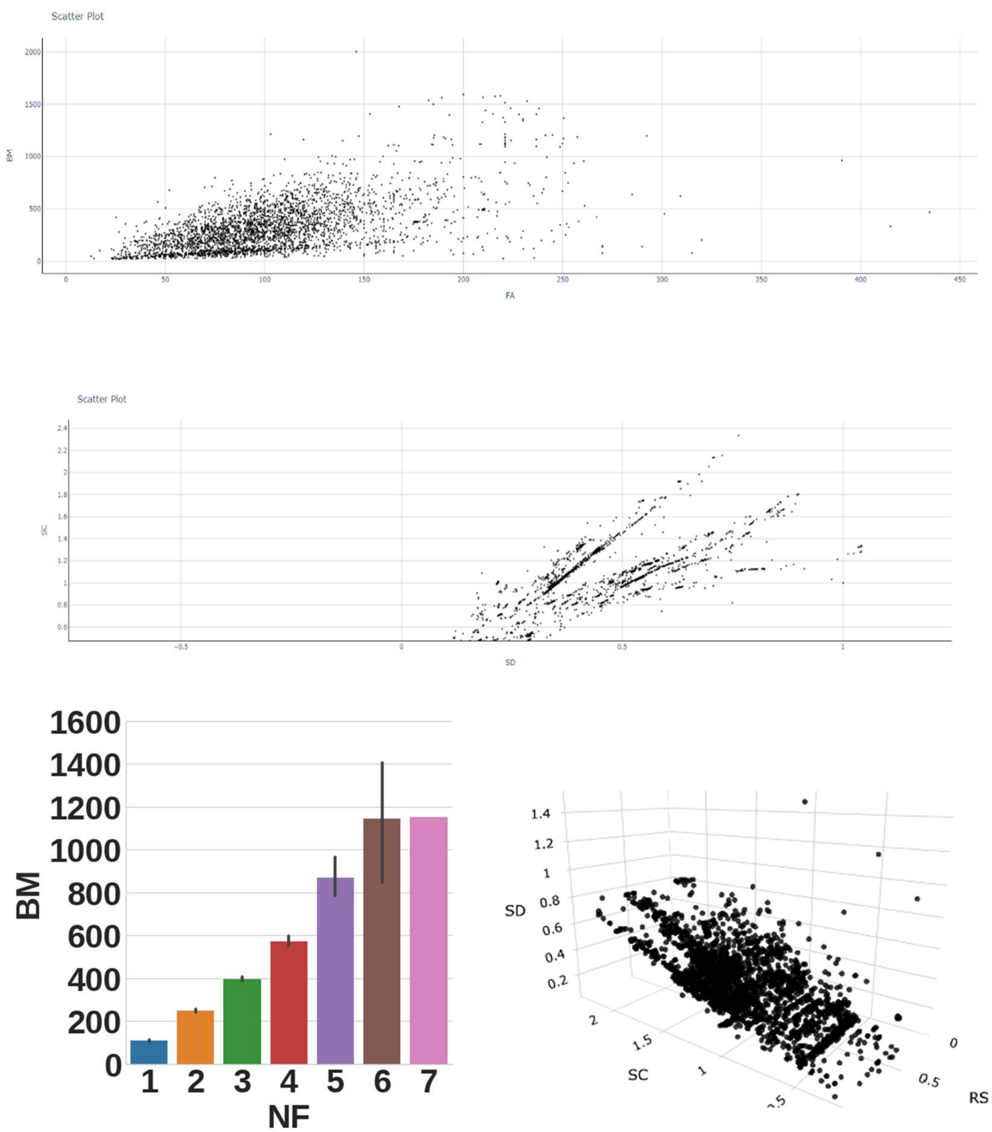


Figure 3.16. Parameters that are highly correlated with each other

3.1.8. Not suitable for rapid detection

In this study, a multi-step approach was employed to reduce the size of the data set by selecting parameters that have a significant impact on seismic analysis results and are easily detectable. Initially, the Backward Feature Elimination and Forward Feature Selection methods were utilized to prioritize and eliminate parameters based on their importance. Additionally, parameters that posed challenges for detection in the field and could potentially slow down human or machine detection processes were also eliminated. Among these, the building mass parameter was removed due to the difficulty in accurately determining its value.. For a visual representation of this process, refer to Figure 3.17.

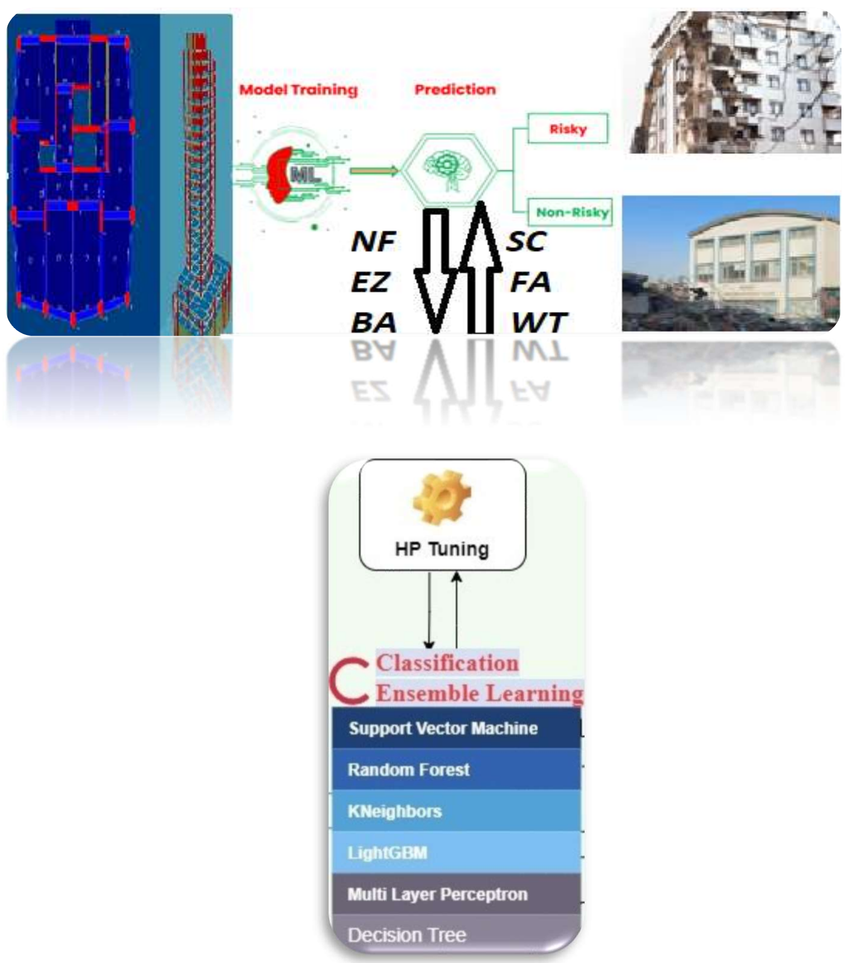


Figure 3.17. Algoritms with dimension reduction

3.1.9. Feature extraction and engineering

The Feature Extraction (FE) method involves generating new features from the original dataset. It proves highly advantageous when the objective is to reduce the computational resources necessary for processing while retaining relevant features from the dataset. FE is instrumental in creating meaningful transformations of the initial features, resulting in more significant and informative features. (Velliangiri et al., 2019). In Figure 3.15, it is observed that the floor area parameter ranks 6th in terms of importance when using the random forest regression algorithm. However, its correlation value in the correlation heat map is quite low, indicated by -0.01, which suggests a weak linear relationship with the target variable (RVS value). To address this issue and increase the correlation with the RVS value, a Feature extraction process was employed. The method involved processing a few parameters together, resulting in the creation of a new variable called "total floor area" (FA_NF). This new variable was obtained by multiplying the number of floors by the ground floor area. After applying this method, the correlation between the "total floor area" (FA_NF) and the RVS value significantly improved and became 0.21, which is ten times higher than the original correlation value. Figure 3.18 visualizes this increase in correlation. As a result of this feature extraction process, a parameter (total floor area) with a higher correlation to the RVS value was produced. This new parameter is expected to have a stronger influence on the seismic risk analysis results as it increases, thereby contributing to a more accurate assessment of the risk situation.

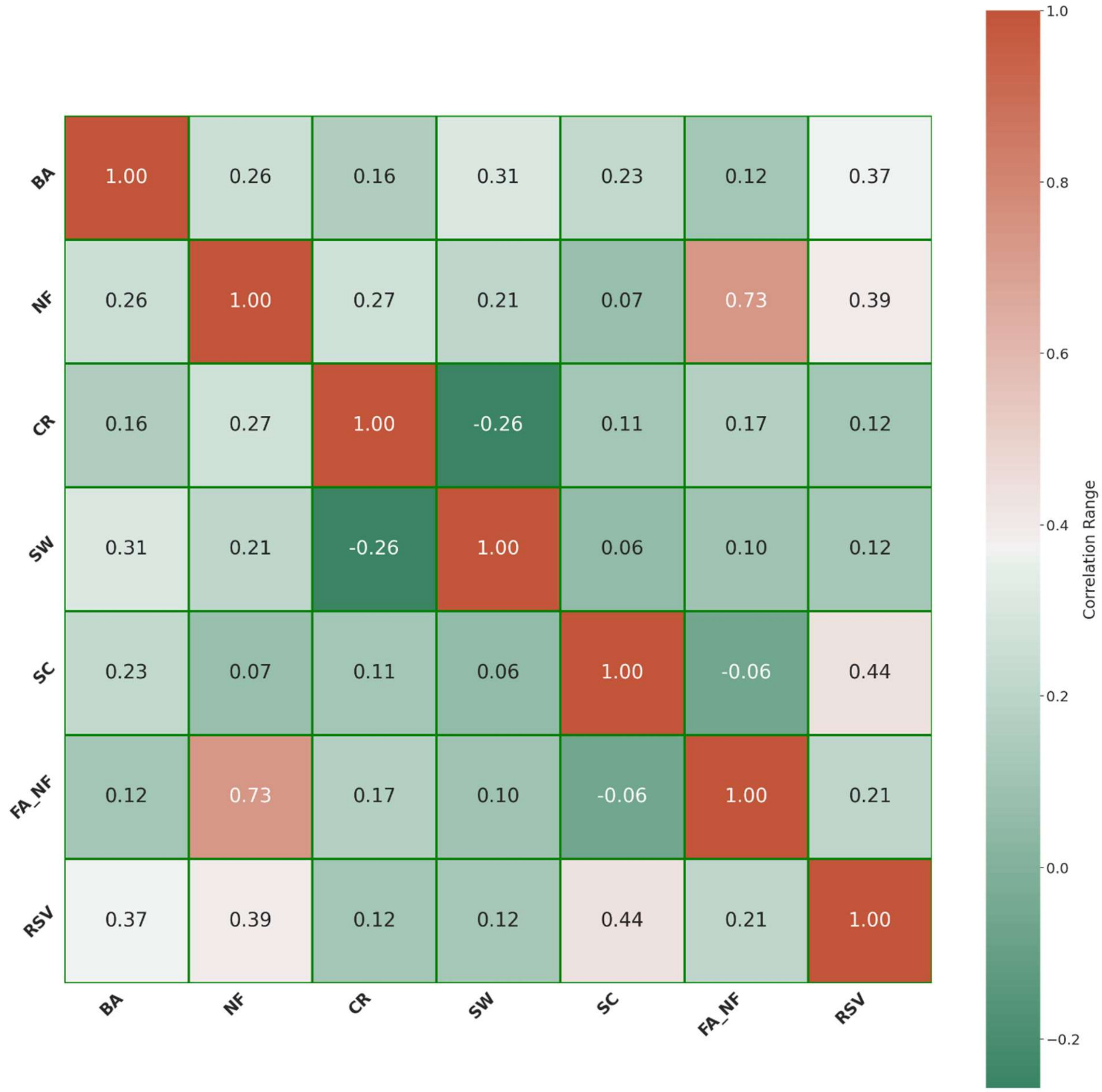


Figure 3.18. Correlation best parameters (EDA Multivariate Analysis)

3.1.10. Results of variable selection with dimension reduction

Choosing the most effective parameters with dimension reduction and feature selection methods contributes to both speeding up the algorithm and increasing the success rate. These parameters are Entire Floor Area (FA_NF), Short time (0.2s) design spectral acceleration coefficient (SC), Compressive Strength (CR), Specific Gravity (SW), Building Age (BA), and Number of Floors (Table 3.7). All these parameters are necessary both for fast detection and to not reduce the success rate.

Table 3.7. The most effective parameters

Variable	Abbreviation
Area of all floors	FA_NF
Building Age	BA
Specific Weight	SW
Compressive Strength	CR
The short period (0.2s) is the design spectral acceleration coefficient	SD
Number of Floors	NF

As a result of these studies in this section, twelve parameters obtained from a data set with detailed seismic analysis results, building photographs and building dimensions were reduced to six parameters with dimension reduction methods. Because this study propose without the need for technical personnel and without entering the building, with the automation methods of the structures, after the parameter selection, the estimations of the RVS values using machine learning methods can be made with high accuracy. Data pre-processing in this section is very important as the procedure will be applied to identify, inventory and sort the most vulnerable buildings that could be damaged by a possible earthquake in a given area with these six parameters.

3.2. Analysis with Machine Learning Algorithms

The core idea behind machine learning (ML) is the ability of data-driven models to learn about a system from observed data itself without requiring prior knowledge of the mechanical relationships that govern the system's behavior. With each new sample of data, machine learning algorithms can improve their performance adaptively, find relationships in complex, diverse, and high-dimensional data, and update their differentiable weights accordingly (Shaikhina et al., 2019). In this study, to achieve the highest success rate, a preference was given to utilizing multiple machine learning algorithms, a concept known as ensemble learning, as opposed to relying on a single algorithm. The ensemble learning approach encompassed a range of supervised machine learning algorithms that leverage labeled input data to acquire a function capable of producing precise outputs when confronted with unlabeled data (Figure 3.19a). To accomplish this, the dataset was initially partitioned into training, validation, and test datasets. Statistical measures, such as parameter correlations, were calculated, and feature engineering tasks were performed, including the definition of categorical variables. Following this, ensemble learning algorithms were applied to carry out the necessary learning procedures. Finally, the machine learning network's performance in predicting the risk state was assessed using the test dataset. (Figure 3.19b, Figure 3.19a).

In order to anticipate speedy seismic analysis findings, the study used a variety of machine learning techniques, such as Decision Tree Classifier, Gradient Boosting (GB), Logistic Regression (LR), Random Forest C., K-Mean Clustering, Support Vector Machine, LGBMClassifier, MLPClassifier, and MARS. Within the scope of the study, a method called the Unified Learning Framework is proposed in which the most effective machine learning algorithm is selected for each building class.

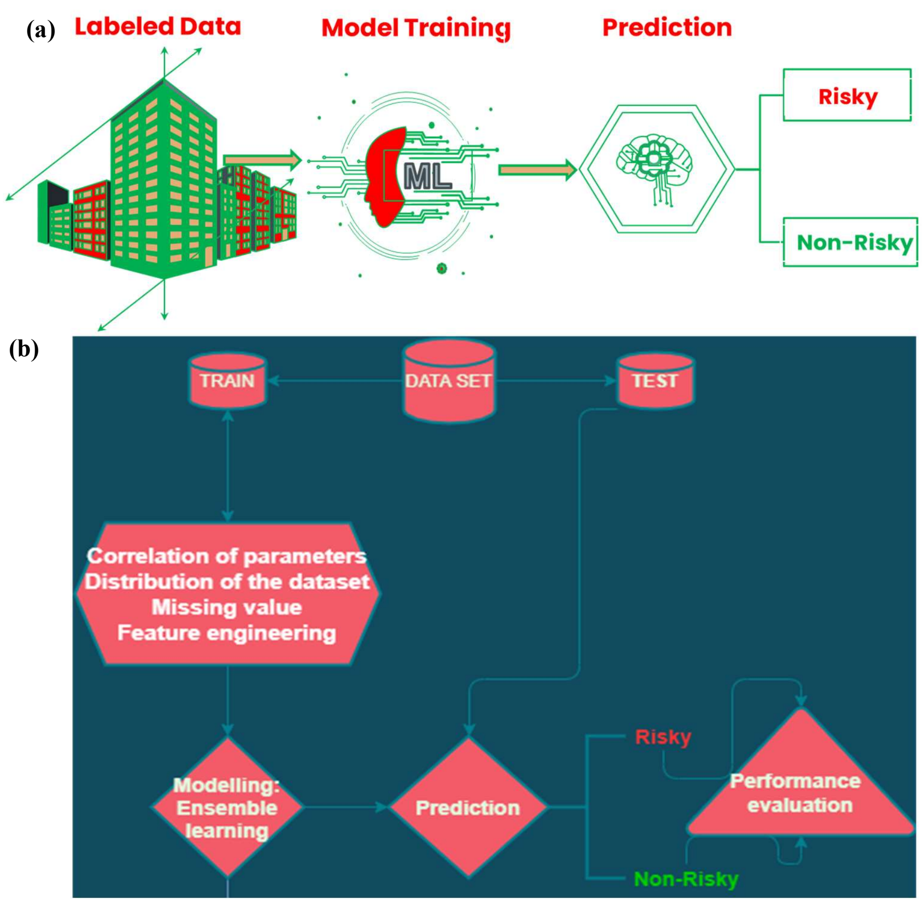


Figure 3.19. (a) Supervised learning example and (b) Flow chart of the method used in this study (Coskun and Aldemir, 2022)

3.2.1. Background on the machine learning methods

The working principles of the algorithms that form the basis of the ML techniques used in this thesis are explained in this section. Nine different ML techniques were applied to regression and classification to compare performances. In this section, the most accurate machine learning algorithms for estimating seismic risk are selected, and the underlying mathematics are explained.

3.2.1.1. Logistic regression

Understanding a binary or proportional response (dependent variable) based on one or more predictors is the aim of a logistic regression model (Hilbe, 2009). Logistic Regression offers the advantage of interpreting the output of the prediction function as a posterior probability. This property is achieved through the sigmoid function, as illustrated in (Equation 3.2):

$$P(Y = k|X = x) = \frac{\exp\,(\,\omega_{k0} + \sum_{i=1}^{n}\,\omega_{ki}x_i)}{1 + \sum_{k}^{K}\exp\,(\,\omega_{k0} + \sum_{i=1}^{n}\,\omega_{ki}x_i)} \qquad (3.2)$$

Logistic Regression (LR) was originally developed for binary classification tasks. However, it can be extended to solve multiclass problems using techniques such as adding a SoftMax layer. In this method, the classification of an example is usually created by considering the class to which the learning model assigns the highest probability (Tan 2022).

In Logistic Regression, for instance, the probability of being at risky is assumed to be the following functional form, which is a sigmoid (logistic) function (Equation 3.3). This value is always between 0-1.

$$The\ sigmoid\ function = 1/(1 + \mathrm{e}^{-\mathrm{z}}) \qquad (3.3)$$

In this study, binary logistic regression will be the focus. since it is more effective to use other algorithms for multiple predictions. It is aimed that the most risky buildings should be determined and new buildings resistant to earthquakes should be built, and this decision should be made by the algorithm.

3.2.1.2. Decision tree classifier

In order to classify instances, decision trees lead them through the tree's structure, from the root to a particular leaf node that finally decides the instance's classification. One of the possible values for the attribute in question is represented by each branch that branches off of a node in the tree. Every leaf node designates a variable or category. The sample is categorized as negative because of its (building age = old, floor height = high, construction area = narrow, mortar quality = weak).

Although decision tree learning can be expanded to learn functions with multiple output values, it is especially well-suited for Boolean classification tasks. When it comes to handling errors, such as misclassifications of training examples and misdescriptions of those examples in the attribute values, decision tree learning techniques demonstrate resilience. Decision tree techniques can also be used in situations where null or missing values are present in training examples. The application of decision tree learning has been widespread, encompassing various fields such as medical studies, equipment maintenance, and the categorization of loan applicants based on their likelihood of non-payment. The most useful feature should be revealed when creating branches of decision trees. The type of measurement required for this is İnformation Gain. In other words, the aim of Information Gain while growing the decision tree is to choose among the candidate features. A measure called entropy is used when calculating the Information Gain. In relation to this Boolean classification, the entropy of a collection S that contains both positive and negative examples of a target concept is (Equation 3.4):

$$Entropy\ (S) = -p_{+}\ \log_2 p_{+} - p_{-}\ \log_2 p_{-} \ldots \qquad (3.4)$$

- S represents a sample of training examples.
- P+ denotes the proportion of positive examples.
- P - denotes the proportion of negative examples.

A measure of the impurity in a set of training examples is called entropy. On the other hand, information gain is a metric that assesses how well an attribute can categorize the training set. It measures the anticipated drop in entropy that occurs when examples are divided or sorted according to a particular characteristic. In essence, information gain measures the effectiveness of an attribute in partitioning and classifying the training data. In short, he decides on the name of the next branch of the Decision tree. The feature with the largest value of this value is placed on the next branch with its features (Equation 3.5).

$$Gain(S, A) = Entropy(S) - \sum_{v} | \left(\frac{Sv}{S}\right) | Entropy(Sv) \qquad (3.5)$$

- S – a collection of examples
- A – an attribute
- Values(A) – possible values of attribute A
- Sv – the subset of S for which attribute A has value v

The process of drawing broader, more general models by extrapolating from specific examples is known as inductive inference. In one setting, the goal is to become proficient at classifying things or circumstances by examining a collection of examples with predefined classes.

A divide-and-conquer strategy can be used to build a decision tree from a collection of instances. The resulting tree will have a leaf node with the class labeled if every instance in the set is part of the same class. A test-defining node is present at the root of the tree, and for every possible result, a corresponding subtree is created by applying the same procedure to the subset of instances linked to that specific result.

From a geometric perspective, a set of x attributes defines an x-dimensional feature space in which each instance is depicted as a point (Quinlan, 1996).

3.2.1.3. Random forest classifier

The random forest classifier functions by choosing a random subset from the training dataset and constructing an ensemble of dt. In this algorithm, the predictions derived from these randomly selected subsets of dt are combined to make the final prediction. As a result, the random forest classifier is categorized as an ensemble learning method. (Shaikhina et al. 2019).

Eq. 3.7 is a widely used formula for computing the Gini index. In this equation, K denotes the number of classes; $\hat{\rho}_{mk}$ represents the impurity measures, and N is the number of samples. Impurity reflects the fraction of observations that pertain to class k within the region

$R_{m,}$ where m signifies the majority class.. he Gini index assumes values between 0 and 1, and the closer it is to 0, the more effective the discrimination, as elaborated on.(Tan 2022).

$$\sum_{k=1}^{K} \hat{\rho}_{mk}(1 - \hat{\rho}_{mk}) \tag{3.6}$$

$$\hat{\rho}_{mk} = \frac{1}{N_m} \sum_{x_i \epsilon R_m} \mathrm{I}(y_i = k) \tag{3.7}$$

The class assignment probabilities calculated by each of the generated trees in a Random Forest classifier are averaged, usually using the arithmetic mean, to determine the final classification. Given the computational efficiency and propensity for non-overfitting of the Random Forest classifier, it is possible to optimize performance by setting the number of trees (Ntree) to a high value (Guan et al. 2013).

Recent years have witnessed an increase in interest in additional RF functions, including the use of variable significance (VS) to optimize feature space, internal proximity matrix measurements to measure correlation between high-dimensional datasets, and preliminary analysis of sample proximities to find outliers in training samples. The VI can be computed internally using a variety of methods, including the Mean Decrease in Accuracy (MDA) and the Mean Decrease in Gini (MDG). The majority of studies reported in this review used the MDA to determine the VS. Then, a search is made for the best division over the linear combinations obtained. If there are only a few inputs, let's say M, taking F a sizeable portion of M may enhance the strength but raise the correlation. A different strategy is defining extra features by choosing random linear combinations of a few of the input variables. This means that, the total number of variables to be merged, determines the feature that is produced. At each node, L variables are chosen at random and added with coefficients that are uniform random values on the range [1, 1]. The number of each tree is determined by this operation (Breiman 2001). Since RF is a classification tree-based algorithm, it can be used in both univariate and multivariate situations. As previously mentioned, an RF introduces the m parameter, a new parameter not present in conventional classification trees. Each node requires the specification of a subset of m predictive variables, ranging in number from 1 to a maximum of 6 in this case study. As the tree develops, this value of m stays constant, and the variable is chosen at random. The generalization error and the accuracy of the classifier are thus also impacted by the definition of this parameter, which also influences the correlation and strength of each tree (Pal and Mather, 2003).

Random Forest (RF) is a robust community-based machine learning model composed of multiple decision trees. Its robustness arises from the fact that the final decision is determined through majority voting, considering the outcomes of all the decision trees in classification tasks. In regression tasks, RF utilizes the average value of the outputs from all the available trees to make its final prediction. The prediction value that receives the most votes is chosen as a consequence of the voting process. For instance, the algorithm interprets the 3 red and 1 black outcome as red because of how the branches are arranged in Figure 3.20.

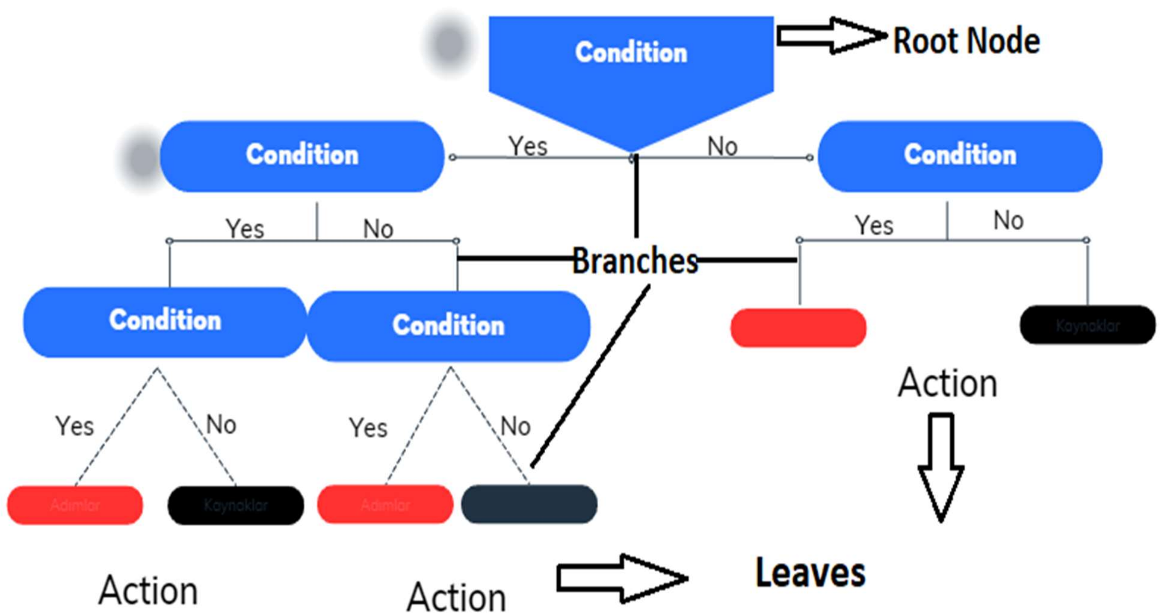


Figure 3.20. Example of a decision tree model

3.2.1.4. Support vector machine (SVM) classifier

For regression prediction and classification, machine learning techniques like Support Vector Machines (SVM) are employed. By automatically avoiding overfitting to the data, it maximizes predictive accuracy by utilizing machine learning principles. Systems that function inside the hypothesis space of linear functions in a high-dimensional feature space are known as support vector machines (SVMs). They are trained using an optimization-theory-derived learning algorithm that includes a statistical learning theory-inspired learning bias (Jakkula, 2006).

The general form of the SVM decision boundary is next obtained as follows:

$$f(x) = sign(\sum_{i\epsilon S} y_i \propto_i K(x_i^T x_j) + b) \qquad (3.8)$$

where S denotes a subset of training samples with Lagrange multipliers that are nonzero. ($\propto_i$), called support vectors, and K is the kernel matrix, induced by a kernel functionk$(x_i x_j)$ whose entries are

$$K_{ij} \equiv \mathrm{k}(x_i x_j)=(\phi(x_i),(\phi(x_j)) \qquad (3.9)$$

Regarding S-V-M, the function Φ typically represents a nonlinear kernel function. The performance of the SVM classifier depends critically on the choice of kernel parameters. Two key parameters that require optimal tuning are the kernel width parameter γ in the Radial Basis Function (RBF) kernel and the polynomial order denoted as d in the polynomial kernel. Unlike the linear kernel, the RBF kernel is particularly effective when dealing with situations where the relationship between class labels and features is nonlinear. In regression tasks, Support Vector Machines (SVM) aim to determine the best hyperplane that maximizes the number of data points falling within the chosen decision boundaries. (Tan 2022).

3.2.1.5. K-Nearest neighbor (KNN)

Instead of instructing the precise definition of the target function, example-based learning methods simply retain training instances. These instances are not generalized until there's a need to classify a new sample. When a newly encountered instance requires classification, a target function value is assigned based on its similarity to the stored instances. K-Nearest Neighbors (KNN) stands out as one of the widely used example-based learning algorithms.. Because instance-based techniques hold off on processing until a fresh sample needs to be classified, they are sometimes referred to as lazy learning techniques. Lazy learning techniques have the advantage of being able to estimate the target function locally and differently for each new instance that needs to be classified, as opposed to estimating it once for the entire instance space. The n-dimensional Rn space is the assumed space for all samples corresponding to points in the K-Nearest Neighbor Learning algorithm. Conversely, the K-Nearest Neighbor classifier is a supervised, non-parametric learning algorithm that uses proximity to classify or predict how a single data point will be categorized. It is a versatile method used for addressing both regression and classification problems.K-nearest neighbor classifier (K-neighbors classifier), on the basis of; It is a classification algorithm that assumes that similar points can be found close to each other. However, the distance must be defined before a classification can be made in the K-nearest neighbors classifier algorithm. At this stage, the euclidean distance is used (Imandoust and Bolandraftar 2016). Here the Euclidean distances between the samples xi = <xi1,…,xin> and xj = <xj1,…,xjn> are calculated with the following formula (Equation 3.10):

$$d(x_i, x_j) = \sqrt[2]{[\sum_r (x_{ir-}x_{jr})^2]} \tag{3.10}$$

The closest distance to each assigned cluster is calculated with Euclidean distances, and the uncertain variable is assigned to its nearest neighbor. For a given query example xq, the function values of f(x_q), x_q's k-nearest neighbors are calculated. If the target function is true, the f values of the k nearest neighbors are averaged.When the target function is discrete, the K-nearest neighbor classification (K-neighbors classifier) involves a voting process among the f values of the k-nearest neighbors. In essence, it determines the class membership as the output.An input set is assigned to the same class as the outputs of the class in which it has the highest number of votes among its k nearest neighbors (Imandoust and Bolandraftar 2016).

$$f(x_q) = \sum_{i=1} \frac{f(x_i)}{k} \tag{3.11}$$

3.2.1.6. Gradient boosting

GB algorithms are a machine operation technique used in regression and storage tasks. This expansion is to .increase the limits of the base model used in the machine learning run and reduce the memory. It works depending on light gradient boosting, reflecting the decision tree (decision tree). In this method, training samples are ordered from largest to smallest (from least untrained to most educated) according to the absolute values of the gradients of the missing values. Then, the first n data with the largest gradients are used in conjunction with randomly selected measures with such small gradients (the most educated measures) that are designed for use later in progression, keeping the limits of prediction (Friedman, 2001).

LightGBM is an algorithm based on histograms, which decreases the computational workload by converting continuous variables into discrete ones. LightGBM is capable of handling categorical features as well. It shares many parameters with XGBoost, including the number of estimators, maximum tree depth, training instance subsampling ratio, number of iterations, and objective functions, making these parameters commonly used for model tuning purposes. (Tan 2022).

GBDT is a widely-used ensemble-classifier algorithm. It operates on a training set containing data samples (x1, y1), (x2, y2), (x3, y3), and so on up to (xn, yn), where x represents the data samples, and y denotes the class labels. The algorithm employs F(x) to represent the estimated function, and its primary objective is to reduce the loss function L the loss function L(y, F(x)) (Chen et al. 2019) :

$$F = argminE_{e,x}[L(y, F_{m-1}(x) + \gamma_m + h_m(x)] \qquad (3.12)$$

$$\gamma_m = argmin \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i) \qquad (3.13)$$

Where, m is iteration number and $h_m(x)$ represents the base decision tree.

LightGradientBoosting (LightGbm) and ExtremeGradientBoosting (XGBoost) are the two most used gradient boosting methods. A machine learning technique with a focus on model performance and computing speed is called XGBoost. It was first presented by Tianqi Chen and is now a part of the Distributed Machine Learning Community's larger toolset. This method can be used for both classification and regression tasks, and it is made to handle big and complex datasets. The dispersed high-performance framework, LightGBM, developed by Microsoft, employs decision trees for sorting, classification, and regression tasks, similar to XGBoost. The algorithm we wish to employ typically depends on the kind of processing unit we have available for running the models. XGBoost is actually faster on the CPU even though it performs comparably worse on the GPU than LightGBM (Bentéjac et al. 2021).

3.2.1.7. Multi-layer perceptron

A Multi-layer Perceptron (MLP) is a type of artificial neural network that incorporates multiple hidden layers and nodes in each layer. Its multi-layered structure enables it to learn non-linear relationships. Multi-layer perceptron is based on logistic regression. Logistic regression is essentially a linear regression model with a sigmoid activation function for classification. It has a single layer and can only learn linear relationships. Logistic regression is easy and fast to train using simple optimization algorithms such as gradient descent. On the other hand, training an MLP is more complex. It requires the backpropagation algorithm for training. In cases with a large number of parameters, overfitting can occur. Regularization techniques (e.g., dropout) and optimization algorithms (e.g., Adam) can be employed in MLP to mitigate these issues. The multi-layered nature of MLP has been used in this study because it allows it to learn complex patterns and nonlinear relationships, resulting in better performance in various fields. A sample network is given in Figure 3.21.

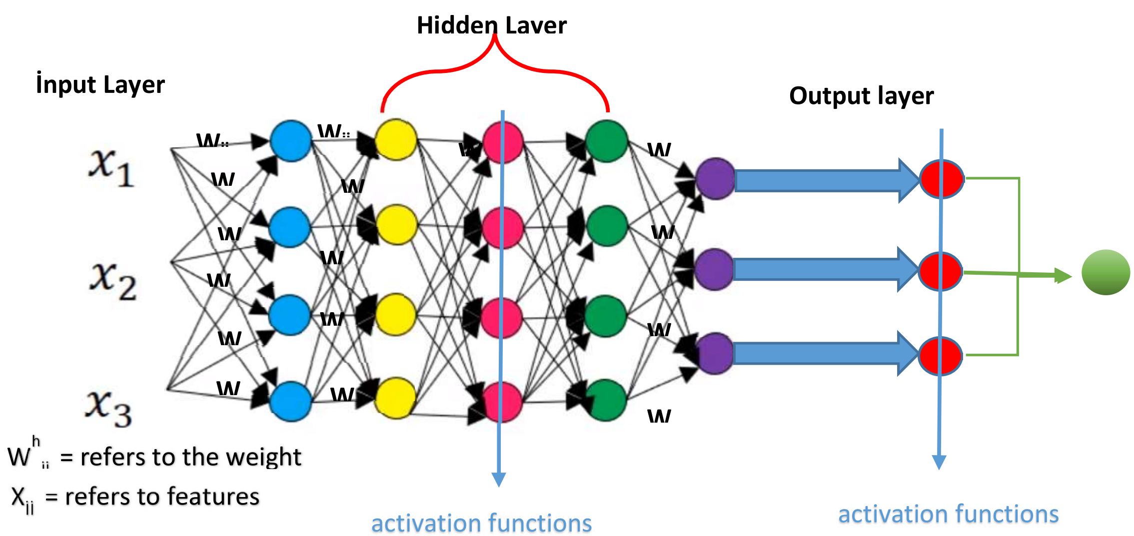


Figure 3.21. Multi-layer perceptron algorithm working principle

3.2.1.8. Multivariate adaptive regression splines

Friedman firstly proposed the multivariate adaptive regression line (MARS). Based on the basis functions that were recovered from the regression data, the MARS algorithm provides a dynamic link between the variables. The construction of a flexible regression model involves the use of basis functions that map to several sets of independent variables. About the fundamental relationships that function between the independent and dependent variables, MARS makes no assumptions. The splines are typically coupled in a smooth manner, and the piecewise curves (polynomials), commonly referred to as basis functions

(BFs), produce a flexible model that can handle both linear and non-linear attitudes (Figure 3.22) (Zhang and Goh 2016).

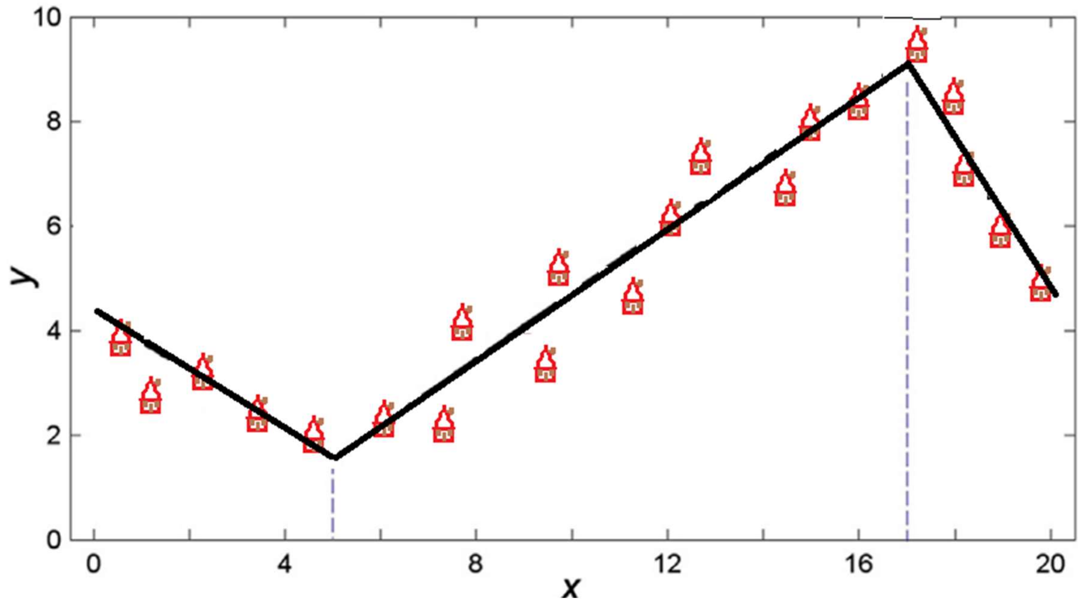


Figure 3.22. Knots and linear splines for a simple MARS example

The research on MARS demonstrates that it is hardly ever used, especially in the field of construction. Mars takes advantage of multicollinearity, which is the interplay of independent variables. So, unlike other algorithms, in the MARS algorithm, a single function is used to define the relationship each other the features of the building identified in the study (Koc 2022).

The Mars algorithm performs analysis by determining the parameters that will have the greatest impact on estimating the outcome. The MARS model does this by employing the two fundamental processes of backward elimination and forward selection. Beginning with all features present in the model, backward elimination then eliminates the least important elements one at a time. The redundant BFs with the lowest contributions are removed during the backward phase. Forward selection starts with an empty model, and it incrementally adds the most crucial features.

3.2.1.9. Ensemble learning

Employing several algorithms and combining their predictions is the basic idea behind the ensemble learning. Theoretically, the average error of a model can be decreased by a factor of M if we have a committee of M models with uncorrelated mistakes (Sewell 2008).

Ensemble methods combine the results of two or more separate machine learning algorithms in an attempt to produce an aggregated result that is more accurate than either of the individual algorithms. In soft voting, the probabilities of each of the classes are averaged to produce a result. For example, if the first algorithm predicts a building with a 30% probability and the second algorithm predicts with 90% probability, the community predicts that the object is a risky building with a weighted mean with respect to the probability. In hard voting, each algorithm has its own vote. In the final vote, the predictions of each algorithm are taken into account, with the community choosing the class with the highest votes. For example, if the predictions of each algorithm are summed up and the majority concludes that the building is risky, the building in question will be classified as risky. In this study, 8 algorithms were combined and the highest success rate was obtained as a result of hard voting (Figure 3.23).

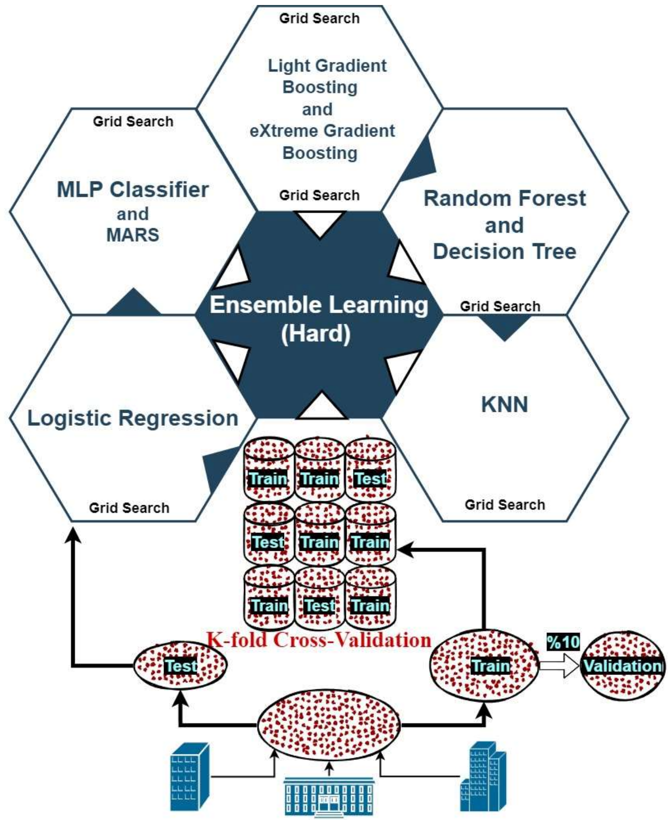


Figure 3.23. Emsemble learning study chart

3.3. Performance and Evaluation Metrics

The objective was to achieve precise predictions of the seismic risk distribution for a substantial building stock by employing machine learning algorithms. In all networks, if the risk score is 1 (0), the risk status is taken as risky (non-risky) from the machine learning algorithm. At the same time, more economical determinations are aimed by dividing the current seismic situation into more risk scores. However, while making these determinations, losses greater than the accuracy percentage of the algorithm should not be given. With this motivation, the performances of the algorithms used will be evaluated not only according to the percentage of success, but also according to metrics such as true positive (TP), true negative (TN), false positive (FP), false negative (FN). These metrics were converted to precision, recall, and combined measure (i.e., $F_{measure}$) given in Equations (Saito and Rehmsmeier 2015).

$$Precision = \frac{TP}{TP + FP} \tag{3.14}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.15}$$

$$F_{measure} = \frac{2\ Precision\ \times Recall}{Presicion + Recall} \tag{3.16}$$

Considering that our dataset is predominantly composed of risky structures, it is difficult and important for them to correctly select risk-free structures. In large data sets, algorithms sometimes make an overfitting error, that is, when a small number of data (i.e., risk-free structures) is encountered in the data set, they give the same answer with a large number of data (i.e., risky structures). In order to overcome this problem, it is aimed to make the necessary controls with confusion matrix (Sammut and Webb 2011). The confusion matrix is a fundamental concept in machine learning that provides a breakdown of the actual and predicted classifications made by a classification system. It has two dimensions, one indexed by the true class to which an object belongs, and the other indexed by the class that the classifier predicts for it. The fundamental confusion matrix for a multi-class classification problem is shown in Figure 3.24. In both columns of the matrix, the number relating to risky structures=1 denotes accurate forecasts, whereas the number related to risk-free structures=0 denotes inaccurate predictions (Deng et al., 2016).

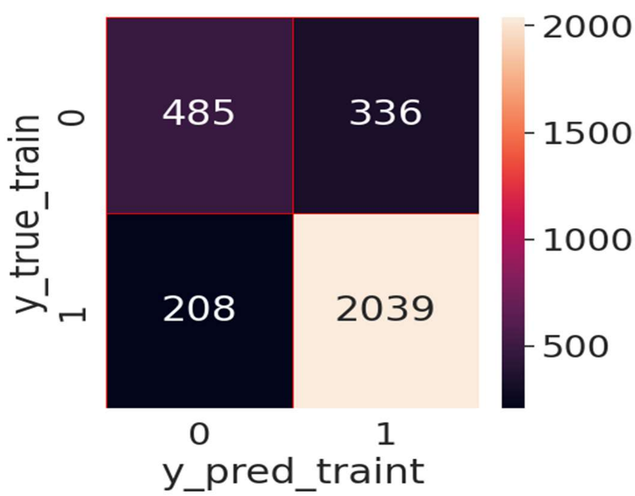


Figure 3.24. Example of confusion matrix used in the study

4. RESULTS

This study sought to create a quick and straightforward visual screening technique utilizing machine learning algorithms to estimate the damage severity of masonry buildings. The goal was to create a calculation network based on external observations of the structures, eliminating the need for technical personnel to enter the buildings for assessment.

The dataset of 4356 buildings was utilized and analyzed with various machine learning algorithms, including regression and classification algorithms. All calculations were conducted using the Google Colaboratory framework, which provides high-speed computations and allows for easy access to large datasets.

The original dataset was split into 3,484 training samples and 872 test samples (Joseph, V. R.2022). To address the issue of imbalanced data, the Smote method was applied to synthetically increase the number of buildings with low earthquake risk in the training dataset. From the original 12 parameters, the study reduced the number to the six most influential and easily detectable parameters, as described in the Data Preprocessing section. K-fold cross-validation was performed to validate the federated models and prevent overfitting.

The study utilized multiple ML algorithms, including Decision Tree Classifier, Gradient Boosting, Logistic Regression, Random Forest, K-Mean Clustering, Support Vector Machine, LGBMClassifier, MLPClassifier, and MARS, to predict rapid seismic analysis results. The Ensemble Learning method was proposed as the first approach to combine predictions from these algorithms. Voting was used as a simple and effective way to aggregate the predictions. The post-voting method was used for the dataset where the algorithms provided seismic analysis results with percentage estimates (soft voting) in binary format (1 or 0) or in multiple format (1.0 or 2, 3). The decision to triple, double, or quadruple classification depends on cost and the number of prioritization layers. If buildings only need to be classified as risky or risk-free, binary classification will be a costly option as more detailed analyses will be made for risky buildings. Other options can be divided into many classes such as risky, medium risk, and no risk, and are less costly options as they can lower the number of buildings that require in-depth inspection.

The components and schematic representations of the Ensemble Learning method were presented in Figure 4.1, showcasing its application in the study to predict seismic analysis results more accurately and efficiently.

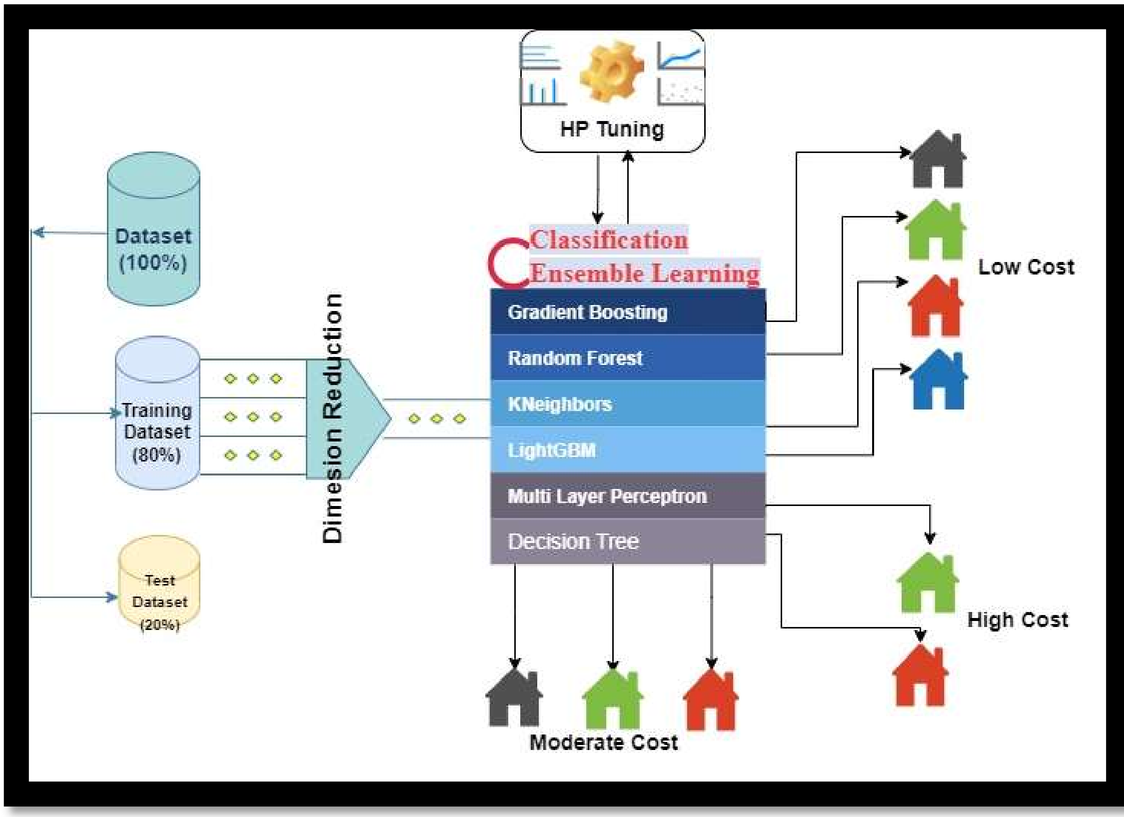


Figure 4.1. Future engineering+dimensiol Reduction+emsemble learning classifier+voting classifier

It's crucial to emphasize the significance of selecting the appropriate hyperparameters for this algorithm to enhance accuracy. Each model has numerous hyperparameters, and an effective approach to finding the best set of hyperparameters is to experiment with various combinations and assess the outcomes. For instance, in logistic regression, the choice of an l2 penalty is preferred because it yields better predictions when the output variable is influenced by all input characteristics. Likewise, the coefficient value was determined based on the values that resulted in the highest success rate, as illustrated in Figure 4.2(a).In the decision tree algorithm, there are many parameters that affect the success percentage. Of these, one of the values "maximum depth" providing the highest percentage of success was selected in the graph in Figure 4.2. In the random forest algorithm, there are many parameters that affect the success percentage. Among these parameters, "minimum samples split" was selected as one of the values that yielded a high success percentage, as shown in the graph in Figure 4.2(c). It's worth noting that this algorithm comprises multiple parameters that influence the results. When implementing the KNN algorithm in this study, the "number of data points (k)" parameter proved to be of utmost importance for improving the success percentage.Therefore, it was clear from Figure 4.2(d) that the worst neighbor values for the dataset were $k < 2$, $3 < k < 4$, and $k > 13$. In this case, these values should be avoided in choosing the k value. It has been observed that all values do not change the percentage of success in the penalty parameter selection of the Support Vector Machine Classifier algorithm (Figure 4.3.). In this study, these graphs were employed, and the optimal values necessary for enhancing the success rate of each algorithm were identified through the grid search method.

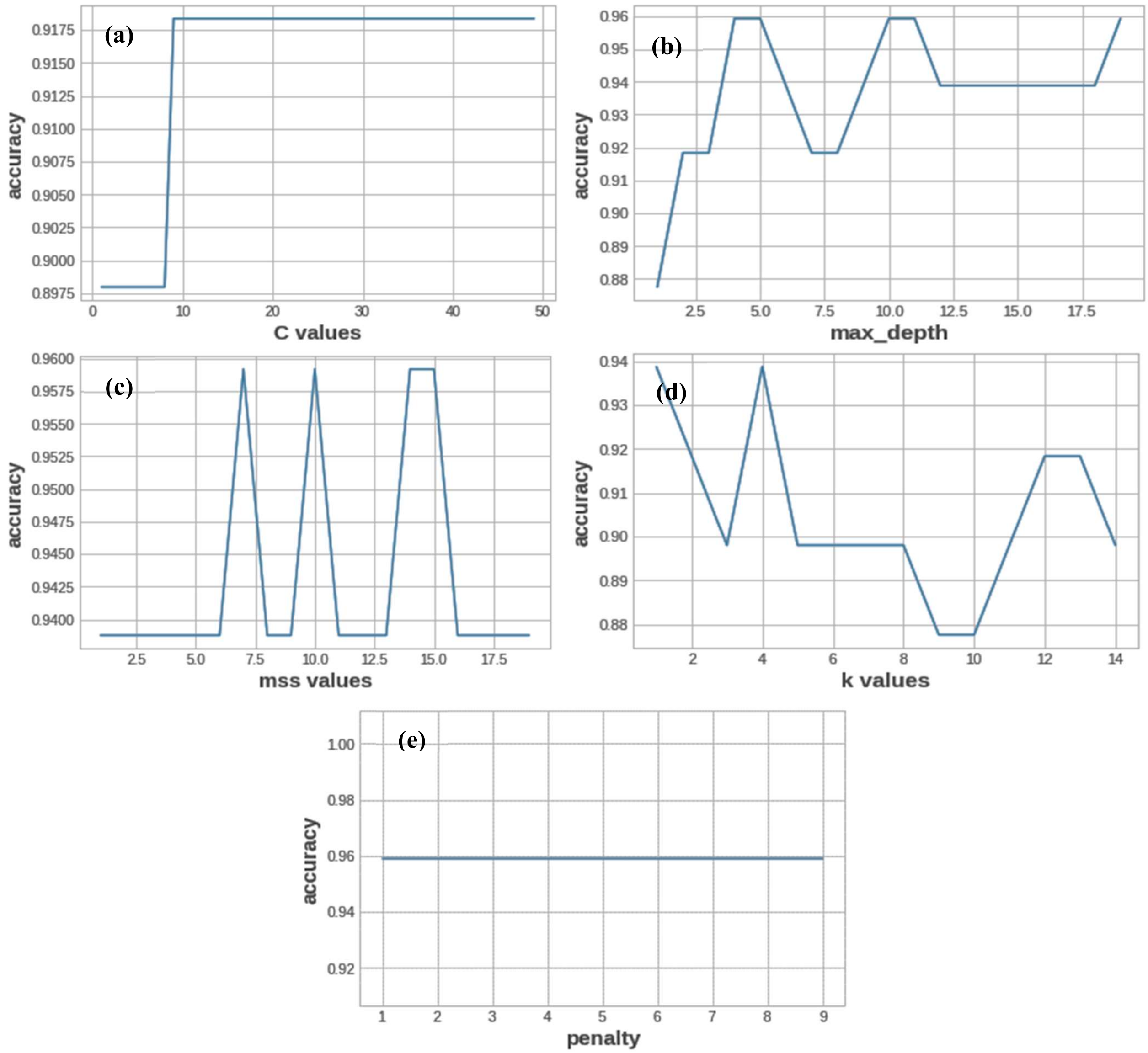


Figure 4.2. Effect of algorithm parameters: (a) LR, (b) DTC, (c) RFC, (d) KNN and (e) SVMC

In this study, the grid-search approach was utilized to optimize the ensemble models, particularly focusing on the stacking model's base model hyperparameters. The optimization process was depicted in Figure 4.3, where the intervals for the hyperparameters were determined based on graphics drawn within the scope of the study.

The steps involved in the meta-model optimization procedure are as follows:

- Determine the Hyperparameter Value Range: The first step is to determine the value range for the hyperparameters that require optimization. In this study, these intervals were selected from the graphs created for each algorithm during the analysis process.
- Iteratively Train and Evaluate Model: After that, the model is iteratively trained by experimenting with different hyperparameter combinations. A model is linked to every combination, and its error or performance metric is found.
- Compare Errors: Through the comparison of errors across various combinations of hyperparameters, the optimization algorithm aims to pinpoint the hyperparameters that align most effectively with the prediction criteria and yield the most accurate results..
- Select Optimal Hyperparameters: Based on the comparison of errors, the hyperparameters that result in the best predictive performance are selected.
- The grid-search algorithm is an effective method to systematically explore different hyperparameter combinations and find the optimal set that leads to improved model performance. It allows for fine-tuning the models, ultimately enhancing the accuracy and reliability of the ensemble approach for seismic analysis prediction in this study ((Qu et al., 2021)).

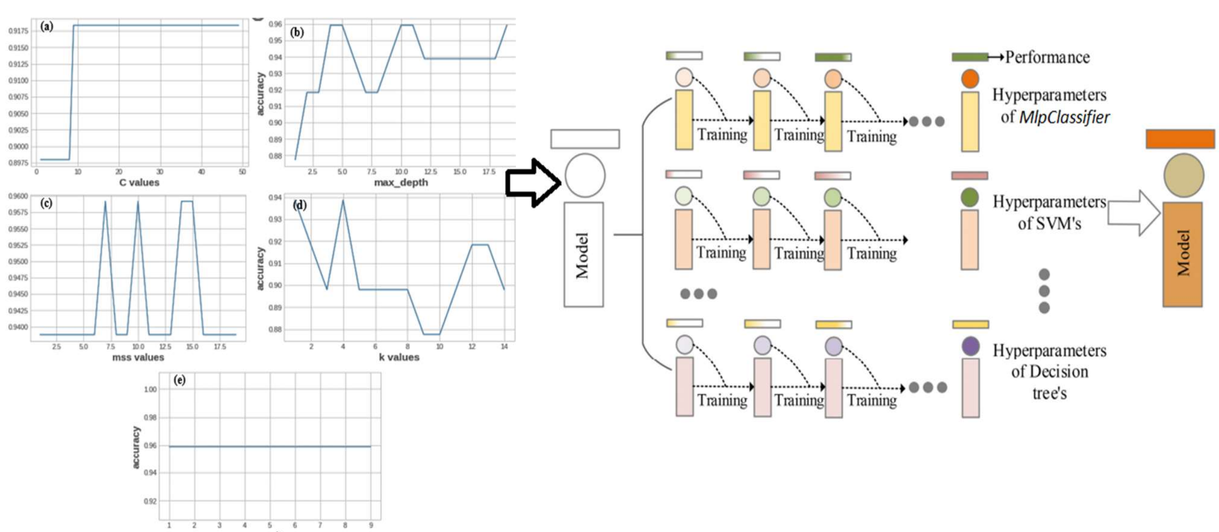


Figure 4.3. Grid search method whose intervals are selected from the graphs

So as to obtain more accurate results, a hybrid method was obtained by using regression method (i.e MARS and Decision Regressor) and classical ML algorithms, Decision Tree Classifier, Logistic Regression, Random Forest, K-Mean Clustering, Support Vector Machine used in regression analysis. LGBTMSClassifier, MLPSClassifier. The working principle of this hybrid system is to estimate the Ve/Vr ratios of the buildings with the MARS algorithm and then classify these values with ensemble learning algorithms. That is, this method is based on the comparison of regression estimates after classification rather than the classification of Ve/Vr values before analysis. This Methot is seen in the summary in Figure 4.4.

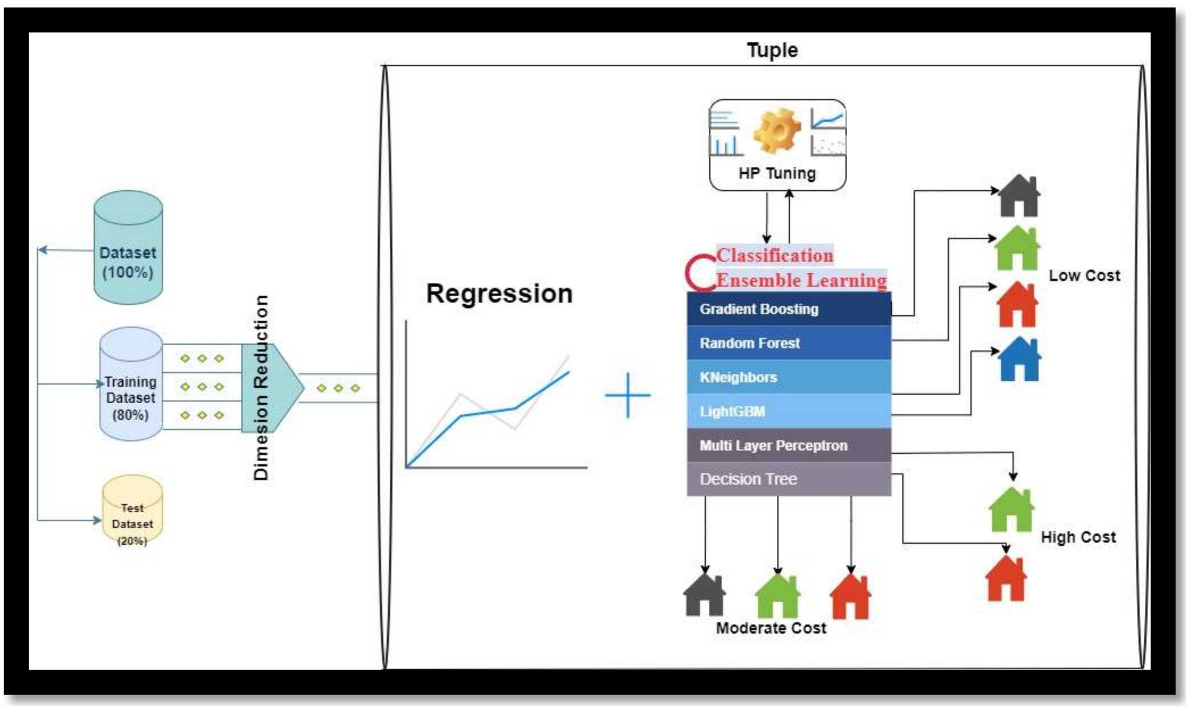


Figure 4.4. Future engineering+dimensiol reduction+tuple (regression+emsemble learning classifier+voting classifier)

If the success rate of the combined MARS (Multivariate Adaptive Regression Splines) and Voting Classifier algorithm approach proves to be unsatisfactory, the utilization of regression algorithms for estimating RVS (Risk Versus Safety) values would be reconsidered."

4.1. Machine – Learning Based Rapid Seismic Risk Estimation Results

Urban transformation in Turkey, which is located in an earthquake-prone zone, is primarily driven by the need to demolish risky buildings and replace them with earthquake-resistant structures. However, due to the large number of at-risk buildings that have not received engineering services, it is not feasible to demolish all of them and construct new buildings in their place. As a result, a prioritization approach is essential, and urban transformation should start from the most risky buildings. In this study, the filtering and classification of masonry structures were performed based on the value (Ve/Vr = RVS), which represents the ratio of shear force on risky walls to the total floor shear force. The study proposed three different risk classes for urban transformation:

Two Risk Classes: Buildings falling under this category are considered difficult and costly to undergo urban transformation due to their high-risk nature.

Three Risk Classes: This method was determined as the most optimum separation method in the study. Buildings are classified into three risk classes based on their RVS values, allowing for a more efficient and targeted urban transformation process.

Four Risk Classes: This classification involves dividing buildings into four risk classes. However, it was found to be less successful compared to the three-risk class method and may not provide as effective results for urban transformation.

By utilizing these risk classifications, urban transformation efforts can be focused on the most vulnerable buildings, ensuring a more strategic and cost-effective approach to seismic risk mitigation in Turkey.

4.1.1. Experimental results of 2 option

The study involved analyzing a dataset comprising 4356 buildings using multiple machine learning algorithms. All computations and analyses were carried out using the Google Colaboratory framework, which employs Python programming. To begin, the dataset was initially split into two sets: a training set of 4356 samples and a test set containing 872 samples. The data division was carried out using the "Train Test Split" function from the "Sklearn model selection" library, ensuring that the data was correctly partitioned for both training and evaluation purposes. To categorize the buildings based on their seismic risk levels, the "cut()" method from two different Pandas libraries was utilized. The method facilitated the creation of two classes within the "rvs" parameter, effectively grouping the data. By implementing this technique, estimates were generated for the dataset, which was then divided into two risk classes: "non-risk" and "risky." The limit RVS values that define the risk classes are detailed in Table 4.11. In summary, the study involved analyzing the dataset using various machine learning algorithms within the Google Colaboratory framework. The data was divided into training and test sets, and the "cut()" method was employed to group the data into two distinct risk classes based on the RVS parameter values. Table 4.1 provides essential information regarding the RVS thresholds defining the risk classes.

Table 4.1. RVS risk classification thresholds

Risk Status	Classification	Threshold
Non-Risky	0	0-0,50
Risky	1	0,51-1

Furthermore, K-fold cross-validation was utilized in the validation phase to reduce the risk of overfitting in the trained dataset resulting from the algorithms. As mentioned earlier, the study incorporated a range of machine learning algorithms, such as MARS, logistic regression, decision tree classifier, gradient boosting, LightGBM, random forest classifier, support vector machine classifier, and K-neighbors classifier. The accuracy of each method and their respective performance metrics are summarized and presented in Table 5.1. This table serves as a comprehensive evaluation of the models' performance, allowing for a direct comparison of the different algorithms used in the study. Researchers can use this table to gain insights into the strengths and weaknesses of each method in predicting the seismic risk situations of structures when divided into two classes. The K-fold cross-validation technique, by partitioning the data into multiple subsets and validating the models on different combinations of training and test data, aids in obtaining more reliable and generalizable performance measures, thereby enhancing the credibility of the study's findings.

Table 4.2. Results of all machine learning methods

<table>
<tr><th colspan="2">Method</th><th colspan="3">Training Error Accuracy (%)</th><th colspan="2">Testing Error Accuracy (%)</th></tr>
<tr><td colspan="2">Logistic regression (LR)</td><td colspan="3">74.54</td><td colspan="2">75.92</td></tr>
<tr><td colspan="2">Decision tree classifier (DTC)</td><td colspan="3">91.44</td><td colspan="2">81.42</td></tr>
<tr><td colspan="2">Random forest classifier (RFC)</td><td colspan="3">99.4</td><td colspan="2">86.8</td></tr>
<tr><td colspan="2">Voting Classifier (VC)</td><td colspan="3">98.63</td><td colspan="2">87.50</td></tr>
<tr><td colspan="2">Mlp Classifier (Mlp)</td><td colspan="3">97.70</td><td colspan="2">86</td></tr>
<tr><td colspan="2">Gradient Boosting (GB)</td><td colspan="3">100</td><td colspan="2">88.30</td></tr>
<tr><td colspan="2">Extreme Gradient Boosting (XGB)</td><td colspan="3">100</td><td colspan="2">88.36</td></tr>
<tr><td colspan="2">Light GBM (LGB)</td><td colspan="3">97.86</td><td colspan="2">87.27</td></tr>
<tr><td colspan="2">Mars and VC with Grid Search (MVC)</td><td colspan="3">99.11</td><td colspan="2">86.99</td></tr>
<tr><td colspan="2">KNeighborsClassifier (KNN)</td><td colspan="3">94</td><td colspan="2">79.2</td></tr>
<tr><td></td><td colspan="2">GB</td><td colspan="2">XGB</td><td colspan="2">VC</td></tr>
<tr><td></td><td>TS=1</td><td>TS=0</td><td>TS=1</td><td>TS=0</td><td>TS=1</td><td>TS=0</td></tr>
<tr><td>TS=1</td><td>97 TP</td><td>55 FN</td><td>90 TP</td><td>62 FN</td><td>98 TP</td><td>54 FN</td></tr>
<tr><td>TS=0</td><td>0 FP</td><td>47 TN</td><td>36 FP</td><td>684 TN</td><td>55 FP</td><td>665 TN</td></tr>
<tr><td>Precision (%)</td><td colspan="2">0.88</td><td colspan="2">0.88</td><td colspan="2">0.88</td></tr>
<tr><td>Recall (%)</td><td colspan="2">0.88</td><td colspan="2">0.88</td><td colspan="2">0.88</td></tr>
<tr><td>Fscore (%)</td><td colspan="2">0.88</td><td colspan="2">0.88</td><td colspan="2">0.88</td></tr>
</table>

When evaluating all models using appropriate evaluation metrics and classifying the seismic risk situations of structures into 2 classes, the Voting Classifier algorithm emerges as the most successful estimator among the models (as depicted in Figures 5.1, 5.2, and 5.3). In these figures, the performance and results of different algorithms are presented, and it is evident that the Voting Classifier algorithm outperforms the others in accurately predicting the seismic risk situations, achieving the highest levels of success in the binary classification task. This observation highlights the effectiveness and robustness of the Voting Classifier in this specific context.

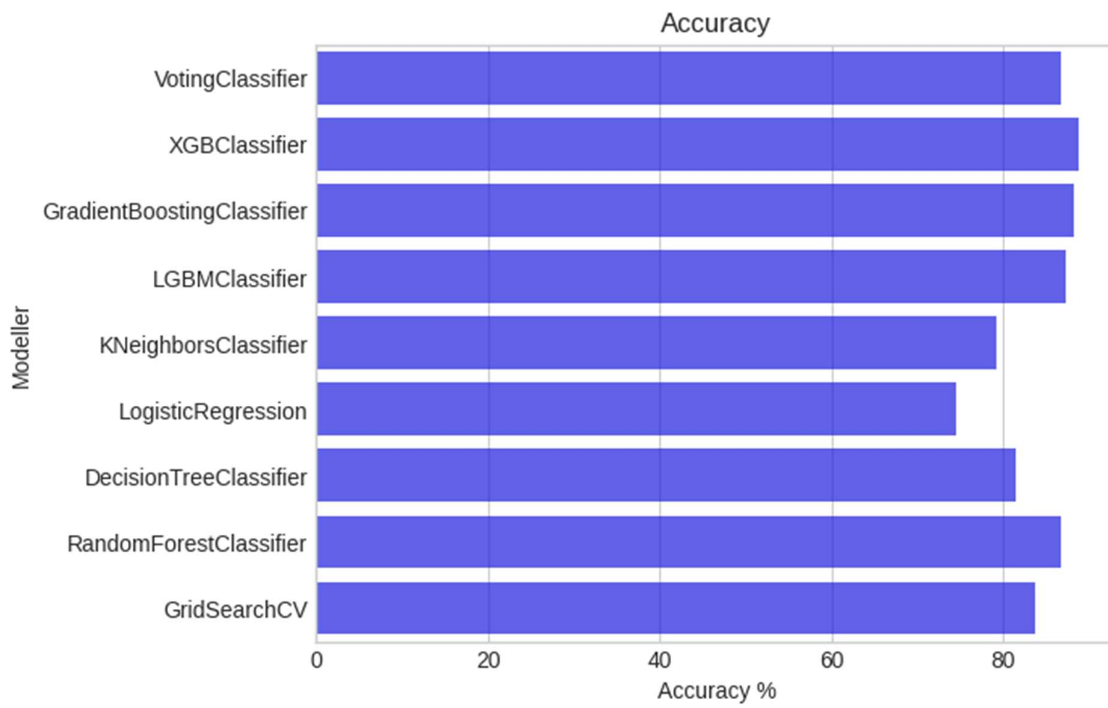


Figure 4.5. Comparison of accuracy rates of algorithms

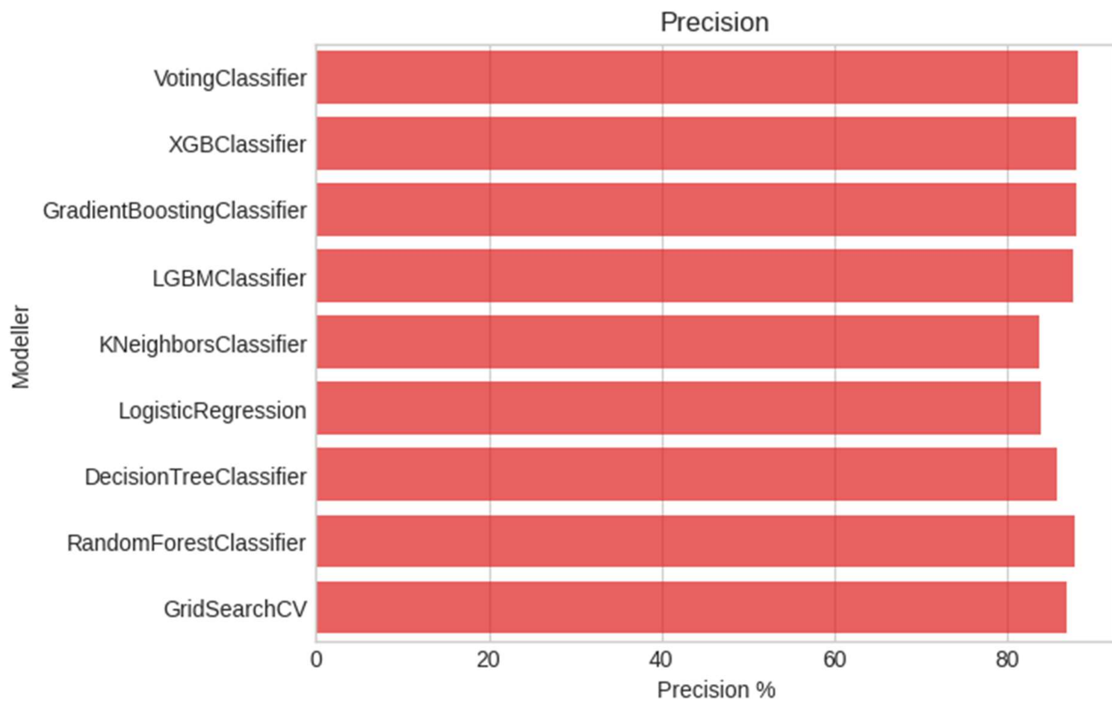


Figure 4.6. Comparison of precision rates of algorithms

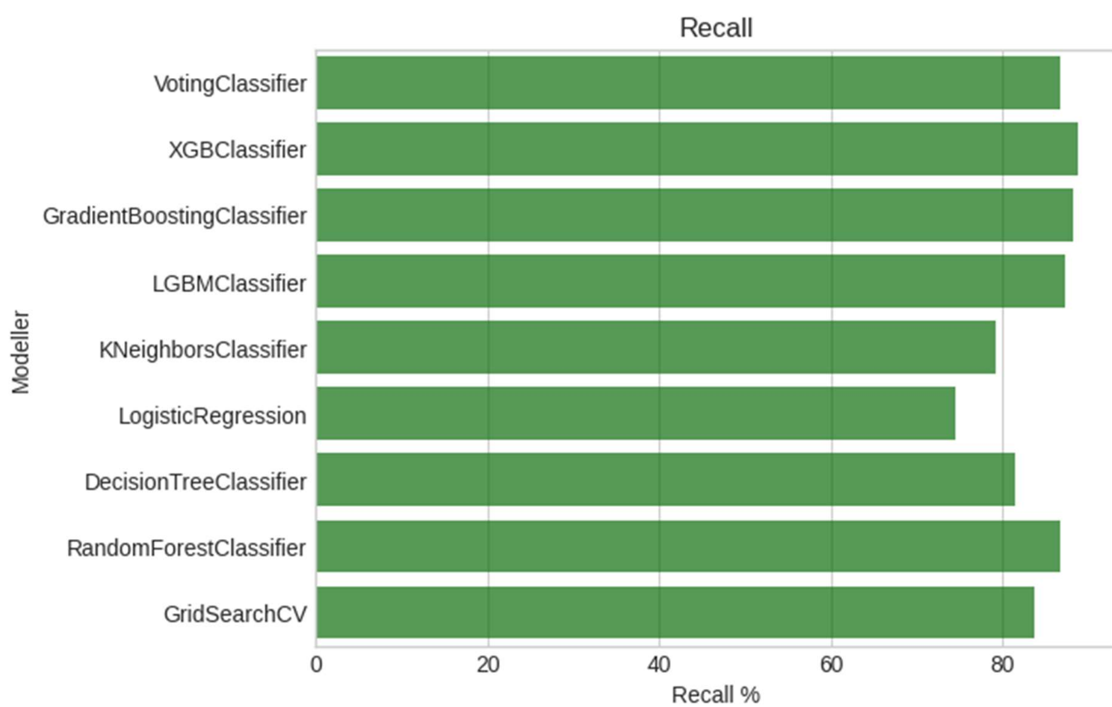


Figure 4.7. Comparison of recall rates of algorithms

Among the algorithms, Voting Classifier, Grading Boosting and Extreme Grading Boosting have very close success rates. However, since it is the most difficult and important to predict the non-risk ones in the data set, the most suitable algorithm to use the more successful Voting Classifier algorithm for the two-risk estimation of structures is selected in this section. When we examined the confusion matrix of Voting Classifier, which is the most successful algorithm, highly accurate prediction results were obtained in both test and train in all risk classes.(Figure 4.8 and 4.9)

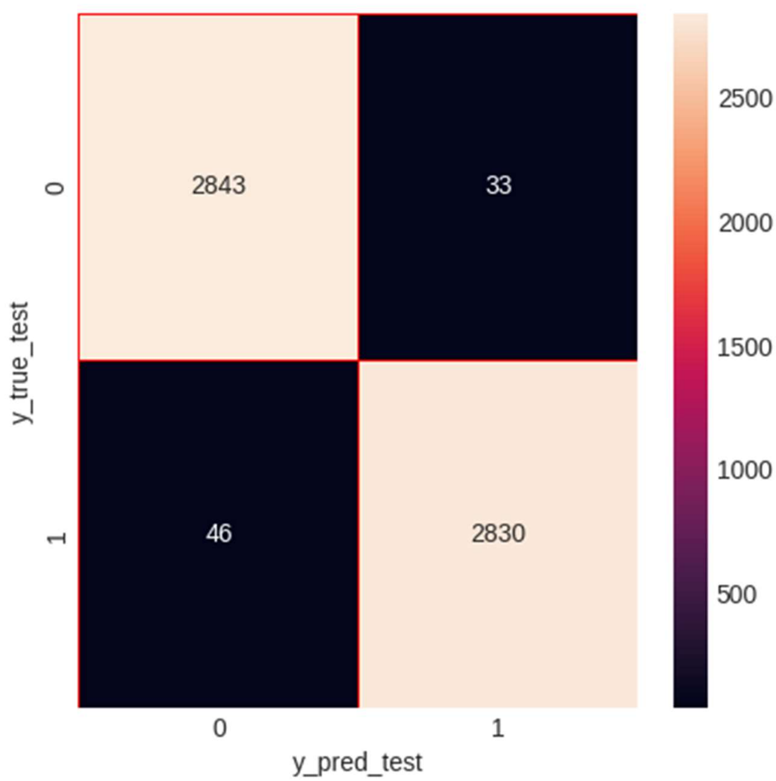


Figure 4.8. Voting classifier algorithm confusion matrix result train

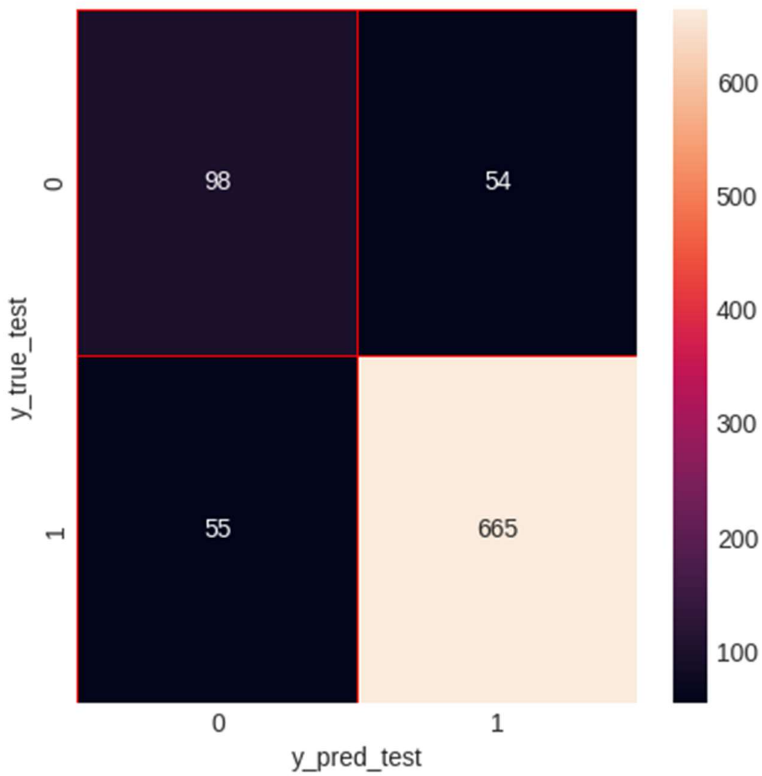


Figure 4.9. Voting classifier algorithm confusuan matrix result test

4.1.2. Experimental results of 3 option

In this study, an analysis was conducted on a dataset comprising 4,356 buildings using various machine-learning algorithms. The calculations and analyses were conducted using the Google Colaboratory framework, which is based on Python. Initially, the dataset was split into two sets: a training set comprising 4,356 samples and a test set containing 872 samples. This partition was carried out using the "Train Test Split" function from the "Sklearn model selection" library, ensuring that the data was divided into suitable subsets for training and evaluation. To categorize the buildings based on their seismic risk levels, the Pandas Library's "cut()" method was utilized. This method allows for the creation of three classes within the "rvs" parameter, effectively grouping the data. The "cut()" method offers two options: dividing data into custom-sized bins and dividing data into equal-sized bins. In this study, both methods were employed. Equal-sized bins provide an easy visualization of the data distribution, while custom bins allow for logical categorization based on specific risk levels.

As a result of applying the "cut()" method, the data set was categorized into three risk classes: "no risk," "risky," and "very risky." The table in the study, referred to as "Table 0-1," shows the limit values of the RVS parameter corresponding to each risk class.In summary, this study used machine learning algorithms in Google Colaboratory to analyze a dataset of buildings. The buildings were then categorized into three risk classes based on their RVS values, providing valuable insights for seismic risk assessment and urban transformation decisions. Table 4.3.

Table 4.3. RVS risk classification thresholds

Risk Status	Classification	Threshold
Non-Risky	0	0-0,3333
Risky	1	0,3333-0.66667
High-Risky	2	0.66667-1

In addition to the successful performance of LightGBM, K-fold cross-validation was applied during the validation stage to mitigate the risk of overfitting in the trained dataset caused by the various algorithms. As previously explained, the study utilized a range of machine learning algorithms, including MARS, logistic regression, decision tree classifier, gradient boosting, LightGBM, random forest classifier, support vector machine classifier, and K-neighbors classifier. To impartially assess the performance of each method, a range of performance metrics, including accuracy, precision, recall, F1-score, and others, were computed and are displayed in Table 5.4. This table serves as a comprehensive summary of the models' effectiveness in predicting the seismic risk situations of structures when divided into three risk classes: "no risk," "risky," and "very risky." K-fold cross-validation plays a pivotal role in the evaluation process by mitigating the impact of particular data splits during training and testing, thereby ensuring that the models' performance is more robust and generalizable. By validating the models on multiple folds of the data, the study attains more robust and generalizable performance metrics, making the results more reliable and applicable in practical scenarios. Researchers and decision-makers can refer to Table 4.4 to gain insights into the strengths and weaknesses of each algorithm in predicting the seismic risk situations across the different risk classes. This information can be used to make informed choices about the most suitable algorithm for seismic risk assessment to three risk class and urban transformation endeavors.

Table 4.4 Results of all machine learning methods

Method	**Training Error Accuracy (%)**	**Testing Error Accuracy (%)**
Logistic regression (LR)	67.09	61.45
Decision tree classifier (DTC)	86.26	77.98
Random forest classifier (RFC)	99.76	84.06
Voting Classifier (VC)	99.31	84.43
Mlp Classifier (Mlp)	97.70	86
Gradient Boosting (GB)	100	85.32
Extreme Gradient Boosting (XGB)	100	86.46
Light GBM (LGB)	99.19	86.58
Regression and VC with G.Search (MVC)	99.60	81.99
KNeighborsClassifier (KNN)	94	79.2

	GB				XGB		LightGBM		
	TS=0	TS=1	TS=2	TS=0	TS=1	TS=2	TS=0	TS=1	TS=2
TS=0	26	1	0	26	0	1	26	1	0
TS=1	2	155	59	1	157	58	3	157	56
TS=2	1	65	563	2	56	571	2	55	572
Precision (%)	0.853			0.864			0.865		
Recall (%)	0.854			0.864			0.865		
Fscore (%)	0.853			0.864			0.865		

Upon comparing all models using evaluation metrics and dividing the seismic risk situations of structures into 3 classes, LightGBM emerges as the algorithm that achieves the most successful estimations (as illustrated in Figures 4.10, 4.11 and 4.12). The evaluation metrics employed in the study likely include accuracy, precision, recall, F1-score, and possibly others to assess the performance of each model in predicting the three risk classes (i.e., "no risk," "risky," and "very risky"). The consistently high performance of LightGBM across these evaluation metrics indicates its effectiveness in accurately classifying buildings into the appropriate risk categories. The superior predictive capabilities of LightGBM make it a promising choice for seismic risk assessment, providing valuable insights for urban transformation decisions and seismic risk mitigation strategies. These figures provide visual evidence of LightGBM's proficiency in handling the complex task of classifying buildings into multiple risk classes, further reinforcing its position as the most successful algorithm in this specific study.

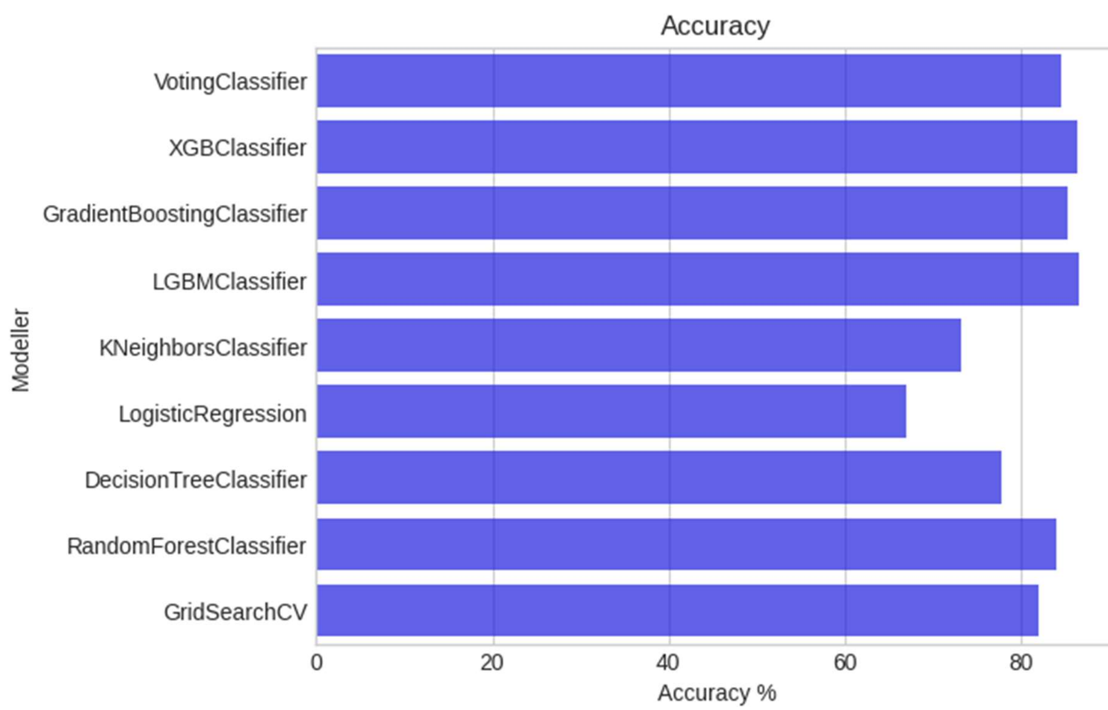


Figure 4.10. Comparison of accuracy rates of algorithms

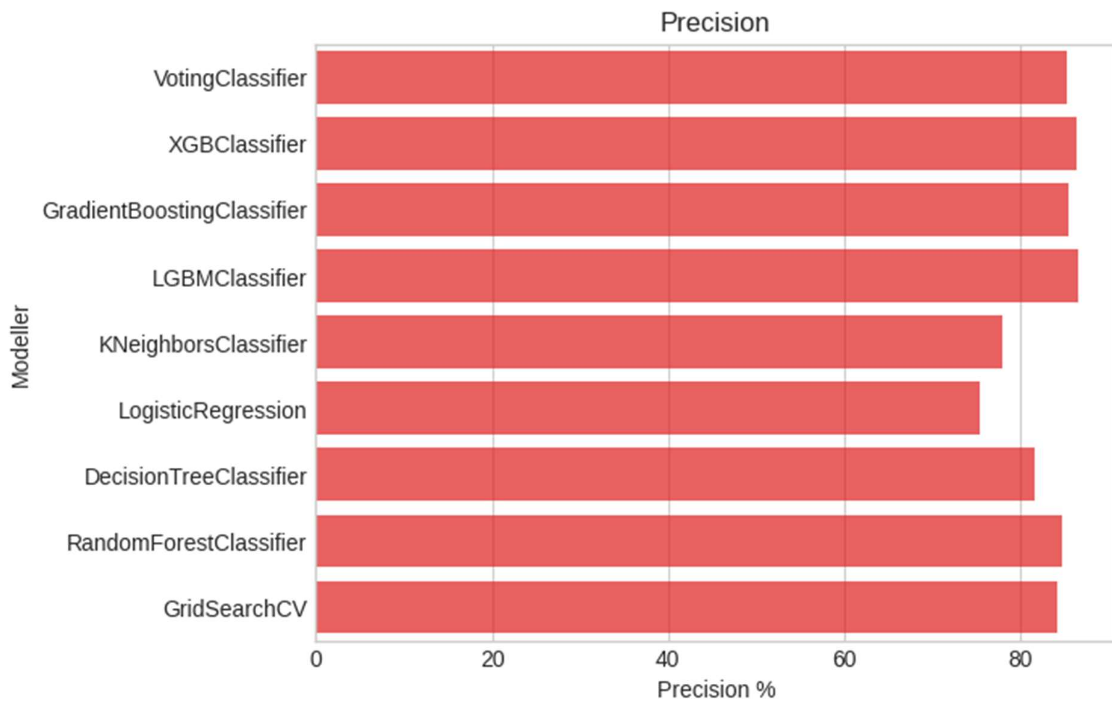


Figure 4.11. Comparison of precision rates of algorithms

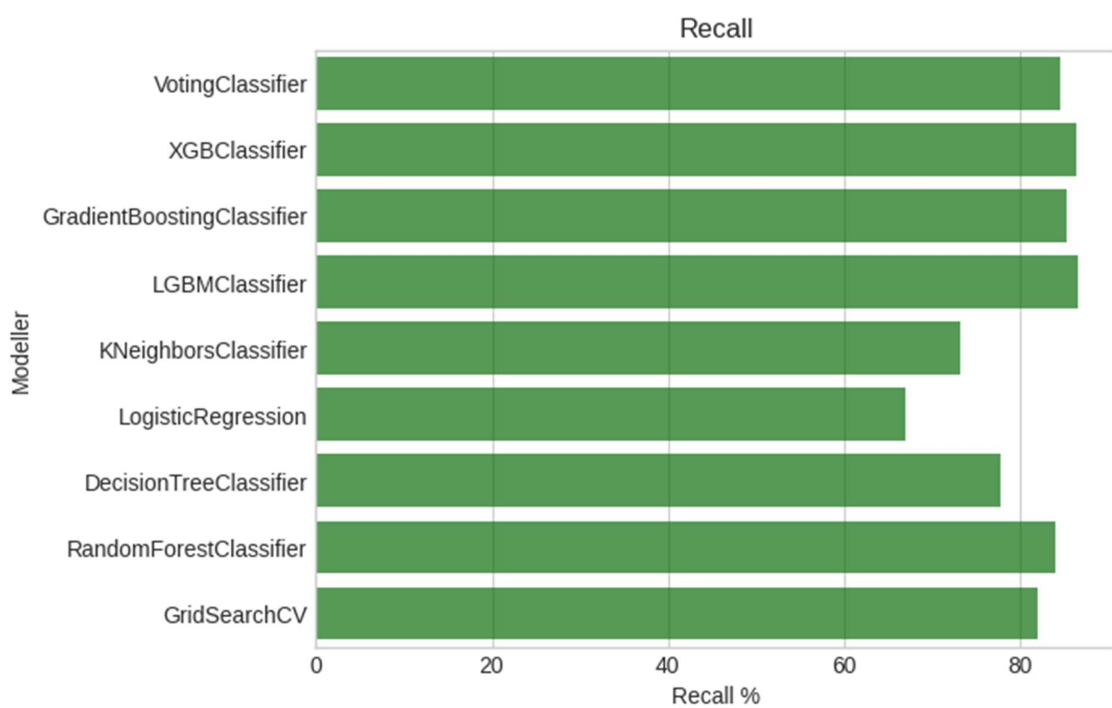


Figure 4.12. Comparison of recall rates of algorithms

Upon examining the confusion matrix of LightGBM, the most successful algorithm, it was observed that highly accurate prediction results were obtained for all risk classes in both the test and train datasets (as depicted in Figures 4.13 and 4.14). An excellent tool for evaluating a classification model's performance is the confusion matrix. It offers a comprehensive breakdown of both correct and incorrect predictions for each class, enabling researchers to evaluate the model's accuracy, precision, recall, and other performance metrics. The excellent performance demonstrated by LightGBM in both the test and train datasets across all risk classes indicates its robustness and effectiveness in accurately predicting seismic risk situations for the buildings in the study. Consequently, these promising results reinforce the credibility of the approach and highlight the potential utility of LightGBM for three seismic risk assessment and urban transformation decisions.

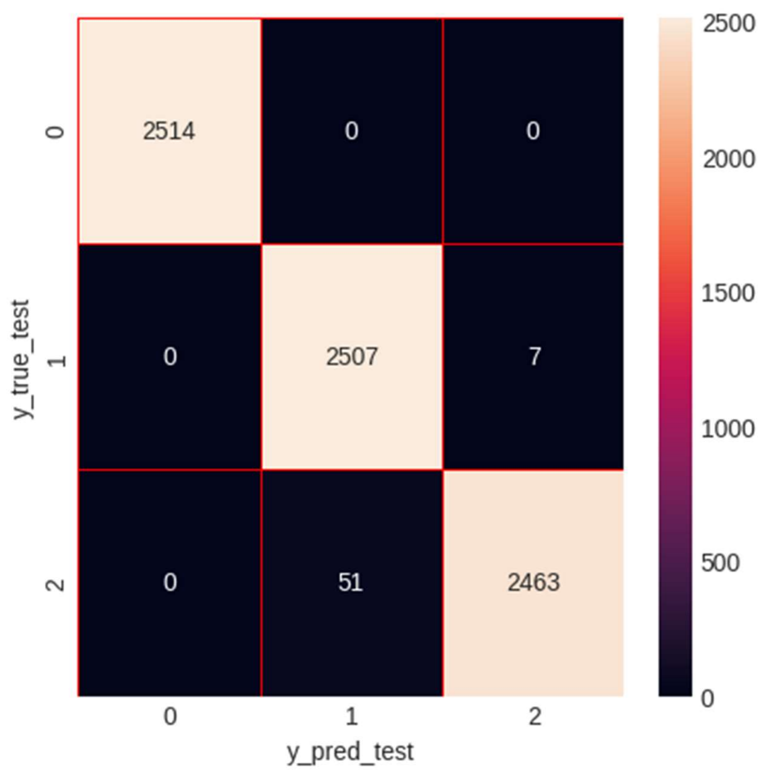


Figure 4.13. LightGBM confusion matrix result train

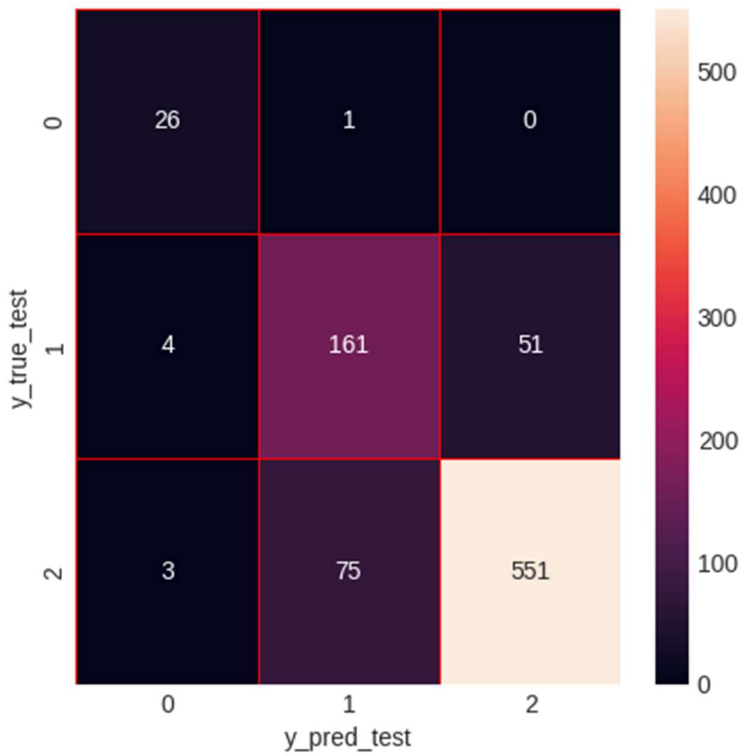


Figure 4.14. LightGBM confusion matrix result test

4.1.3. Experimental results of 4 option

The research involved the analysis of a dataset containing 4,356 buildings using a variety of machine-learning algorithms. All calculations and analyses were conducted using the Google Colaboratory framework, which utilizes Python programming. To begin, the dataset was initially split into two sets: a training set consisting of 4356 samples and a test set containing 872 samples. This division was achieved using the "Train Test Split" function from the "Sklearn model selection" library, ensuring that the data was appropriately divided for training and evaluation purposes. To categorize the buildings based on their seismic risk levels, the "cut()" method from two different Pandas libraries was utilized. The method facilitated the creation of four classes within the "rvs" parameter, effectively grouping the data. By implementing this technique, estimates were generated for the dataset, which was then divided into four risk classes: "non-risk," "medium risky," "risky," and "high risky." The limit RVS values that define the risk classes are detailed in Table 4.5.

In summary, the study involved analyzing the dataset using various machine learning algorithms within the Google Colaboratory framework. The data was divided into training and test sets, and the "cut()" method was employed to group the data into four distinct risk classes based on the RVS parameter values. Table 4.5 provides essential information regarding the RVS thresholds defining the risk classes. These findings offer valuable insights for seismic risk assessment and urban transformation decisions, providing a comprehensive classification scheme for buildings based on their level of seismic risk.

Table 4.5. RVS risk classification thresholds

Risk Status	Classification	Threshold
Non-Risky	0	0-0,25
Medium Risky	1	0,25-0.50
Risky	2	0.50-0.75
High Risky	3	0.75-1

With this number of classification, urban transformation costs will be reduced to the minimum and the success rates will be kept at the lowest level. This means a rapid risk transformation for an endangered area. Moreover, K-fold cross-validation was employed during the validation stage to mitigate the risk of overfitting the algorithm to the training dataset. As previously mentioned, this study utilized various algorithms, including Mars, logistic regression, decision tree classifier, gradient boosting, LightGBM, random forest classifier, support vector machine classifier, and K-nearest neighbors classifier.The accuracy of each method and performans metrics is presented in Table 4.6.

Table 4.6. Results of all machine learning methods

Method	Training Error Accuracy (%)	Testing Error Accuracy (%)
Logistic regression (LR)	57.45	49.06
Decision tree classifier (DTC)	68.69	77.09
Random forest classifier (RFC)	99.64	75.68
Voting Classifier (VC)	98.95	77.6
Mlp Classifier (Mlp)	90	75.2
Gradient Boosting (GB)	100	77.6
Extreme Gradient Boosting (XGB)	100	79.01
Light GBM (LGB)	98.05	78.2
Mars and VC with Grid Search (MVC)	99.69	73.74
KNeighborsClassifier (KNN)	92.61	62.84

	VC				XGB				LightGBM			
	TS=0	TS=1	TS=2	TS=3	TS=0	TS=1	TS=2	TS=3	TS=0	TS=1	TS=2	TS=3
TS=0	20	3	1	2	22	2	1	1	22	2	1	1
TS=1	0	61	37	28	3	63	31	29	6	63	32	25
TS=2	0	24	52	45	0	29	52	40	0	32	53	36
TS=3	2	24	29	544	1	20	26	552	6	35	24	534
Precision (%)	0.77				0.78				0.78			
Recall (%)	0.77				0.79				0.78			
Fscore (%)	0.77				0.79				0.78			

When we compare all models with evaluation metrics and divide the seismic risk situations of structures into 4 classes, XGB is the algorithm that makes the most successful estimations. Figure (4.15,4.16,4.17)

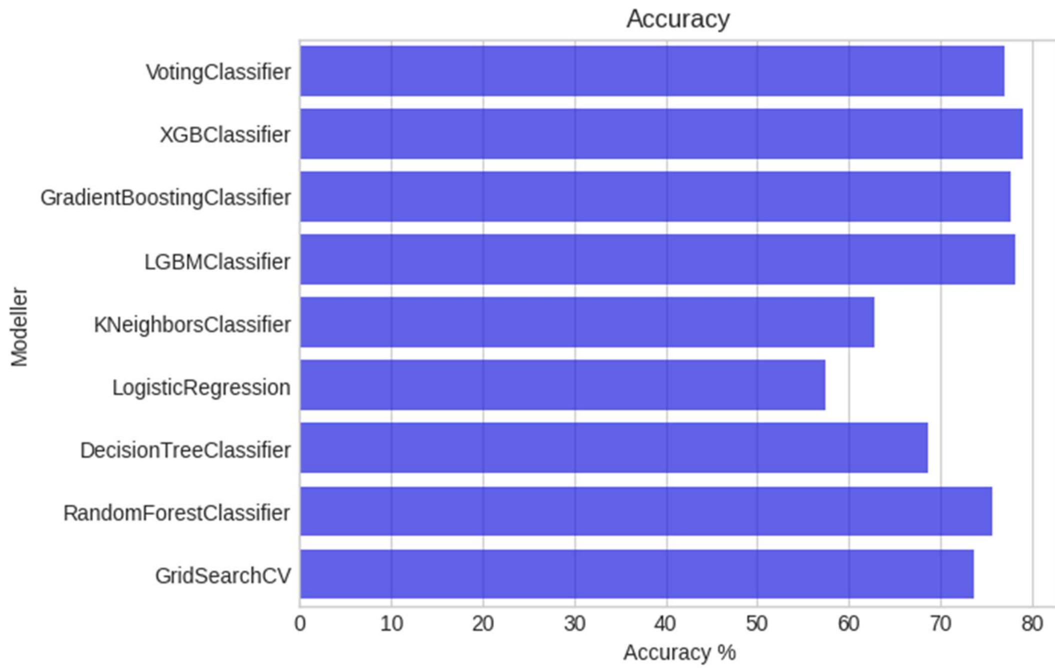


Figure 4.15. Comparison of accuracy rates of algorithms

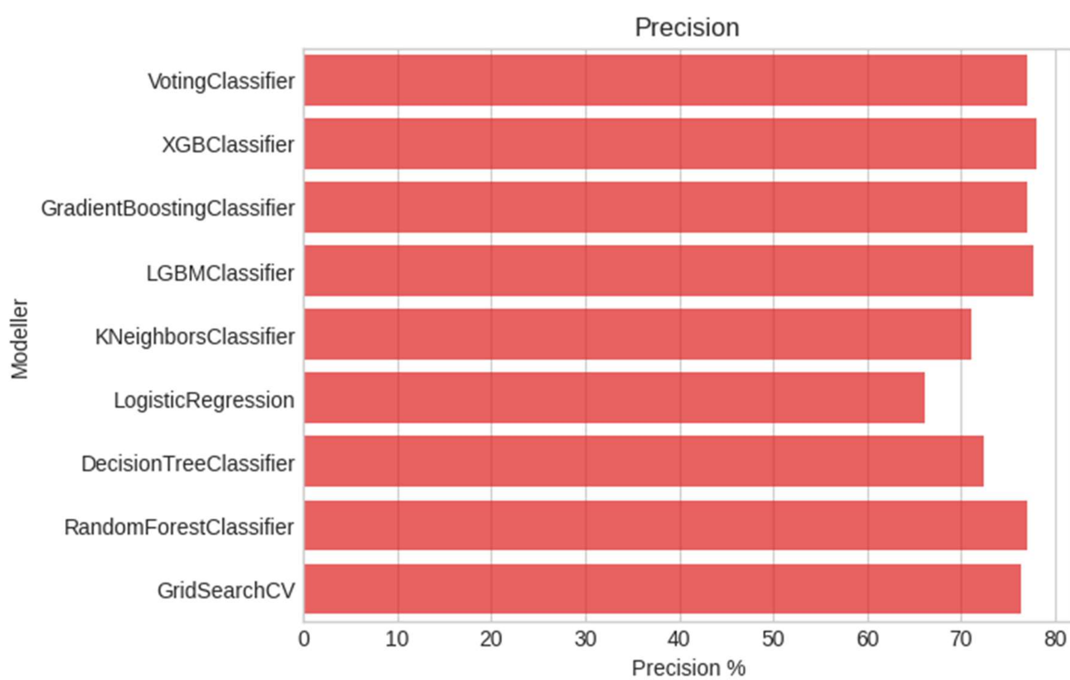


Figure 4.16. Comparison of precision rates of algorithms

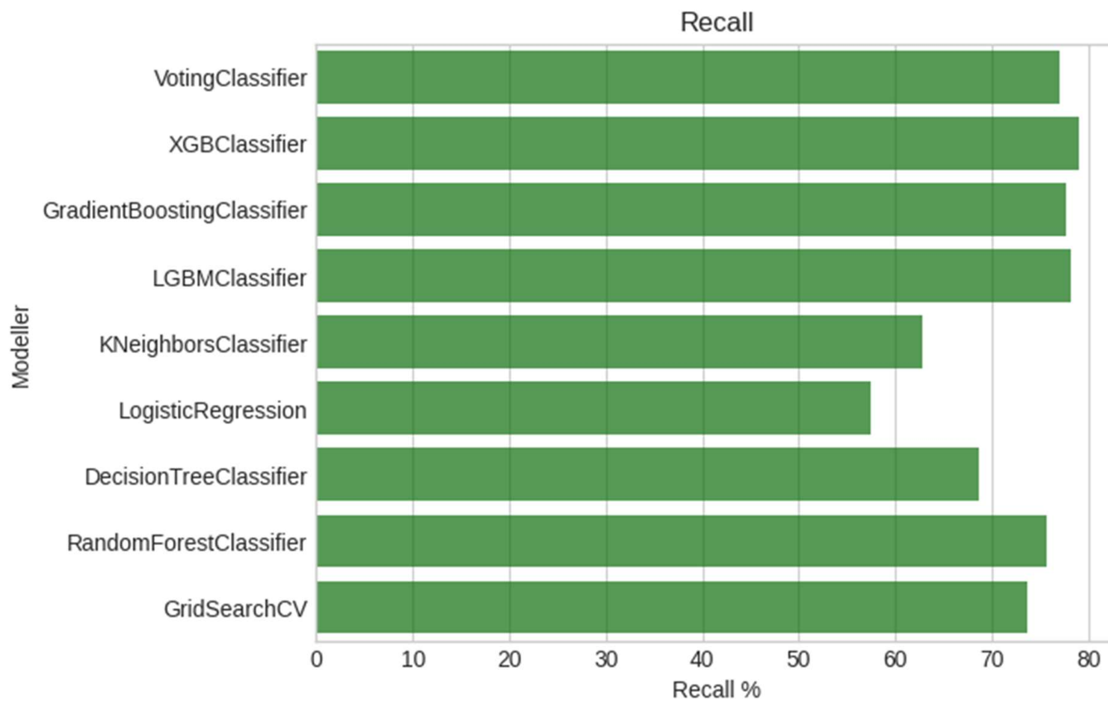


Figure 4.17. Comparison of recall rates of algorithms

Upon examining the confusion matrix of Extreme Gradient Boosting (XGBoost), which is identified as the most successful algorithm, it was evident that highly accurate prediction results were achieved in both the test and train datasets across all risk classes (as illustrated in Figures 4.18 and 4.19). For each risk class, the confusion matrix offers comprehensive information on the model's prediction accuracy, including true positive, true negative, false positive, and false negative values. The high accuracy of XGBoost in both test and train datasets for all risk classes signifies its exceptional performance in accurately classifying buildings based on their seismic risk levels.These findings reinforce the efficacy of XGBoost as a powerful machine learning algorithm for seismic risk assessment. The robustness and consistency of its accurate predictions across different risk classes make it a promising choice for practical applications in urban transformation decisions and seismic risk mitigation strategies.

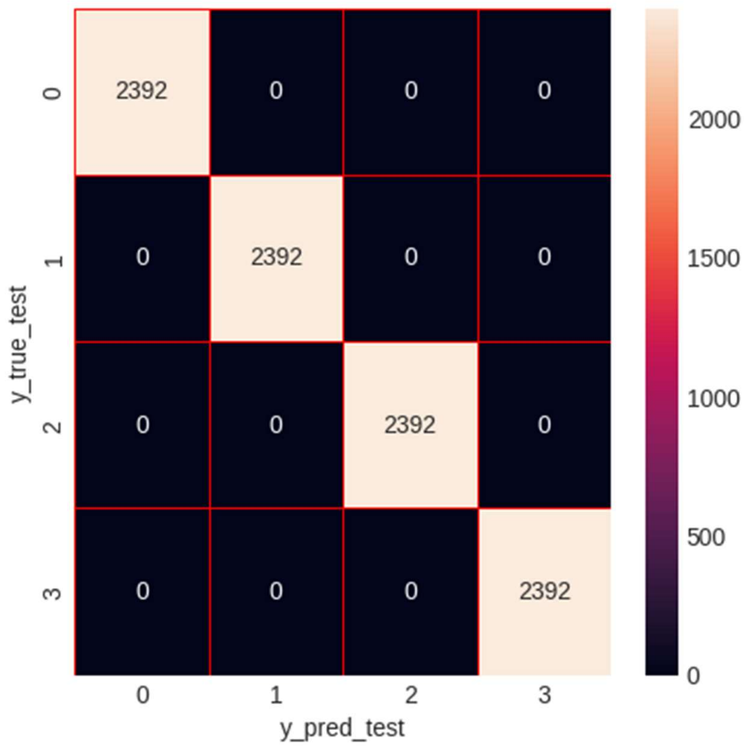


Figure 4.18. Extreme gradient boosting algorithm confusion matrix result train

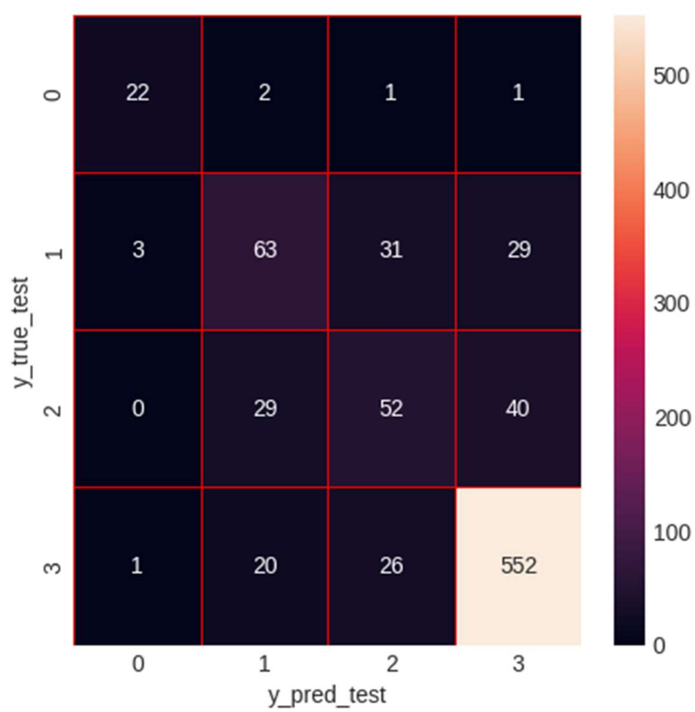


Figure 4.19. Extreme gradient boosting algorithm confusion matrix result test

5. SUMMARY OF ML-BASED SEISMIC RISK PRIORITIZATION METHOD

In light of the findings from the study, the following steps can be applied when computing the rapid seismic analysis results for a group of buildings (Figure 6.1):

Data Preparation: Gather the necessary data for the buildings in the group, including relevant parameters that influence seismic risk.

Data Preprocessing: Ensure that the data is properly cleaned, normalized, and preprocessed before feeding it into the selected algorithm. (i.e Feature Extraction and Dimension Reduction)

Algorithm Selection: Depending on the desired classification of seismic risk, choose the appropriate algorithm. For classifying seismic risk into two degrees, the Voting Classifier algorithm is recommended. For three degrees, use the lightGBM algorithm, and for four degrees, use Extreme Gradient Boosting (XGBoost). The balance of economy or success rate will be important for the selection to be made in this section.

Techniques Prevent Overfitting: Hyperparameter Tuning, Model Evaluation with Cross-Validation and Grid Search with Cross-Validation:

Model Training: Train the chosen algorithm with the prepared training dataset. This process entails adjusting the model to understand the connections between the input parameters and the seismic risk categories. If there is an imbalance in the outcome variables within the dataset, balance the training dataset by employing synthetic data generation methods, such as the SMOTE method.

Model Evaluation: Evaluate the trained model's performance using cross-validation methods or the test dataset. Assess the model's performance using a variety of metrics, such as accuracy, precision, recall, F1-score, and other pertinent indicators.

Deployment and Prediction: After the model is trained and assessed, it can be put into operational use to provide swift seismic analysis predictions for new groups of buildings. Provide the necessary input parameters, and the model will classify the buildings into their respective seismic risk classes (two, three, or four degrees).

Result Interpretation: Interpret the model's predictions and assess the seismic risk of each construction in the group based on the assigned risk class.

By following these steps and using the appropriate algorithm for the desired classification, one can efficiently perform rapid seismic analysis for a building group, providing valuable insights for decision-making in urban transformation and seismic risk mitigation efforts.

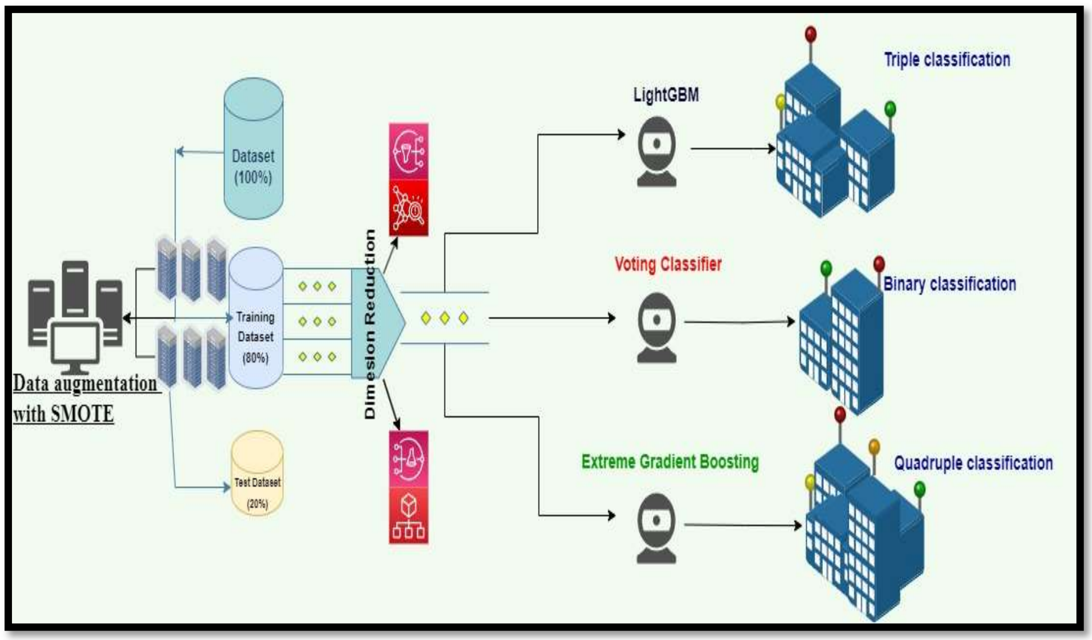


Figure 5.1. Unified learning framework working chart

6. DISCUSSION and CONCLUSIONS

6.1. Discussion of Results

The existing rapid screening methods in the literature typically assess earthquake risk in large building stocks by penalizing certain structural deficiencies without a calculated correlation to detailed seismic risk analysis. In this novel approach, a significant advancement is introduced to the rapid screening method. It leverages a large database to establish a correlation between seismic risk analysis results and rapid screening scores. Unlike conventional rapid screening methods found in the literature, this proposed approach enables the creation of a risk distribution map for extensive building stocks that exhibits a substantial correlation with detailed seismic risk analysis results. As a result, the machine learning network introduced in this study can be effectively utilized in the development of seismic risk mitigation systems. The reliability of the proposed machine learning network is further strengthened through an evaluation of its performance using both extensive datasets and numerical seismic analysis results.

The performance of the proposed network is enhanced by incorporating ensemble learning, relying on both the hard voting scheme and hybrid methods. Increasing with SMOTE the number of low-risk buildings applied to the training data set contributed to the success rate of the study. In addition, by getting rid of unnecessary parameters with dimension reduction methods, the success rate has been increased as well as speeding up algorithms and field detections. In other words, These manipulations significantly improved the success rate of estimations for both the training and test databases. This is because; risk score estimation performance (i.e., percentage of the failed elements or percentage of the shear capacities of failed specimens, etc.) as the proposed network was detected to have some deficiencies regarding the seismic risk score estimations. There are studies that classify buildings as risky and non-risky before, but there is no study that divides the risk class into 3 or 4. When the results of the study, in which the risk class is divided into 3, are analyzed, only 1 out of 27 non-risky structures is estimated as medium risk and this is calculated in seconds without detailed analysis. This means a success rate of over 98% in non-risky structures. In addition, it has a success rate of over 75%, even in the estimation of medium-risk structures with the lowest success rate. In addition, only 4 out of 216 medium-risk constructs were labeled as risk-free. This shows that even when the algorithm is faulty, it provides over 98% success in naming a risky structure as non-risky. This observation also boosted confidence in the proposed network. However, when considering its actual application, the scenario appears different.

The outcomes attained within the purview of the study can be compared with the damage results obtained after the earthquake. For example, a risky structure can be used to predict a heavily damaged structure after an earthquake, while a medium-risk structure for a moderately damaged structure can be used to predict a no-risk, low-east or undamaged structure. Since the success rate of the MARS integrated hybrid system is low, it was concluded that this method should not be used in estimating the seismic risk level of masonry buildings.

6.2. Conclusions

Identifying the most vulnerable buildings and taking preventive measures before major earthquakes occur is crucial for saving lives. Nevertheless, the current detailed assessment procedures are demanding in terms of resources and are not well-suited for evaluating a large number of buildings. Moreover, current rapid visual screening methods used to estimate the seismic risk of large building stocks often lack reliability and accuracy. In light of these challenges, this study aimed to introduce a fresh perspective on filtering buildings based on their seismic risk. To achieve this, a computational network was developed, which could assess the risk of structures using external observations. This network was trained using data from known detailed seismic risk analyses. Subsequently, the algorithm's performance in estimating seismic risk was evaluated on buildings that were not part of the training dataset.

In this study, the approach was to not rely solely on a single method for estimation but rather to employ ensemble learning and hybrid methods to achieve the highest level of success. In the context of ensemble learning, a range of distinct supervised machine learning algorithms were employed. These algorithms included logistic regression, decision tree classifier, random forest classifier, support vector machine classifier, K-neighbors classifier, MlpClassifier, GradientBoosting, Mars, and LightBossting, all utilized for predicting building damage levels.Then, hard voting was utilized as the outcome of the proposed method was designed to be composed of risky and non-risky buildings (i.e., 1-0 or 1-0-2, or 1-0-2-3). Unfortunately, recent earthquakes have shown that most of the structures in Turkey are structures that have not received engineering service. Specifically, on February 6, 2023, two earthquakes, one with a magnitude of 7.8 Mw (± 0.1) and the other with a magnitude of 7.5 Mw, occurred just nine hours apart. These earthquakes had epicenters in the Pazarcık and Ekinözü districts of Kahramanmaraş, and they caused extensive damage to most of the structures in the area. Considering this point, within the purview of the study, buildings were not only classified as risky/non-risky, but also prioritized for the number of risky buildings by classifying risky structures (according to Ve/Vr estimates). This method will be able to provide a solution to the risky building detection, which is very costly to be done all at once in Turkey, and which of these risky buildings should be firstly demolished. By using the method in the study, risky buildings can be divided into 2 or 3 or 4 classes and the selection of the most critical structures that need to be demolished immediately can be performed with high accuracy. The success rates of the proposed ensemble learning for the training and test databases were 88.36%, 86.58%, and 79.01%, respectively. As mentioned above, the method that we classify as the Hybrid method is Reinforced Learning. It classifies all its solutions with reinforcement learning algorithms like Random Forest Classifier with feature importance to increase the success rate of a particular problem type. In addition, by combining the most used Regression, Gradient Boosting algorithms and Ensemble Learning algorithms in the literature, it has achieved a classification and success rate that has not been presented before. In total, they observe real risk-free observations, including the positive contributions they have accumulated.

This procedure is used to find, list, and rank the most vulnerable buildings in a given area that might sustain damage in the event of an earthquake. This makes the approach crucial for identifying and fortifying Turkey's weak structures as well as for reducing the amount of casualties and property damage.

In addition, more accurate results can be obtained with a more precise assessment of the years of manufacture for the determination of the engineering service and building inspection service that the residence has received since the recent past in Turkey.
This approach could be more effective if mobile apps or web-based software are developed to facilitate on-site data input.

Automation systems can be established by determining the parameters of Duronla buildings using a vision computer, and then estimating the seismic risk analysis result with this method.

In recent studies, it has been claimed that ADASYN (an innovative, adaptive synthetic sampling methodology designed to extract information from unbalanced data sets) could be utilized to generate more accurate synthetic data. Since ADASYN can produce more synthetic data for minority class examples that are harder to learn than for minority examples that are easier to learn (Haibo et al. 2008). In future studies, SMOTE could be replaced with ADASYN to compare the performance of the machine-learning networks.

7. REFERENCES

Achs, G., & Adam, C. (2012). Rapid seismic evaluation of historic brick-masonry buildings in Vienna (Austria) based on visual screening. *Bulletin of Earthquake Engineering*, *10*(6), 1833–1856. https://doi.org/10.1007/s10518-012-9376-5

Aldemir, A., Guvenir, E., & Sahmaran, M. (2020). Rapid screening method for the determination of regional risk distribution of masonry structures. *Structural Safety*, *85*, 101959. https://doi.org/10.1016/J.STRUSAFE.2020.101959

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bholowalia, P., & Kumar, A. (2014). EBK-means: A clustering technique based on the elbow method and k-means in WSN. *International Journal of Computer Applications*, *105*(9).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2020). *Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning*. Machine Learning Mastery.

Cardinali, V., Tanganelli, M., & Bento, R. (2022). A hybrid approach for the seismic vulnerability assessment of the modern residential masonry buildings. *International Journal of Disaster Risk Reduction*, *79*, 103193. https://doi.org/10.1016/j.ijdrr.2022.103193

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, C., Zhang, Q., Ma, Q., & Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemometrics and Intelligent Laboratory Systems*, *191*, 54–64. https://doi.org/10.1016/j.chemolab.2019.06.003

Cooner, A., Shao, Y., & Campbell, J. (2016). Detection of Urban Damage Using Remote Sensing and Machine Learning Algorithms: Revisiting the 2010 Haiti Earthquake. *Remote Sensing*, *8*(10), 868. https://doi.org/10.3390/rs8100868

Coskun, O., & Aldemir, A. (2022). Machine learning network suitable for accurate rapid seismic risk estimation of masonry building stocks. *Natural Hazards*. https://doi.org/10.1007/S11069-022-05553-Y

Coskun, O., Aldemir, A., & Sahmaran, M. (2020). Rapid screening method for the determination of seismic vulnerability assessment of RC building stocks. *Bulletin of Earthquake Engineering*, *18*(4), 1401–1416. https://doi.org/10.1007/s10518-019-00751-9

Demertzis, K., Kostinakis, K., Morfidis, K., & Iliadis, L. (2023). An interpretable machine learning method for the prediction of R/C buildings' seismic response. *Journal of Building Engineering*, *63*, 105493. https://doi.org/10.1016/j.jobe.2022.105493

Deng, X., Liu, Q., Deng, Y., & Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, *340–341*, 250–261. https://doi.org/10.1016/j.ins.2016.01.033

Federal Emergency Management Agency. (1998). *FEMA 155-ATC-21-1, Rapid Visual Screening of Buildings for Potential Seismic Hazards: Supporting Documentation,*.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

Geiß, C., Aravena Pelizari, P., Marconcini, M., Sengara, W., Edwards, M., Lakes, T., & Taubenböck, H. (2015). Estimation of seismic building structural types using multi-sensor remote sensing and machine learning techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, *104*, 175–188. https://doi.org/10.1016/j.isprsjprs.2014.07.016

Geon Yong Park, Heeseong Kim, Hwi Woon Jeong, & Hee Yong Youn. (2013). A Novel Cluster Head Selection Method based on K-Means Algorithm for Energy Efficient Wireless Sensor Network. *2013 27th International Conference on Advanced Information Networking and Applications Workshops*, 910–915. https://doi.org/10.1109/WAINA.2013.123

Halûk Sucuoğlu. (2007). *Kentsel Yapı Stoklarında Deprem Risklerinin Sokaktan Tarama Yöntemi ile Belirlenmesi*.

Hassan, A. F., & Sozen, M. A. (1997). Seismic vulnerability assessment of low-rise buildings in regions with infrequent earthquakes. *ACI Structural Journal*, *94*(1), 31–39.

Hilbe, J. M. (2009). *Logistic regression models*. CRC press.

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, *37*(2.5), 3.

Kaur, G. (2020). A comparison of two hybrid ensemble techniques for network anomaly detection in spark distributed environment. *Journal of Information Security and Applications*, *55*, 102601. https://doi.org/10.1016/j.jisa.2020.102601

Kerim KOÇ. (2022). *DEVELOPMENT OF A MACHINE LEARNING PREDICTION MODEL FOR CONSTRUCTION SAFETY MANAGEMENT*. YILDIZ TECHNICAL UNIVERSITY.

Kerle, N., Nex, F., Gerke, M., Duarte, D., & Vetrivel, A. (2019). UAV-Based Structural Damage Mapping: A Review. *ISPRS International Journal of Geo-Information*, *9*(1), 14. https://doi.org/10.3390/ijgi9010014

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kumari, V., Harirchian, E., Lahmer, T., & Rasulzade, S. (2022). Evaluation of Machine Learning and Web-Based Process for Damage Score Estimation of Existing Buildings. *Buildings*, *12*(5), 578. https://doi.org/10.3390/buildings12050578

LI, C., ZHANG, G., LEI, T., & GONG, A. (2011). Quick image-processing method of UAV without control points data in earthquake disaster area. *Transactions of Nonferrous Metals Society of China*, *21*, s523–s528. https://doi.org/10.1016/S1003-6326(12)61635-5

Li, L., Liu, X., Chen, Q., & Yang, S. (2018). Building damage assessment from PolSAR data using texture parameters of statistical model. *Computers & Geosciences*, *113*, 115–126. https://doi.org/10.1016/j.cageo.2018.01.018

Luo, M., Xie, J., Yan, Y., Ke, Z., Yu, P., Wang, Z., & Zhang, J. (2020). Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II. *Energy and Buildings*, *210*, 109776. https://doi.org/10.1016/j.enbuild.2020.109776

M. T. Beyhan. (2023). *A MACHINE LEARNING APPROACH TO SEISMIC RISK ASSESSMENT OF REINFORCED CONCRETE STRUCTURES*. Middle East Technical University.

Macedo, F., Rosário Oliveira, M., Pacheco, A., & Valadas, R. (2019). Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing*, *325*, 67–89. https://doi.org/10.1016/j.neucom.2018.09.077

Mehmet TAN. (2022). *DEVELOPMENT OF A MACHINE LEARNING BASED FRAMEWORK FOR FAILURE FORECAST AND SHEAR STRENGTH ESTIMATION OF CONVENTIONAL RC SHEAR WALLS*. ISTANBUL TECHNICAL UNIVERSITY .

Milosevic, J., Cattari, S., & Bento, R. (2020). Definition of fragility curves through nonlinear static analyses: procedure and application to a mixed masonry-RC building stock. *Bulletin of Earthquake Engineering*, *18*(2), 513–545. https://doi.org/10.1007/s10518-019-00694-1

Muhammet Yuşa EKİCİ. (2022). *MEVCUT BETONARME BİNALARIN HIZLI SİSMİK DEĞERLENDİRİLMESİNDE KULLANILAN BAZI GEOMETRİK PARAMETRELERİN ÖN EĞİTİMLİ KONVOLÜSYONEL SİNİR AĞLARI İLE BELİRLENMESİ*. KONYA TEKNİK ÜNİVERSİTESİ.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sensing of Environment*, *86*(4), 554–565. https://doi.org/10.1016/S0034-4257(03)00132-9

Qu, Z., Xu, J., Wang, Z., Chi, R., & Liu, H. (2021). Prediction of electricity generation from a combined cycle power plant based on a stacking ensemble and its hyperparameter optimization with a grid-search method. *Energy*, *227*, 120309. https://doi.org/10.1016/j.energy.2021.120309

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, *28*(1), 71–72.

Joseph, V. R., Optimal ratio for data splitting, Stat. Anal. Data Min.: ASA Data Sci. J.. 15 (2022), 531–538. https://doi.org/10.1002/sam.11583

Rajarathnam, S., & Santhakumar, A. R. (2015). Assessment of seismic building vulnerability based on rapid visual screening technique aided by aerial photographs on a GIS platform. *Natural Hazards*, *78*(2), 779–802. https://doi.org/10.1007/s11069-014-1382-2

Rashidi, A., Sigari, M. H., Maghiar, M., & Citrin, D. (2016). An analogy between various machine-learning techniques for detecting construction materials in digital images. *KSCE Journal of Civil Engineering*, *20*(4), 1178–1188. https://doi.org/10.1007/s12205-015-0726-0

R.M., S. P., Maddikunta, P. K. R., M., P., Koppu, S., Gadekallu, T. R., Chowdhary, C. L., & Alazab, M. (2020). An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. *Computer Communications*, *160*, 139–149. https://doi.org/10.1016/j.comcom.2020.05.048

Ruggieri, S., Cardellicchio, A., Leggieri, V., & Uva, G. (2021). Machine-learning based vulnerability analysis of existing buildings. *Automation in Construction*, *132*, 103936. https://doi.org/10.1016/j.autcon.2021.103936

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

Santos, A. R., Santos, M. A., Baumbach, J., McCulloch, J. A., Oliveira, G. C., Silva, A., Miyoshi, A., & Azevedo, V. (2011). A singular value decomposition approach for improved taxonomic classification of biological sequences. *BMC Genomics*, *12*(S4), S11. https://doi.org/10.1186/1471-2164-12-S4-S11

Sewell, M. (2008). Ensemble learning. *RN*, *11*(02), 1–34.

Shah, M. F. and A. A. and K.-B. O. K. (2016). *A Case Study Using Rapid Visual Screening Method to Determine the Vulnerability of Buildings in two Districts of Jeddah, Saudi Arabia*. Proceedings of the 15th international symposium on new technologies for urban safety of mega cities in Asia, Tacloban, Philippines.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2019). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, *52*, 456–462. https://doi.org/10.1016/j.bspc.2017.01.012

S.Imandoust and M.Bolandraftar. (2016). Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. *S B Imandoust et al. Int. Journal of Engineering Research and Applications*, *3*(5), 605–610.

Sözen, M. A. (2014). *Surrealism in Facing the Earthquake Risk* (pp. 1–13). https://doi.org/10.1007/978-3-319-00458-7_1

Stepinac, M., & Gašparović, M. (2020). A Review of Emerging Technologies for an Assessment of Safety and Seismic Vulnerability and Damage Detection of Existing Masonry Structures. *Applied Sciences*, *10*(15), 5060. https://doi.org/10.3390/app10155060

Sublime, J., & Kalinicheva, E. (2019). Automatic Post-Disaster Damage Mapping Using Deep-Learning Techniques for Change Detection: Case Study of the Tohoku Tsunami. *Remote Sensing*, *11*(9), 1123. https://doi.org/10.3390/rs11091123

Standard for Seismic Evaluation of Existing Reinforced Concrete Buildings, Technical Manual for Seismic Evaluation and Seismic Retrofit of Existing Reinforced Concrete Buildings (2001).

Velliangiri, S., Alagumuthukrishnan, S., & Thankumar joseph, S. I. (2019). A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science*, *165*, 104–111. https://doi.org/10.1016/j.procs.2020.01.079

Wang, C., Yu, Q., Law, K. H., McKenna, F., Yu, S. X., Taciroglu, E., Zsarnóczay, A., Elhaddad, W., & Cetiner, B. (2021). Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management. *Automation in Construction*, *122*, 103474. https://doi.org/10.1016/j.autcon.2020.103474

Wang, X., Wittich, C. E., Hutchinson, T. C., Bock, Y., Goldberg, D., Lo, E., & Kuester, F. (2020). Methodology and Validation of UAV-Based Video Analysis Approach for Tracking Earthquake-Induced Building Displacements. *Journal of Computing in Civil Engineering*, *34*(6). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000928

Xu, J. Z., Lu, W., Li, Z., Khaitan, P., & Zaytseva, V. (2019). *Building Damage Detection in Satellite Imagery Using Convolutional Neural Networks*.

Xu, Z., Wu, L., & Zhang, Z. (2018). Use of active learning for earthquake damage mapping from UAV photogrammetric point clouds. *International Journal of Remote Sensing*, *39*(15–16), 5568–5595. https://doi.org/10.1080/01431161.2018.1466083

Xue, P., Jiang, Y., Zhou, Z., Chen, X., Fang, X., & Liu, J. (2019). Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms. *Energy*, *188*, 116085. https://doi.org/10.1016/j.energy.2019.116085

Yakut, A. (2004). Preliminary seismic performance assessment procedure for existing RC buildings. *Engineering Structures*, *26*(10), 1447–1461. https://doi.org/10.1016/j.engstruct.2004.05.011

Yu, Q., Wang, C., McKenna, F., Yu, S. X., Taciroglu, E., Cetiner, B., & Law, K. H. (2020). Rapid visual screening of soft-story buildings from street view images using deep learning classification. *Earthquake Engineering and Engineering Vibration*, *19*(4), 827–838. https://doi.org/10.1007/s11803-020-0598-2

Yücemen, M. S., Özcebe, G., & Pay, A. C. (2004). Prediction of potential damage due to severe earthquakes. *Structural Safety*, *26*(3), 349–366. https://doi.org/10.1016/j.strusafe.2003.09.002

Zhang, W., & Goh, A. T. C. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, *7*(1), 45–52. https://doi.org/10.1016/j.gsf.2014.10.003

Zhang, Y., Burton, H. v., Sun, H., & Shokrabadi, M. (2018). A machine learning framework for assessing post-earthquake structural safety. *Structural Safety*, *72*, 1–16. https://doi.org/10.1016/j.strusafe.2017.12.001

Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks, Hong Kong, 2008, pp. 1322-1328, https://doi.org/10.1109/IJCNN.2008.4633969

APPENDICES

APPENDIX 1- Publications Derived from the Thesis and Code

Natural Hazards (2023) 115:261–287
https://doi.org/10.1007/s11069-022-05553-y

ORIGINAL PAPER

Check for updates

Machine learning network suitable for accurate rapid seismic risk estimation of masonry building stocks

Onur Coskun[1] · Alper Aldemir[2]

Received: 26 April 2022 / Accepted: 1 August 2022 / Published online: 17 August 2022

Abstract

Most losses from earthquakes are associated with fully collapsed buildings. So, determining the seismic risk of buildings is essential for building occupants in active earthquake zones. Unfortunately, current methods used to estimate the risk state of large building stocks are insufficient for reliable, fast, and accurate decision-making. In addition, the risk classifications of buildings after major natural disasters depend entirely on the experience of the technical team of engineers. Therefore, the decision on risk distributions of building stocks before and after hazards requires more sustainable and accurate methods that include other means of technological advancement. In this study, the building characteristics dominating the seismic risk outcome were determined using a database of 543 masonry buildings. Later, for the first time in the literature, a new, fast and accurate seismic evaluation method is proposed. The proposed method is thoroughly associated with detailed evaluation results of structures with the help of machine learning algorithms. This study utilized an approach in which six machine learning algorithms work together (i.e., Logistic Regression, Decision Tree, Random Forest, K-Mean Clustering, Support Vector Machine, and Ensemble Learning Method). As a result of the analysis of these algorithms, the correct prediction rates for the learning database (i.e., 434 buildings) and the test database (i.e., 109 buildings) of the proposed method were determined as approximately 96.67% and 95%, respectively. Lastly, machine learning algorithms trained by structures with known after seismic risk results are developed. The proposed method managed to classify risk states with the accuracy of 84.6%.

Keywords Seismic risk estimations · Masonry structures · Machine learning · Seismic risk classification

✉ Alper Aldemir
alperaldemir@hacettepe.edu.tr

Onur Coskun
onur.coskun.dr@gmail.com

1 Graduate School of Science and Engineering, Hacettepe University Beytepe, Ankara, Turkey

2 Department of Civil Engineering, Hacettepe University, Ankara, Turkey

Springer

1 Introduction

The risk of heavy damage or failure of buildings during a seismic event is a significant concern for communities, as the affected area is generally large. For this reason, determining the earthquake risk of structures has become an essential issue among structural engineers (FEMA356). Most losses from earthquakes are associated with the local or total collapse of buildings. In this context, continuous assessment and monitoring of buildings' seismic safety and vulnerability are challenging, especially when extensive area assessments are required. In addition, examining structural cracks or determining the wall type in a masonry building damaged due to earthquake events is a very dangerous task for site engineers. For this reason, it has become possible to determine the seismic risk of buildings by observations that can be made from outside the building. But, the construction industry is one of the slowest professions to adapt to new technologies, and this situation needs to change. In this study, a contemporary method was proposed to accurately predict the seismic performance of buildings with the help of machine learning algorithms eliminating the need for any technical personnel to enter the building.

The methodologies designed to determine the seismic risk of individual buildings are well matured. However, they necessitate the use of complex analysis tools along with detailed inputs like material testing, plan drawings, quality of material, etc. (FEMA2456, Eurocode 6, TEC2018, GABHR 2019; Dejong 2009; Aldemir et al. 2013; Penna et al. 2014; Beyer et al. 2014; Penna 2015 and Ahmad and Ali 2017). These detailed seismic assessment techniques become dysfunctional if the number of concerned buildings reaches thousands or more. This is mainly because the inputs and procedures take significant time, manpower of experienced engineers, and computational power. Therefore, these methods could not be employed when the seismic risk of the large building stock is aimed to be determined. Consequently, the current state-of-the-art on detailed seismic assessment of structures is limited by human and infrastructure resources. Therefore, Sozen (2014) stated that the seismic risk assessment of building inventories could only be accomplished by changing the strategy from seeking safety to filtering out vulnerable buildings from the large building stock (i.e., low-pass filtering). Thus, a versatile and accurate method should be formulated to enable decisions to be made using inexpensively acquired building data and the evaluation process to be implemented quickly (Sozen 2014).

Although several researchers have tried to generate simple methods to assess the seismic risk of reinforced concrete (RC) structures (Yakut 2004; Yucemen and Ozcebe 2004; Askan and Yucemen 2010; Maziliguney et al. 2012; Al-Nimry et al. 2015; Perrone et al. 2015; Kumar et al. 2017; FEMA P154; Coskun et al. 2020; Harirchian et al. 2020a; b); the literature shows a limited number of efforts to propose rapid screening (or filtering) methods applicable to unreinforced masonry (URM) building stocks (GABHR 2019; D'Ayala 2013; Shah et al. 2016; Achs and Adam 2012; Grünthal 1998; Achs 2011; Rajarathnam and Santhakumar 2015; Aldemir et al. 2020). D'Ayala (2013) attempted to correlate damage states with fragility curves to determine the seismic vulnerability of masonry structures. However, this method requires the fragility curve for the location of the building, which reduces the applicability of this process, as fragility curves are scarce in number. Shah et al. (2016) used a building classification with respect to the masonry material used, the state of the building, construction quality, building shape irregularity, and the level of earthquake-resistant design. They used the European Macroseismic Scale (Grünthal 1998) and applied defined vulnerability classes (A–F) to determine the risk level of masonry structures. However, the outcome of this method still lacked correlation with actual performance. In another approach, Achs and Adams (2012) proposed a rapid visual screening method that used penalty scores for structural parameters, including seismic hazard, regularity in the plan, regularity in elevation, horizontal stiffness, local failure, secondary structures, soil condition, foundation, and state of preservation. The penalty scores were derived from comprehensive preliminary in situ inspections and measurements of Viennese brick masonry buildings (Achs 2011). Rajarathnam and Santhakumar (2015) used aerial photographs on a geographic information system (GIS) platform to accelerate rapid visual screening. However, none of these methods is based on a large database of masonry structures with detailed seismic assessment results. In other words, all previous methods lack a correlation between rapid screening scores and detailed analysis results. Aldemir et al. (2020) proposed a new rapid visual screening method applicable to masonry structures. They aimed to increase the accuracy of the seismic risk estimation by correlating the rapid visual screening scores with the detailed seismic risk analysis results. To this end, they generated a linear relationship between the risk and the considered parameters. They concluded that this approach resulted in a promising method with some accuracy problems in predicting the test database. The complex relationships between the seismic risk estimations and the selected parameters could be resolved by implementing machine learning algorithms. Similarly, researchers have recently given some effort to incorporate the machine learning algorithms 1- to find a better relationship between the observed damage and seismic events (Mangalathu et al. 2020 and Zhang et al. 2018); 2- to propose methods to predict the fragility curves (Kiani et al. 2019; Ruggieri et al. 2021); 3- to estimate the seismic risk (Zhang et al. 2019; Harirchian et al. 2020a; b). However, none of these studies proposed a versatile seismic risk filtering method for masonry structures incorporating machine learning algorithms. In addition, some recent studies tried to propose contemporary strategies to efficiently evaluate the seismic performance of structures (Javidan and Kim 2022a; b).

Several studies in the literature have succeeded in creating damage maps using unmanned aerial vehicles for post-earthquake damage assessment (Wang et al. 2021; Cooner et al. 2016; Li et al. 2018; Xu et al. 2019; Xu et al. 2018; Sublime and Kalinicheva 2019; Kerle et al. 2020; Stepinac et al. 2020; etc.). These studies have also evolved into studies that include post-earthquake permanent displacement estimations in order to derive new algorithms to accelerate data processing time (Li et al. 2011) and to classify post-earthquake damages (Wang et al. 2020). Finally, studies are carried out to determine the physical properties of buildings and their structural deficiencies, such as soft floors, using unmanned aerial vehicle photographs (Yu et al. 2020). However, none of these studies was designed to be applied before disasters in order to facilitate pre-hazard applications (i.e., to increase preparedness). In addition, machine learning methods have been developed for the estimation of structural systems using photographic data. Geiß et al. (2015) showed in their research that structural systems (masonry, confined masonry, reinforced concrete frame, steel frame, etc.) could be predicted with a high success rate by machine learning methods. In their methods, random forest and support vector machine algorithms are used.

Therefore, this study focused on developing a simple, rapid visual screening method to predict the damage level of masonry buildings using machine learning algorithms. The parameters required for the procedure were aimed to be collectible without the need for the entrance of technical personnel into the risky building. The parameters that could be determined externally are specified as follows:

- Number of Stories (NS)
- Floor system type (FT)

- Visual damage (VD)
- Wall material type (WT)
- Typical story height (TY)
- Vertical irregularity (VI)
- Typical plan area (TA)
- Earthquake zone (EZ)

It should be stated that the age of the structure is not included as an independent parameter in this proposed network. However, the selected parameters managed to correlate the existing physical and mechanical properties with the risk state. Although the age of the building is an important parameter, the visual damage parameter has a significant correlation with the age of the building. Therefore, the proposed network has high accuracy in estimating the risk state. In other words, this study aimed to develop a calculation network based on the properties of any structures from external observations. This network was trained with the known detailed seismic risk analysis results. To this end, machine learning algorithms were trained with a large stock of buildings whose seismic risk analysis results were available (i.e., 543 different real buildings with seismic risk assessment analyses). Then, the seismic risk analysis estimation performance of this algorithm was tested with untrained buildings (i.e., 109 different real buildings with seismic risk assessment analyses). The formed machine learning algorithm estimates the risk and damage level of the analyzed structure during a possible earthquake event.

2 Definition of the URM building database

The database used within the scope of this study was obtained from the Risky Buildings Department of the Ministry of Environment and Urbanization. The earthquake risk analyses of the buildings in the database were determined by the detailed seismic risk analysis calculation method included in the provisions of the Urban Transformation Law No. 6306 (GABHR 2012) or Turkish Earthquake Code (TEC 2007). In this context, the plan drawings, material strengths, etc., of all buildings were available, along with all the physical properties. Therefore, the necessary technical analysis has been done on these structures to train machine learning algorithms. The selected parameters are presented in Table 1. Before using this raw database, data engineering was performed to filter out unnecessary or misleading information by deleting null values, categorizing the selected parameters, etc. In addition, the distribution of parameters is given in Fig. 1. It is known that masonry structures are commonly constructed to have less than four stories in Turkey. However, in some regions of Turkey, the seismicity is low, promoting the use of masonry structures up to 8 stories. In addition, the number of these exceptional cases is low. Thus, the selected database is intentionally formed to have a limited number of masonry buildings with more than four stories (only %25 of the entire database).

The detailed seismic assessment analysis of buildings was performed as per GABHR (2012). In the numerical models, all piers were simulated using 2-node 3-D frame elements, whereas all slabs were modeled with 4-node thin shell elements. In the numerical models, the modulus of elasticity was calculated using the expression (i.e., 200 fm) given in GABHR (2012). In each story, a rigid diaphragm was defined, provided that a reinforced concrete slab existed. After that, a response spectrum analysis was performed under the effect of a reduced design spectrum (i.e., R = 2). The analysis was performed for two orthogonal directions separately. During the response spectrum analysis, 95% of mass

Table 1 Selected Parameters and their definitions

Parameters	Abbreviation	Unit	Selected intervals	Ranges
Number of stories	NS	–	Any integer value	1–10
Floor system Type	FT	–	1: *RC Slab with RC bond beam,* 2: *RC Slab without RC bond beam,* 3: *Others*	1–3 1–4
Earthquake zone	EZ	–	1: *PGA** $\geq$ *0.75 g,* 2: *0.50 g* $\leq$ *PGA* $<$ *0.75 g,* 3: *0.25 g* $\leq$ *PGA* $<$ *0.50 g,* 4: *PGA* $<$ *0.25 g*	
Wall material type	WT	–	1: *Solid clay brick,* 2: *Hollow clay brick,* 3: *Stone,* 4: *Solid concrete block,* 5: *Others*	1–5
Typical story height	TH	m	1: TH $\leq$ *2.4,* 2: *2.4* $<$ *TH* $\leq$ *3.2* 3: TH $>$ *3.2*	1–3
Typical plan area	TA	m^2	1: TA $\leq$ *50,* 2: *50* $<$ *TA* $\leq$ *200* 3: TA $>$ *200*	1–3
Vertical irregularities	VI	–	1: *Yes,* 0: *No*	0–1
Visual damage	VD	–	1: *Yes,* 0: *No*	0–1

*PGA stands for peak ground acceleration

participation was satisfied in each orthogonal direction. In the detailed seismic assessment analysis, the performance limits for slab elements were not calculated. The performance of each pier was determined as Minimum Damage (MD) provided that the pier had enough capacity to resist the reduced design spectrum and gravity demands. On the contrary, the performance of each pier was classified as Collapse Damage (CD) if the pier did not have enough capacity to resist the reduced design spectrum and gravity demands. The capacity of each pier was estimated by considering all the failure modes given in TEC (2018) (i.e., diagonal tension and base sliding). Also, the axial load demands were compared with the axial load capacities of each pier. In these calculations, a correction was made depending on the slenderness ratio. The correction factor was taken as 1 for slenderness ratios less than 6, 0.8 for slenderness ratios between 6 and 10, 0.7 for slenderness ratios between 10 and 15, and 0.5 for slenderness ratios greater than 15. If the pier was found to have less axial load capacity than the demand, it would be classified as Collapse Damage (CD). The performance of each masonry building was claimed to be satisfying the life safety performance level provided that less than 50% of the total base shear at the first story is resisted by masonry piers with a Collapse Damage (CD) performance level.

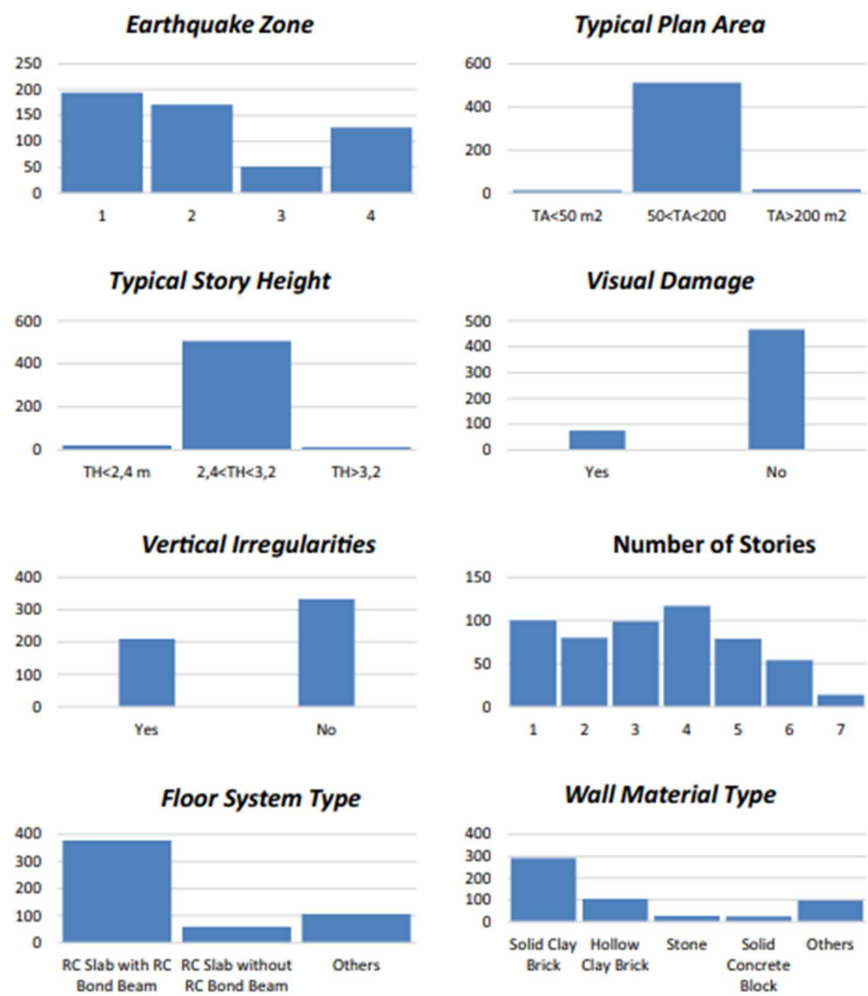


Fig. 1 Frequency of data set

3 Correlation of parameters with detailed seismic assessment analysis results

Detailed seismic assessment analysis of buildings is critical in determining the behavior of the building during seismic events. For this reason, detailed seismic assessment analysis is generally required for the safety check of all buildings over 20 years old and located in close proximity to earthquake zones. On the contrary, rapid screening methods should be used in countries where the filtering of risky buildings is aimed to be performed to take action about retrofitting operations. Therefore, this filtering operation is vital as structures at risk of collapse should be strengthened immediately, or new earthquake-resistant structures should be built instead of these structures. However, current rapid screening methods in the literature have a minimal correlation with the detailed assessment analysis results as none of the methods was calibrated with the detailed analysis results. Thus, it is difficult to depend on the risk estimations of the rapid screening methods while taking actions at the seismic risk mitigation level. Consequently, in this study, it was aimed to form a network to correlate the seismic risk with the properties of buildings. For this reason, it was important to determine the most influential variables to be used in the detection of risky or non-risky structures in correlation with the seismic risk results from the detailed analysis. This operation could yield to dissociate the unnecessary parameters from the network and reduce the bias. The correlation of the variables in the available data set with the class (i.e., the risk state from the detailed analysis, RS) is shown below (Table 2 and Fig. 2). In addition, the relationships between parameters, i.e., visual damage, vertical irregularity, number of stories and typical plan area, and the detailed seismic assessment analysis results are shown in Fig. 3. In Fig. 3, it is apparent that the risk state probability has a robust correlation with the increasing values of story number, vertical irregularity, and visual damage. In addition, total floor area has some limited correlation with the risk state determined from detailed analysis.

In Table 2, the correlation coefficients were calculated using "grouping and mean value determination processes" proposed by McKinney (2011). Grouping and mean value determination processes are used to group the subject data according to some identifiers, to examine the effect of input parameters on the output, to combine these data, or to transform data. This process is called learning through groups (McKinney 2011). In this method, only one independent variable is selected in each case (i.e., vertical irregularities), and the dependent variable is always the state of risk (i.e., 0 or 1). Then, the correlation coefficient for each subcategory in each independent variable is calculated by dividing the number of risky structures with subcategory i to the total number of risky structures. In Table 2, values above 0.5 contribute to the risk of the

Table 2 Correlation of variables with the risk state

FT	RS	TY	RS	TA	RS	
2	0.848	1	0.770	1	0.778	
1	0.807	0	0.667	2	0.562	Visual Damage = VD
3	0.551	2	0.667	0	0.375	Vertical Irregularities = VI
						Earthquake Zone = EZ
VI	RS	VD	RS	NF	RS	
						Wall Material Type = WT
1	0.899	1	1.000	7	1.000	Number of Floors = NF
0	0.678	0	0.727	6	1.000	Floor System Type = FT
				5	0.933	
WT	RS	EZ	RS	3	0.891	
1	0.798	3	0.898	2	0.8859	Typical Story Height = TY
5	0.793	2	0.845	4	0.830	Typical Plan Area = TA
						Result of Seismic Risk = RS
4	0.727	1	0.775	1	0.146	
2	0.717	4	0.584			
3	0.500					

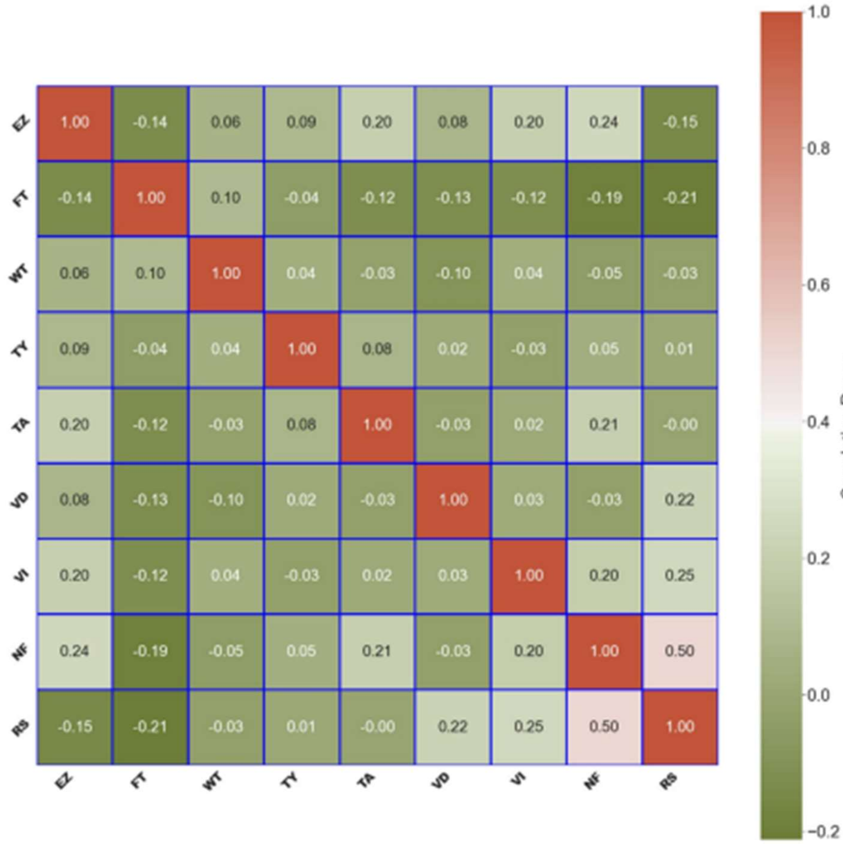


Fig. 2 Correlation relationship of parameters

building, while values below 0.5 contribute to the non-risk of the building. For example, if there is vertical irregularity (VI = 1), the RS value will increase to 0.899, indicating that the building contributes to its risky nature. If there is no vertical irregularity (VI = 0), the RS value will decrease by 0.678, contributing less to the building's risk. Since vertical irregularity in structures is an undesirable situation as it adversely affects load transfer, this obtained result could be claimed to be reasonable. When the heat map is examined (Fig. 2), the parameters with the highest correlation that affect the detailed seismic assessment analysis result are vertical irregularity (VI), visual damage (VD), and the number of floors (NF). The correlation of these parameters is also technically rational. Because the likelihood that the structure becomes risky for seismic disturbances will increase if the structure is damaged. Likewise, as the number of floors increases, the horizontal drift demands of the building will increase inherently, which increases the seismic risk. Besides, the presence of structural cracks and vertical irregularities in the structure (i.e., VD = 1, VI = 1) causes the risk state (RS) to converge to 1. In contrast, the absence of structural damages (VD = 0, VI = 0) causes the RS to converge to 0.

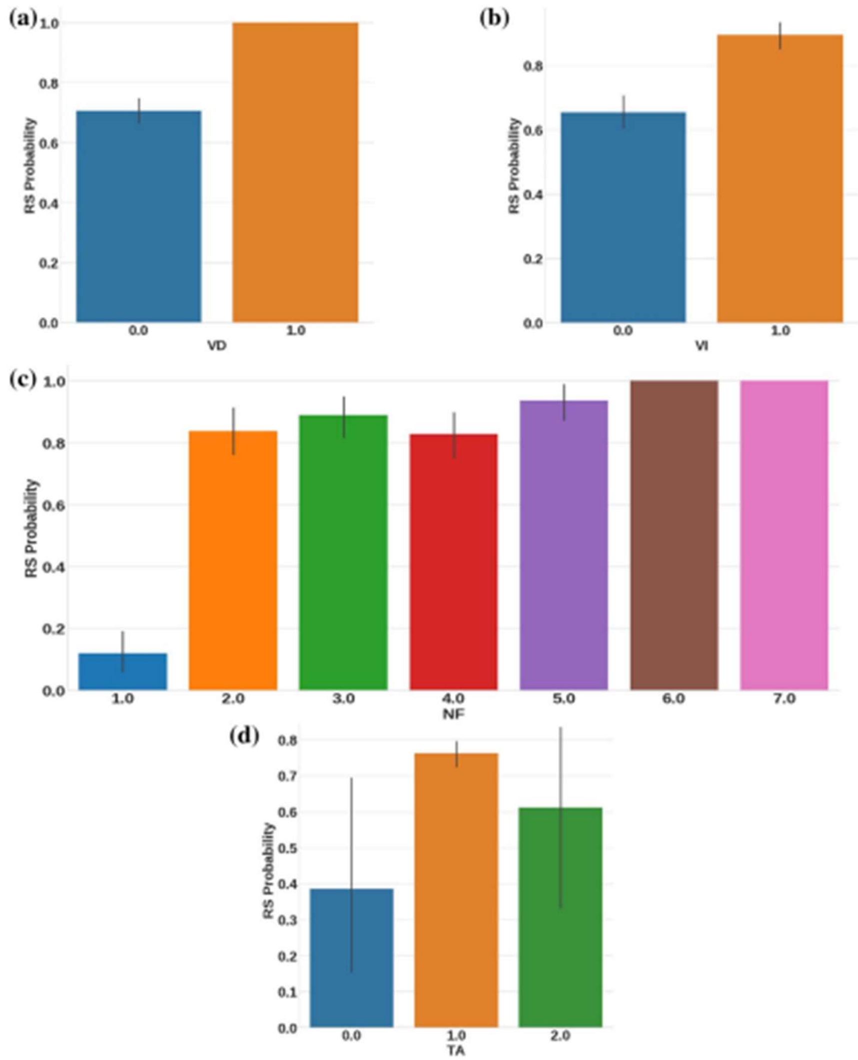


Fig. 3 Relationships between **a** visual damage, **b** vertical irregularity, **c** number of stories, and **d** typical plan area and the detailed seismic assessment analysis results

In order to further investigate the selected parameters from different perspectives, another graph representing the relation between buildings with vertical irregularities and visual damage is plotted (Fig. 4a). From Fig. 4a, it could be inferred that vertical irregularity (VI = 1) is observed more in risky and structurally undamaged structures (i.e., RS = 1 and VD = 0) than in non-risky and structurally undamaged structures (RS = 0 and VD = 0). This leads to the conclusion that vertical irregularity and visual damage correlate well with each other. Figure 4b shows the distributions of the risk states of buildings, the number of stories, and visual damage parameters and their

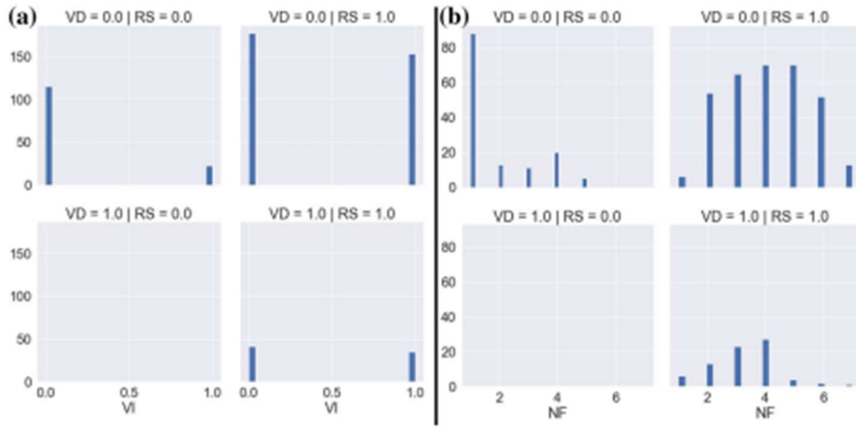


Fig. 4 Relationship between **a** visual damage—vertical irregularities and **b** number of stories—visual damage and the detailed seismic analysis results

interrelationships. The structures with no structural damage and no risk (i.e., VD = 0 and RS = 0) in the data set are usually one story. Therefore, it is understood that the probability of single-story masonry buildings being non-risky is high, and this situation may change toward risky as the number of floors increases. Higher lateral displacement in high-rise buildings may be the reason for this situation.

4 Machine learning algorithms

Machine learning (ML) is concerned with the ability of data-driven models to learn information about a system from directly observed data without predetermining the mechanical relationships. ML algorithms can adaptively improve their performance with each new data sample, update their differentiable weights according to the new data and discover relationships in complex heterogeneous and high-dimensional data (Shaikhina et al. 2019). In this study, instead of sticking to a single method, it was preferred to use multiple machine learning methods (i.e., ensemble learning) in order to achieve the highest success percentage. As part of ensemble learning, different supervised machine learning algorithms were utilized since supervised machine learning algorithms rely on labeled input data to learn a function that produces an appropriate output when given new unlabeled data (Fig. 5a). Logistic regression (Peng et al. 2002); decision tree classifier (Kotsiantis et al. 2007); random forest classifier (Shaikhina et al. 2019); support vector machine (SVM) classifier (Widodo and Yang 2007); and K-neighbors classifier (Imandoust and Bolandraftar 2013) are used to predict building damage levels. To this end, the dataset was divided into training and test datasets first. Then, the statistical measures, like the correlation of parameters, are determined along with the feature engineering operations, i.e., categorical variable definitions. Then, the ensemble learning algorithms were codified to perform the necessary learning operations. Finally, the performance of the ML network on the estimation of the risk state was checked with the test database (Fig. 5b).

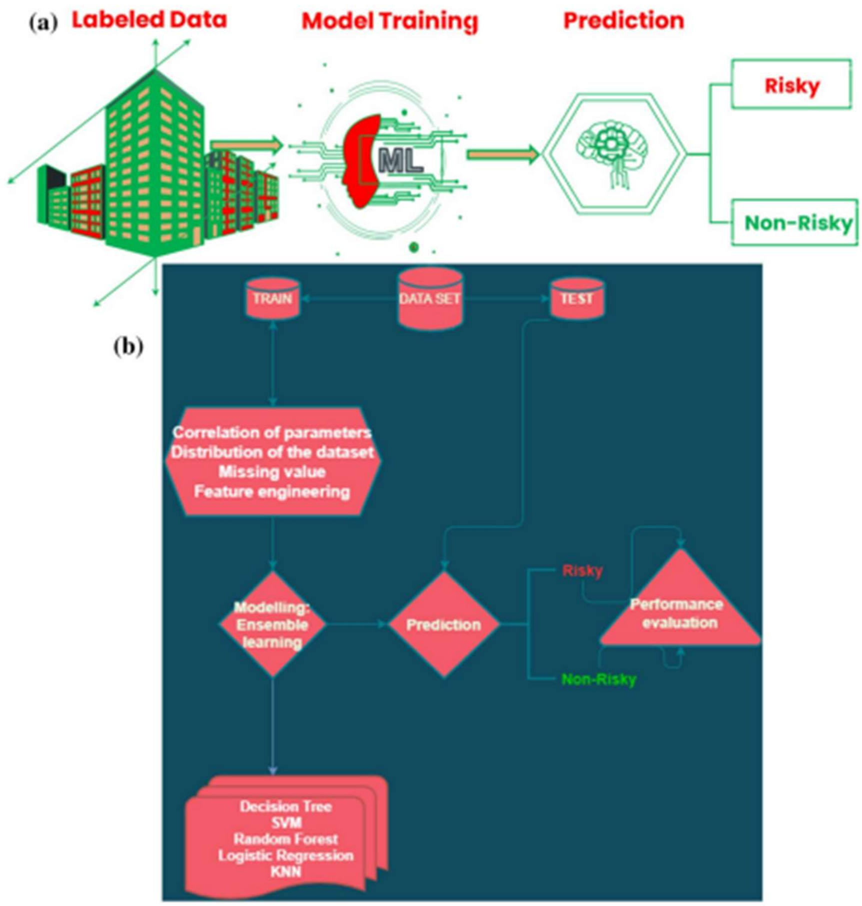


Fig. 5 **a** Supervised learning example and **b** Flowchart of the method used in this study

5 Performance metrics

The seismic risk distribution of a large building stock was aimed to be accurately estimated using the aforementioned machine learning algorithms. In all networks, the risk state from the machine learning algorithm was taken as risky (non-risky) if the risk score became 1 (0). In this part, the performances of the used algorithms will be evaluated not only by the success percentage but also by the metrics like true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These metrics were transformed into precision, recall, and the combined measure (i.e., $F_{measure}$) given in Eqs. (1–3) (Saito and Rehmsmeier 2015).

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$F_{\text{measure}} = \frac{2\ \text{Precision} \times \text{Recall}}{\text{Presicion} + \text{Recall}} \tag{3}$$

The utilized data set consisted of 543 buildings and was analyzed using multiple machine learning algorithms. All the calculations were made in the Jupyter environment (i.e., formally known as IPython). Initially, the dataset was divided into 434 trains and 49 tests with the “Train Test Split” function from the “Sklearn model selection” library. Next, another dataset with additional 60 buildings excluded from the initial stage was formed for the second test stage. In other words, the total training dataset (i.e., the training dataset plus validation dataset) was equal to 483 buildings, whereas the test dataset comprised 60 buildings. In addition, K-fold cross-validation was applied at the validation stage, which prevents the trained data set from being overfitted by the algorithm. As explained before, logistic regression, decision tree classifier, random forest classifier, support vector machine (SVM) classifier, and K-neighbors classifier were all used in this study. The accuracy of each method is presented in Table 3. In addition, the confusion matrices are summarized in Table 4.

It should be noted that it is essential to choose the correct hyperparameters of the logistic regression to increase the percentage of accuracy. There were many hyperparameters for each model, so an excellent way to identify the best set of hyperparameters was to try different combinations and compare the results. The penalty value of l2 is chosen because l2 provides a better prediction when the output variable is a function of all input properties. The coefficient value, on the other hand, was chosen as one of the values providing the highest percentage of success in the graph in Fig. 6a. In the decision tree algorithm, there are many parameters that affect the success percentage. Of these, one of the values “maximum depth”, providing the highest percentage of success, was selected from Fig. 6b. In the random forest algorithm, there are many parameters that affect the success percentage. Of these, “minimum samples split” was chosen as one of the values that provide the high success percentage in the graph in Fig. 6c. It should be noted that this algorithm has many parameters that affect the results. While applying the KNN algorithm in this study, the “number of data points: k” parameter is one of the most important parameters in increasing the success percentage. Therefore, it was clear from Fig. 6d that the worst neighbor values for the dataset were $k<2$, $3<k<4$, and $k>13$. In this case, these values should be avoided in choosing the k value. It has been observed that all values do not change the percentage of success in the penalty parameter selection of the Support Vector Machine Classifier

Table 3 Correlation of variables with the risk state

Method	Training error accuracy (%)	Testing error accuracy (%)
Logistic regression (LR)	93.88	91.88
Decision tree classifier (DTC)	95.00	95.90
Random forest classifier (RFC)	93.33	95.90
Support vector machine classifier (SVMC)	96.67	93.80
K-neighbors classifier (KNN)	93.00	88.50
Ensemble learning (EL)	96.67	95.90

Table 4 Confusion matrices of each algorithm

	LR		DTC		RFC		SVMC		KNN		EL	
	TS = 1	TS = 0	TS = 1	TS = 0	TS = 1	TS = 0	TS = 1	TS = 0	TS = 1	TS = 0	TS = 1	TS = 0
TS = 1	**20 TP**	**3 FN**	**21TP**	**2FN**	**19 TP**	**4 FN**	**19**	**4**	**19**	**4**	**21**	**2**
TS = 0	**0 FP**	**37 TN**	**0 FP**	**37 TN**	**0 FP**	**37 TN**	**0**	**37**	**0**	**37**	**0**	**37**
Precision (%)	**94**		**94**		**94**		**94**		**94**		**97**	
Recall (%)	**93**		**94**		**93**		**93**		**93**		**97**	
Fscore (%)	**93**		**94**		**93**		**93**		**93**		**97**	

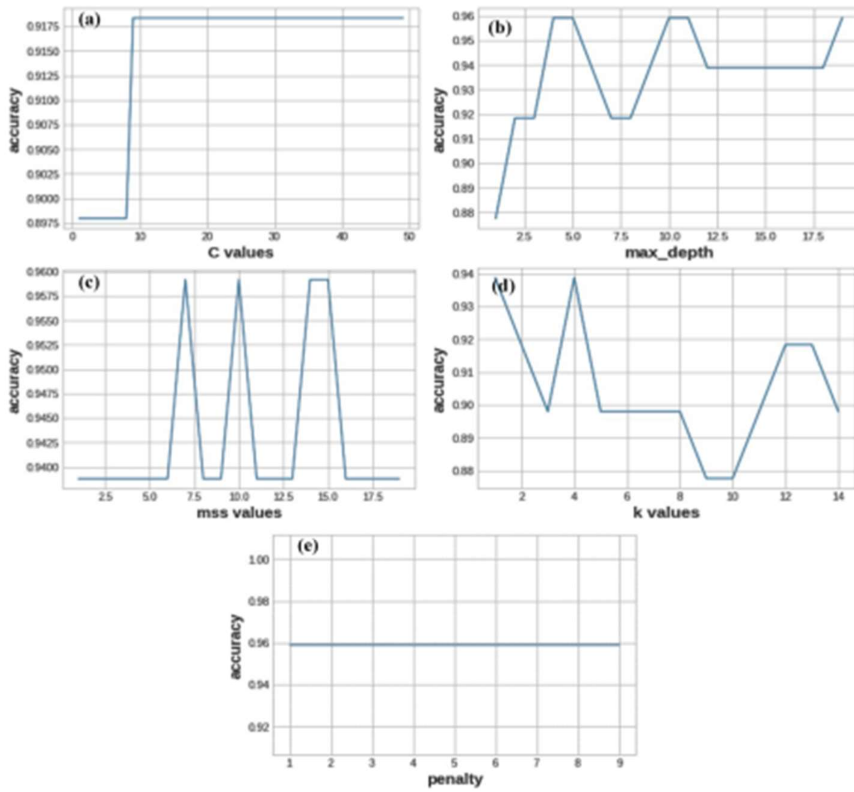


Fig. 6 Effect of algorithm parameters: **a** LR, **b** DTC, **c** RFC, **d** KNN and **e** SVMC

algorithm (Fig. 6e). In this study, the optimum values required to increase the success percentage of each algorithm were determined by the grid search method.

Finally, the number of possible outcomes of the variable RS that lead to the most accurate estimations was investigated. To this end, the Euclidian distances were plotted against the data points (i.e., dendrogram). The least number of possible outcomes could be determined by counting the least number of intersecting points with any possible horizontal lines drawn on the dendrogram. For the risk state variable, this number equaled two (Fig. 7). Therefore, this cross-check also verified the validity of the selected possible outcomes of the RS variable (i.e., risky or non-risky). In other words, the formed machine learning network could only distinguish between the risky and non-risky buildings. It could not classify the buildings according to their possible damage rates like minor damage, moderate damage, or collapse.

In Ensemble Learning, voting is one of the simplest ways of combining the predictions from multiple machine learning algorithms. In this study, a single machine learning algorithm was not used to estimate the seismic analysis result. By using more than one machine learning algorithm, the majority of their predictions are based on votes. For this method, "hard" or "soft" voting can be done. Here, the algorithms give the seismic analysis result with two options: 1–0 (hard) or percentage ratio estimates (soft). In the proposed

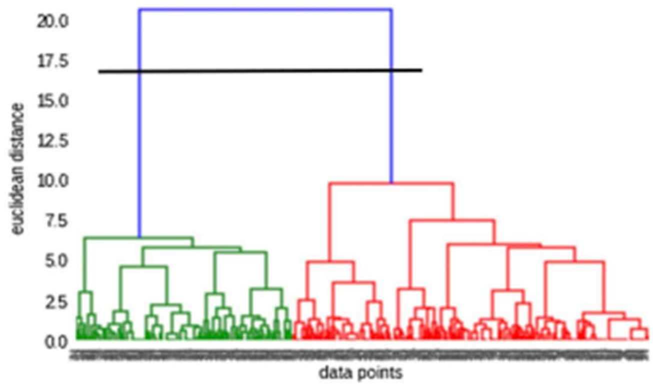


Fig. 7 Dendrogram representing the least number of outcomes for the variable RS

methodology, it was decided to use hard voting for the data set. Components of the ensemble learning, performance results, and schematic representations are shown in Figs. 8 and 9. In addition, the success rate of this machine learning algorithm is shown in Table 5. In Fig. 9, the mean accuracy is obtained by taking the mean accuracy across k-folds for each algorithm.

6 Performance of the proposed method with damage observations after dinar EQ and elazig-kovancilar EQ in Turkey

In this part, the risk status estimations of the buildings whose actual damages after real earthquakes in Turkey were compared. For this purpose, the Dinar earthquake ($M_L = 6.1$) and the Elazig-Kovancilar earthquake ($M_w = 6.1$) were utilized. During the Dinar EQ, it was reported that 2,043 buildings were completely destroyed, and approximately 4,500 buildings were severely damaged (EERI 1995). It was also reported that 2,549 buildings collapsed, and approximately 50 buildings were severely damaged (Akkar et al. 2011). The pseudo-spectral accelerations in the constant acceleration region for the Dinar EQ and the Elazig-Kovancilar EQ were reported as 0.90 g and 0.82 g, respectively. To test the prediction performance of the machine learning algorithms proposed in this study, 13 buildings (5 buildings from the Elazig-Kovancilar EQ and eight buildings from the Dinar EQ) were used. Details on risk estimates are presented in Tables 6,7. From Tables 6,7, it could be concluded that the seismic risk status of 11 out of 13 buildings was correctly estimated by the proposed method. In summary, the machine learning algorithm gave correct results by estimating six undamaged structures as non-risky and five damaged structures as risky. But, it failed to classify two damaged structures as non-risky (i.e., Building 1 in Table 6 and Building 3 in Table 8). With these results, the proposed ML algorithms correctly estimated the damage status of 13 buildings and achieved 84.6% accuracy.

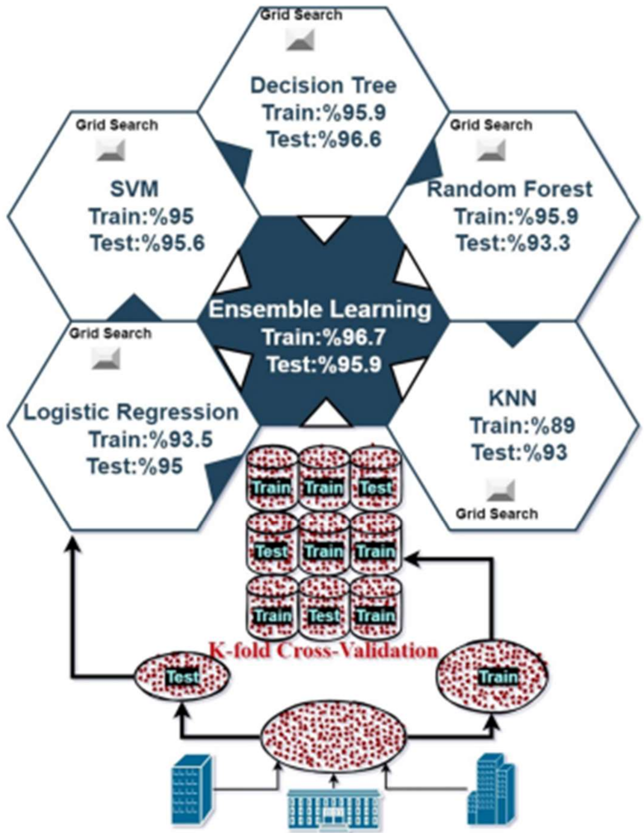


Fig. 8 Comparison of scores algorithms

Fig. 9 Mean accuracy results of each algorithm

7 Discussion of results

Rapid screening methods in the literature classify the earthquake risk of large building stocks by penalizing the existence of some physical deficiencies of structures. The decision is made depending on these penalty scores, which have no intended or calculated correlation to detailed seismic risk analysis. Therefore, in this proposed approach, new insight was

Table 5 Confusion matrix of ensemble learning

	Estimated risk state = 1	Estimated risk state = 0
Real risk state = 1	**21 TP** *	**2 FN**
Real risk state = 0	**0 FP**	**37 TN**
Precision =	**21 / (21 + 0) = 100%**	
Recall =	**21 / (21 + 2) = 91%**	
$F_{measure}$ =	**2 * (1 * 0.91) / (1 + 0.91) = 95%**	

*FN: False-negative, FP: False-positive, TP: True-positive and TN: True-negative

brought to the rapid screening method. To this end, a large database is utilized to correlate the seismic risk analysis results to the rapid screening scores. Therefore, unlike the literature-available rapid screening methods, the proposed method is useful for generating a risk distribution map of large building stocks having a significant correlation with the detailed seismic risk analysis results. Consequently, the proposed machine learning network could be employed while generating seismic risk mitigation systems. The confidence in the proposed machine learning network is enhanced by examining the performance of the proposed network with both post-earthquake reconnaissance and numerical seismic analysis results.

The performance of the proposed network is improved by implementing ensemble learning depending on the hard voting scheme. This manipulation significantly enhanced both the estimation success rate of the training and test database. In addition, the risk state estimation after EQs is aimed to be accurately predicted by the proposed method. To this end, the proposed network's performance was investigated by comparing its risk state estimations with damage states of buildings after real earthquakes. Thirteen buildings damaged during the Dinar EQ ($M_L = 6.1$) and the Elazig-Kovancilar EQ ($M_w = 6.1$) were utilized. The risk estimation performance of the proposed network was found to be as large as 84.6%. This observation also increased the confidence in the proposed network.

In Table 5, it is apparent that the false-negative ratio of the proposed model is 3.33%, whereas false-positive ratio is 0% for the test database. However, the observation for the real EQ application of the proposed method is different. The model estimated two damaged structures as non-risky, which could be a drawback in the practical application of this model. Therefore, the false-positive ratio of the proposed model should be improved in order to have a more dependable model.

8 Conclusion

It is essential to determine the most vulnerable buildings and take precautions before major earthquakes hit in order to eliminate the loss of lives. However, the available detailed procedures require too much human power and resources to be possibly applied in large stocks of buildings. In addition, current rapid visual screening methods used to estimate the risk state of large building stocks do not result in reliable and accurate filtering. Therefore, in this study, it was aimed to bring a new perspective to the building seismic risk filtration. To this end, a calculation network based on the properties of any structures from external

Table 6 Comparison of Risk Estimations of the proposed method with the actual damage observations from Elazig-Kovancilar EQ

Photo of building damage	Risk state from the proposed method		Risk state after EQ
	S_{DS} = 0.82 g NF = 1 story FT = RC Slab without RC Bond Beam VI = No VD = No WT = Stone TY = 3.25 m TA = 86.9 m^2	ML Non-Risky	Heavy Damage after Elazig-Kovancilar EQ
	S_{DS} = 0.82 g NF = 2 stories FT = RC Slab without RC Bond Beam VI = Yes VD = No WT = Mixed TY = 2.80 m TA = 90.95 m^2	ML Risky	Heavy Damage after Elazig-Kovancilar EQ
	S_{DS} = 0.82 g NF = 1 story FT = Timber Slab VI = No VD = No WT = Stone TY = 3.25 m TA = 290.15 m^2	ML Risky	Heavy Damage after Elazig-Kovancilar EQ

Table 6 (continued)

Photo of building damage	Risk state from the proposed method		Risk state after EQ
	S_{DS} = 0.82 g NF = 2 stories FT = Timber Slab VI = Yes VD = No WT = Mixed TY = 2.85 m TA = 340.86 m^2	ML Risky	Heavy Damage after Elazig-Kovancilar EQ
	S_{DS} = 0.82 g NF = 1 story FT = Timber Slab VI = No VD = No WT = Stone TY = 3.30 m TA = 381.16 m^2	ML Risky	Heavy Damage after Elazig-Kovancilar EQ

Table 7 Comparison of risk estimations of the proposed method with the actual damage observations from Dinar EQ: Set 2

Photo of building damage	Risk state from the proposed method (BRS Value)		Risk state after EQ
	S_{DS} = 0.90 g NF = 1 story FT = Timber Slab VI = No VD = No WT = Concrete Block TY = 2.75 m TA = 91.08 m^2	ML Non-risky	Minor Damage after Dinar EQ
	S_{DS} = 0.90 g NF = 2 stories FT = RC Slab with RC Bond Beam VI = No VD = No WT = Mixed TY = 2.85 m TA = 200.81 m^2	ML Non-risky	Minor Damage after Dinar EQ
	S_{DS} = 0.90 g NF = 1 story FT = Timber Slab VI = No VD = No WT = Concrete Block TY = 2.75 m TA = 228.06 m^2	ML Risky	Heavy Damage after Dinar EQ

Table 7 (continued)

Photo of building damage	Risk state from the proposed method (BRS Value)		Risk state after EQ
	S_{DS} = 0.90 g NF = 1 story FT = RC Slab with RC Bond Beam VI = No D = No M = Concrete Block H = 2.60 m A = 90.96	ML Non-risky	Minor Damage after Dinar EQ

Table 8 Comparison of Risk Estimations of the proposed method with the actual damage observations from Dinar EQ: Set 1

Photo of building damage	Risk state from the proposed method		Risk state after EQ
	S_{DS} = 0.90 g NF = 3 stories FT = RC Slab with RC Bond Beam VI = No VD = No WT = Solid Clay Brick TY = 3.05 m TA = 123.88 m^2	ML Non-risky	Minor Damage after Dinar EQ
	S_{DS} = 0.90 g NF = 1 story FT = Timber Slab VI = No VD = No WT = Concrete Block TY = 2.5 m TA = 91,80 m^2	ML Non-risky	Minor Damage after Dinar EQ
	S_{DS} = 0.90 g NF = 1 story FT = RC Slab with RC Bond Beam VI = No VD = No WT = Concrete Block TY = 2.65 m TA = 142.79 m^2	ML Non-risky	Moderate Damage after Dinar EQ

Table 8 (continued)

Photo of building damage	Risk state from the proposed method		Risk state after EQ
	S_{DS} = 0.90 g NF = 1 story FT = Timber Slab VI = No VD = No WT = Hollow Clay Brick TY = 2.80 m TA = 88.16 m^2	ML Non-risky	Minor Damage after Dinar EQ

observations was collected. This network was trained with the known detailed seismic risk analysis results. Then, the seismic risk analysis estimation performance of this algorithm was tested with untrained buildings.

In this study, instead of depending on the estimation of a single method, it was preferred to use ensemble learning in order to achieve the highest success rate. As part of ensemble learning, different supervised machine learning algorithms were utilized. Logistic regression, decision tree classifier, random forest classifier, support vector machine classifier, and K-neighbors classifier are used to predict building damage levels. Then, hard voting was utilized as the outcome of the proposed method was designed to be composed of risky and non-risky buildings (i.e., 1–0). The successful percentage estimations of the proposed ensemble learning for the training and test database were 96.7% and 95.9%, respectively.

The proposed network's performance was also investigated by comparing real EQ damages. In summary, the proposed ML algorithm gave correct results by classifying six undamaged structures as non-risky and five damaged structures as risky. However, the method failed to estimate the risk state of two damaged structures by defining them as non-risky (i.e., Building 1 in Table 6 and Building 3 in Table 8). Therefore, the proposed ML algorithm achieved 84.6% accuracy. This methodology could serve better, provided that mobile applications or web-based software are designed to enable data entry in the field.

Acknowledgements The authors thank the Ministry of Environment and Urbanization for its kind support during the database collection. This publication is a part of doctoral dissertation work by the first author in the Academic Program of Civil Engineering, Institute of Science, Hacettepe University.

Author contributions All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Onur Coskun. The first draft of the manuscript was written by Alper Aldemir and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Achs G, Adam C (2012) Rapid seismic evaluation of historic brick-masonry buildings in vienna (austria) based on visual screening. Bull Earthq Eng 10:1833–1856. https://doi.org/10.1007/s10518-012-9376-5

Achs, G (2011) Seismic hazard of historic residential buildings: evaluation, classification and experimental investigations. Ph.D. thesis (in German), Vienna University of Technology

Ahmad N, Ali Q (2017) Displacement-based seismic assessment of masonry buildings for global and local failure mechanisms. Cogent Eng 4:1414576. https://doi.org/10.1080/23311916.2017.1414576

Akkar S, Aldemir A, Askan A, Bakır S, Canbay E, Demirel IO, Erberik MA, Guulerce Z, Gülkan P, Kalkan E, Prakash S, Sandıkkaya MA, Sevilgen V, Ugurhan B, Yenier E (2011) 8 March 2010 elazıg-kovancılar (turkey) earthquake: observations on ground motions and building damage. Seismol Res Lett 82(1):42–58

Aldemir A, Erberik MA, Demirel IO, Sucuoglu H (2013) Seismic performance assessment of unreinforced masonry buildings with a hybrid modeling approach. Earthq Spectra 29(1):33–57

Aldemir A, Guvenir E, Sahmaran M (2020) Rapid screening method for the determination of regional risk distribution of masonry structures. Struct Saf 85:101959

Al-Nimry H, Resheidat M, Qeran S (2015) Rapid assessment for seismic vulnerability of low and medium rise infilled rc frame buildings. Earthq Eng Eng Vib 14:275–293. https://doi.org/10.1007/s11803-015-0023-4

Askan A, Yucemen MS (2010) Probabilistic methods for the estimation of potential seismic damage: application to reinforced concrete buildings in turkey. Struct Saf 32:262–271. https://doi.org/10.1016/j.strusafe.2010.04.001

Beyer K, Petry S, Tondelli M, Paparo A. Towards displacement-based seismic design of modern unreinforced masonry structures. InPerspectives on European Earthquake Engineering and Seismology 2014 (401-428) Springer Cham

Cooner AJ, Shao Y, Campbell JB (2016) Detection of urban damage using remote sensing and machine learning algorithms: revisiting the 2010 Haiti earthquake. Remote Sensing 8:868

Coskun O, Aldemir A, Sahmaran M (2020) Rapid screening method for the determination of seismic vulnerability assessment of RC building stocks. Bull Earthq Eng 18:1401–1416

D'Ayala D. Assessing the seismic vulnerability of masonry buildings. In:Handbook of seismic risk analysis and management of civil infrastructure systems 2013 (pp 334-365) Woodhead publishing.

Dejong, MJ (2009) Seismic assessment strategies for masonry structures. Ph.D. Thesis, Massachusetts Institute of Technology, Boston, USA.

EERI (1995) 1 Ekim 1995 Dinar earthquake engineering report (in Turkish), METU Press.

European committee for standardization (CEN) (2003) Eurocode 6: Design of masonry structures. prEN 1996–1, Brussels, Belgium

Federal Emergency Management Agency (FEMA P154) (2015) Rapid visual screening of buildings for potential seismic hazards: A handbook. Washington, D.C, USA.

Federal Emergency Management Agency (FEMA356) (2000) Prestandard and commentary for the seismic rehabilitation of buildings. Washington, D.C, USA.

GABHR (2019) Guidelines for the assessment of buildings under high risk. Ministry of Environment and Urbanization, Ankara, Turkey.

GABHR (2012) Guidelines for the assessment of buildings under high risk, Ministry of Environment and Urbanization, Government of Republic of Turkey (in Turkish).

Geiß C, Pelizari PA, Marconcini M, Sengara W, Edwards M, Lakes T, Taubenböck H (2015) Estimation of seismic building structural types using multi-sensor remote sensing and machine learning techniques. ISPRS J Photogramm Remote Sens 104:175–188

Grünthal, G., ed. (1998) European macroseismic scale 1998. Cahiers du Centre Européen du Géodymamique et de Séismologie 15 Luxembourg: Centre Européen de Géodynamique et de Séismologie, 99 pps

Harirchian E, Kumari V, Jadhav K, Das RR, Rasulzade S, Lahmer T (2020) A machine learning framework for assessing seismic hazard safety of reinforced concrete buildings. Appl Sci 10(20):7153

Harirchian E, Lahmer T, Kumari V, Jadhav K (2020) Application of support vector machine modeling for the rapid seismic hazard safety evaluation of existing buildings. Energies 13(13):3340

Imandoust SB, Bolandraftar M (2013) Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. Int J Eng Res Appl 3(5):605–610

Javidan MM, Kim J (2022) Fuzzy-based method for efficient seismic performance evaluation of structures with uncertainty. Comput Aided Civil Infrastruct Eng 37(6):781–802

Javidan MM, Kim J (2022b) An integrated system for simplified seismic performance evaluation and life-cycle cost analysis. J Build Eng 45(1):103655

Kerle N, Nex F, Gerke M, Duarte D, Vetrivel A (2020) UAV-based structural damage mapping: a review. Int J Geo Inform 9(1):14

Kiani J, Camp C, Pezeshk S (2019) On the application of machine learning techniques to derive seismic fragility curves. Comput Struct 218(1):108–122

Kotsiantis SB, Zaharakis I, Pintelas P (2007) Supervised machine learning: a review of classification techniques. Emerg Artif Intell Appl Comput Eng 160:3–24

Kumar SA, Rajaram C, Mishra S, Kumar RP, Karnath A (2017) Rapid visual screening of different housing typologies in Himachal Pradesh, India. Nat Hazards 85:1851–1875. https://doi.org/10.1007/s11069-016-2668-3

Li C, Zhang G, Lei T, Gong A (2011) Quick image-processing method of UAV without control points data in earthquake disaster area. Trans Nonfer Metal Soc China 21:523–528

Li LL, Liu XG, Chen QH, Yang S (2018) Building damage assessment from polsar data using texture parameters of statistical model. Comput Geosci 113:115–126

Mangalathu S, Sun H, Nweke CC, Yi Z, Burton HV (2020) Classifying earthquake damage to buildings using machine learning. Earthq Spectra 36(1):183–208

Mazılıgüney L, Yakut A, Kadaş K, Kalem İ. Evaluation of Preliminary Assessment Procedures for Reinforced Concrete School Buildings in Turkey. InTenth International Congress on Advances in Civil Engineering (ACE2012) 2012 (17-19)

Peng CYJ, Lee KL, Ingersoll GM (2002) An introduction to logistic regression analysis and reporting. J Educ Res 96(1):3–14

Penna A (2015) Seismic assessment of existing and strengthened stone-masonry buildings: critical issues and possible strategies. Bull Earthq Eng 13:1051–1071. https://doi.org/10.1007/s10518-014-9659-0

Penna A, Lagomarsino S, Galasco A (2014) A nonlinear macroelement model for the seismic analysis of masonry buildings. Earthq Eng Struct Dynam 43(2):159–179

Perrone D, Aiello MA, Pecce M, Rossi F (2015) Rapid visual screening for seismic evaluation of rc hospital buildings. Structures 3:57–70. https://doi.org/10.1016/j.istruc.2015.03.002

Rajarathnam S, Santhakumar AR (2015) Assessment of seismic building vulnerability based on rapid visual screening technique aided by aerial photographs on a gis platform. Nat Hazards 78:779–802. https://doi.org/10.1007/s11069-014-1382-2

Ruggieri S, Cardellicchio A, Leggieri V, Uva G (2021) Machine-learning based vulnerability analysis of existing buildings. Autom Constr 132(1):103936

Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10(3):e0118432

Shah MF, Ahmed A, Kegyes-B OK. (2016) A case study using rapid visual screening method to determine the vulnerability of buildings in two districts of Jeddah, Saudi Arabia. In: Proceedings of the 15th international symposium on new technologies for urban safety of mega cities in Asia, Tacloban, Philippines (pp 7-9)

Shaikhina T, Lowe D, Daga S, Briggs D, Higgins R, Khovanova N (2019) Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed Signal Process Control 52:456–462

Sozen MA (2014) Surrealism in facing the earthquake risk. Springer, Seismic Evaluation and Rehabilitation of Structures

Stepinac M, Gašparović M (2020) A review of emerging technologies for an assessment of safety and seismic vulnerability and damage detection of existing masonry structures. Appl Sci 10:5060

Sublime J, Kalinicheva E (2019) Automatic post-disaster damage mapping using deep-learning techniques for change detection: case study of the Tohoku tsunami. Remote Sens 11:1123

TEC (2007) Turkish Earthquake Code Specification for the Buildings to be Constructed in Disaster Areas. Ministry of Public Works and Settlement, Ankara, Turkey

TEC (2018) Turkish Earthquake Code Specification for structures to be built in disaster areas, Ministry of Environment and Urbanization, Ankara, Turkey.

Wang C, Yu Q, Law KH, McKenna F, Yu SX, Taciroglu E, Zsarnoczay A, Elhaddad W, Cetiner B (2021) Machine learning-based regional scale intelligent modeling of building information for natural hazard risk management. Autom Constr 122(1):103474

Wang X, Wittich CE, Hutchinson TC, Bock Y, Goldberg D, Lo E, Kuester F. (2020) Methodology and validation of UAV-based video analysis approach for tracking earthquake-induced building displacements. J Comput Civil Eng 34 (6)

Widodo A, Yang BS (2007) Support vector machine in machine condition monitoring and fault diagnosis. Mech Syst Signal Process 21(6):2560–2574

Xu ZH, Wu LX, Zhang ZX (2018) Use of active learning for earthquake damage mapping from uav photogrammetric point clouds. Int J Remote Sens 39:5568–5595

Xu, JZ, Lu, W, Li, Z, Khaitan, P, Zaytseva, V (2019) Building damage detection in satellite imagery using convolutional neural networks. 33rd Conference on Neural Information Processing Systems, Vancouver, Canada

Yakut A (2004) Preliminary seismic performance assessment procedure for existing rc buildings. Eng Struct 26:1447–1461. https://doi.org/10.1016/j.engstruct.2004.05.011

Yu Q, Wang C, McKenna F, Yu SX, Taciroglu E, Cetiner B, Law KH (2020) Rapid visual screening of soft-story buildings from street view images using deep learning classification. Earthq Eng Eng Vib 19(4):827–838

Yucemen MS, Ozcebe G, Pay AC (2004) Prediction of potential damage due to severe earthquakes. Struct Saf 26:349–366. https://doi.org/10.1016/j.strusafe.2003.09.002

Zhang Y, Burton HV, Sun H, Shokrabadi M (2018) A machine learning framework for assessing post-earthquake structural safety. Struct Saf 72(1):1–16

Zhang Z, Hsu TY, Wei HH, Chen JH (2019) Development of a data-mining technique for regional-scale evaluation of building seismic vulnerability. Appl Sci 9(7):1502

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 1. Dendogram, Elbow rule and K-means clustering methods (page:79)

```
Dendrogram
from scipy.cluster.hierarchy import linkage, dendrogram
merg = linkage(xC_train,method="ward")
dendrogram(merg,leaf_rotation = 90)
plt.xlabel("data points")
plt.ylabel("euclidean distance")
plt.show()
-----------------------------------------------------------------------
--
Kmeans and Elbow rule
from sklearn.cluster import KMeans
wcss = []

for k in range(1,10):
kmeans = KMeans(n_clusters=k)
kmeans.fit(xC_train)
wcss.append(kmeans.inertia_)

plt.plot(range(1,10),wcss)
plt.xlabel("number of k (cluster) value")
plt.ylabel("wcss")
plt.show()
#kmeans uygulama   6 cluster but we apply 2
kmeans2 = KMeans(n_clusters=3)
clusters3 = kmeans2.fit_predict(xC_train)
xkC_train=xC_train

#xkC_train["label"]= clusters3
data=xkC_train
#but 7 cluster
#10 centers plot
plt.scatter(kmeans2.cluster_centers_[:,0],kmeans2.cluster_centers_[:,
1],color = "yellow")
```

```
plt.show()
```

Chapter 2. Synthetic Minority Oversampling Technique (page:80)

```
print("Before Undersampling, counts of label '1':
        {}".format(sum(yC_train == 1)))
print("Before Undersampling, counts of label '0': {}
        \n".format(sum(yC_train == 0)))
print("Before Undersampling, counts of label '2':
        {}".format(sum(yC_train == 2)))
print("Before Undersampling, counts of label '3': {}
        \n".format(sum(yC_train == 3)))
# apply near miss
from imblearn.under_sampling import NearMiss
from imblearn.over_sampling import SMOTE
sm = SMOTE(random_state = 99)

xC_train_miss, yC_train_miss = sm.fit_resample(xC_train,
        yC_train.ravel())


print('After Undersampling, the shape of train_X:
        {}'.format(xC_train_miss.shape))
print('After Undersampling, the shape of train_y: {}
        \n'.format(yC_train_miss.shape))

print("After Undersampling, counts of label '0':
        {}".format(sum(yC_train_miss == 1)))
print("After Undersampling, counts of label '1':
        {}".format(sum(yC_train_miss == 0)))
print("After Undersampling, counts of label '2':
        {}".format(sum(yC_train_miss == 2)))
print("After Undersampling, counts of label '2':
        {}".format(sum(yC_train_miss == 3)))
```