ORIGINAL RESEARCH

Prediction of Earthquake Vulnerability for Low-Rise RC Buildings Using Probabilistic Random Forest

Khodaparasti M.[1], Alijamaat A.[2], Pouraminian M.[3,*]

Abstract:

Assessing the seismic vulnerability of existing buildings is one of the major concerns of governments in the world. Reducing the destructive and catastrophic consequences of earthquakes is necessary and inevitable. So far, various techniques have been presented to evaluate the seismic vulnerability of buildings. One of the fast and effective assessment techniques is the Rapid Visual Screening (RVS) technique with fastly identify high-risk buildings for a more accurate assessment. Among the RVS methods, the Hassan-Sozen PI method is the simplest method to evaluate the seismic vulnerability of low-rise RC buildings. The value of the priority index (PI) is determined from the simple geometric features of the building such as the number of stories, floors area, column area, area of concrete walls and infilled in the main directions of the building. In this article, the data collection have been gathered from Elyasi et al.'s reference such as geometrical information (with geometrical features provided by Hassan-Sozen) and earthquake features (peak ground acceleration and earthquake magnitude) for 658 low-rise RC buildings. The number of considered input features includes seven geometric features and two earthquake features (9 features in total) and the predicted output of Hassan-Sozen priority index. The machine learning technique utilized in this article for prediction seismic vulnerability is a probabilities random forest in which a simple Bayesian method is used to create forest trees. This method has had a slight improvement in accuracy criteria and considerable improvement in accuracy and recall criteria compared to other traditional random forest and machine learning methods. Improved machine learning algorithms implemented in the RVS prediction techniques can help to achieve sustainable cities and promotion of more resilience cities.

Keywords:
seismic vulnerability, rapid visual screening, machine learning (ML), random forest (RF), simple Bayesian, resilience cities, reinforced concrete buildings.

*Corresponding author Email: majid.pouraminian@iau.ac.ir; https://orcid.org/0000-0001-5648-8365
1 Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran
2 Department of Computer Engineering, Abhar Branch, Islamic Azad University, Abhar, Iran
3 Department of Civil Engineering, Ramsar Branch, Islamic Azad University, Ramsar, Iran
(majid.pouraminian@iau.ac.ir)

1. Introduction

Heavy migration to urban areas and the desire of workers and skilled people for more income and an urban lifestyle has increased demand for housing. There are many old buildings in operation in different parts of countries, which has caused the formation of a building stock with a high degree of vulnerability, especially in countries prone to earthquakes and high seismicity. Screening the buildings of a region, prioritizing, retrofitting to reduction the consequences of earthquakes, financial losses and loss of life is inevitable. Today, screening of buildings before an earthquake is used in many earthquake-prone countries, and it is being developed and improved. Complex structural analysis methodologies have been designed to define screening problems at the level of a building device, which is often time-consuming, costly, and impractical to perform such nonlinear analysis by considering details for individual buildings in an urban area. Therefore, there is a need for a rapid screening method to prioritize and filter stock buildings with a higher degree of vulnerability, which is called Rapid Visual Screening (RVS). The primary motivation hidden in RVS methods is that their design based on simple visual observations such as Sidewalk Surveys fastly detects buildings at risk. With the help of RVS techniques, the number of buildings from a building complex located in an area that requires a more detailed and comprehensive investigation is being reduced.

Feeling the need for such techniques, a handbook titled rapid visual screening of buildings for Potential Seismic Hazards was published in 1988 by the Federal Emergency Management Agency [FEMA] in the United States of America, which was revised in 2002 and the last edition in 2015 under the title of FEMA P-154 was released. The FEMA guideline provides a detailed framework for applying the RVS technique for various building typologies. Using the information collected from the sidewalks, a vulnerability score was calculated to estimate the seismic performance of the building. The scoring system also includes two evaluation levels. The first level is manual (mandatory) which consists of the general information of the building and the main features of the building, and the second level is optional and deals with the information of the first level in more detail. In the first level assessment, the basic of structural score is based on the Seismic Hazard Intensity of the structural system. A lower evaluation score obtained for a building indicates that the building is more vulnerable. The calculated score S will indicate the probability of collapse equal to 1/10S. The methodology provided by FEMA has been applied in several countries after localization (Haririchian et al., 2021). Some of the RVS methods include the method provided in the regulations of New Zealand and the JBDPA guidelines of Japan. (Shaheryar et al., 2022). In addition to existing RVS methodology based on rules and policies, other methods have also been presented based on researchers' research.

In 1997, Hassan and Sozen invented a simple method for quick evaluation of low-rise RC buildings, which is known as the Priority Index (PI). The PI index is empirically calibrated based on data collected from buildings with damage at different levels after the Erzincan earthquake in 1992. For the process of evaluating and determining the PI index, it is only necessary to consider information about the geometric characteristics of the building, including the number of stories, the area of the floors, the cross-sectional area of the columns above the ground floor, and the area of the walls. The characteristics of the construction soil type and the geometric irregularity of the building are not considered in calculating the PI index, and the presented index is obtained based on the specific region.

Niloofar Elyasi et al. In 2023, by adding two features of seismic intensity, calculation the first one is the peak acceleration of the ground PGA: Peak Ground Acceleration and, the other is the magnitude of the earthquake in the modified Mercalli intensity scale to the seven geometric features used in the PI index

(with a total of 9 features) of the seismic vulnerability of the building. It was presented low-rise RC. The database use by them includes Survey impressions of 9 characteristics in 658 buildings with different levels of damage that occurred in 6 earthquakes such as Duzce 1999, Bingol 2003, Nepal 2015, Taiwan 2016, Ecuador 2016, and Pohang 2017 in different countries. The RVS problem is binary classification and the random forest algorithm is directly used to predict the damage state based on the PI input parameter. The results show that the presented methodology has given a significant improvement in the PI index and there is no need to judge in definition of damage border categories. Also, the non-region-specific model has a good performance for low-rise RC buildings in areas with seismic risk potential.

In this article, considering the Hassan-Sozen PI index and based on the research database of Elyasi et al. Including 658 buildings with geometric and seismic characteristics, a new machine learning model is presented for vulnerability assessment. The innovation of this article shows how to use the new algorithm in a random forest of probabilities by using the simple Bayesian method to create forest trees for modeling seismic damage prediction of buildings.

2. Dataset

The dataset includes inputs and outputs. Inputs include seven geometric features according to the Hassan-Sozen PI index. Also, the entrance has two seismic features.

2.1. Geometrical Features

To determine the seismic vulnerability rating of low-rise RC buildings in the Hassan-Sozen PI index, seven features include the number of stories, the floor area, the column area, the area of the infill wall in the main direction of the structure and the area of the reinforced concrete walls in the main direction of the structure. According to equation 1 to 4, the PI index is calculated from the sum of the wall index and the column index, and the lowest value calculated in the north-south and east-west direction will be the calculation criterion.

$$PI = \min\left\{PI^{N-S}; PI^{E-W}\right\} \tag{1}$$

$$PI^{N-S} = WI^{N-S} + CI; PI^{E-W} = WI^{E-W} + CI; \tag{2}$$

$$WI^{N-S} = \frac{\left[\left(A_{cw}^{N-S} + A_{mw}^{N-S}\right)/10\right]}{A_{ft}} \times 100; \tag{3a}$$

$$WI^{E-W} = \frac{\left[\left(A_{cw}^{E-W} + A_{mw}^{E-W}\right)/10\right]}{A_{ft}}; \tag{3b}$$

$$CI = \frac{\left[\left(A_{col}\right)/2\right]}{A_{ft}} \times 100; \tag{4}$$

Which A_{mw} and A_{cw} are respectively the total cross-sectional area of reinforced concrete walls and non-reinforced masonry filling walls in the horizontal direction at the ground floor, respectively.

A_{ft} is also the total area of the floor above the ground floor and A_{col} is the total area of the cross-section of the columns above the ground floor. Buildings with a lower PI score are more vulnerable to earthquakes and it should be prioritized for more detailed studies and evaluation and retrofitted if needed. In the other words, with $PI \to 0$ building suffered severe vulnerability during the Erzincan earthquake.

2.2. Seismic Features

To According to the research of Elyasi et al. Two characteristics of the peak acceleration of the earth $[0.15g \le PGA \le 0.89g]$ and the modified Mercalli intensity $[5.05 \le MMI \le 8.78]$ have been considered. Also, the database of features is extracted from the earthquake information in Table 1 and it includes a total number of 658 buildings under six earthquake events in 5 different countries of the world.

Table 1. Seismic features of machine learning model

	Event Name	Data	M_w	*Sample Size*
1	Duzce ,Turkey	Nov 12,1999	7.2	116
2	Bingol , Turkey	May 1,2003	6.4	55
3	Gorkha , Nepal	Apr 15,2015	7.8	135
4	Meinong,Taiwan	Feb 6,2016	6.4	106
5	Muisne , Ecuador	Apr 16,2016	7.8	172
6	Pohang,South Korea	Nov 15,2017	5.4	74
				Sum=658

3. Machine Learning Technique

ML is one of the subsets of artificial intelligence that allows computers to automatically obtain enough information and learn and improve themselves based on the data without directly writing a program by it. One of the main goals of machine learning is the access to data for better learning, the more data, the better learning can be created and ultimately produce better results. There are many algorithms for machine learning and each Hundreds of new algorithms are added daily. These algorithms have various methods that can be referred to as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

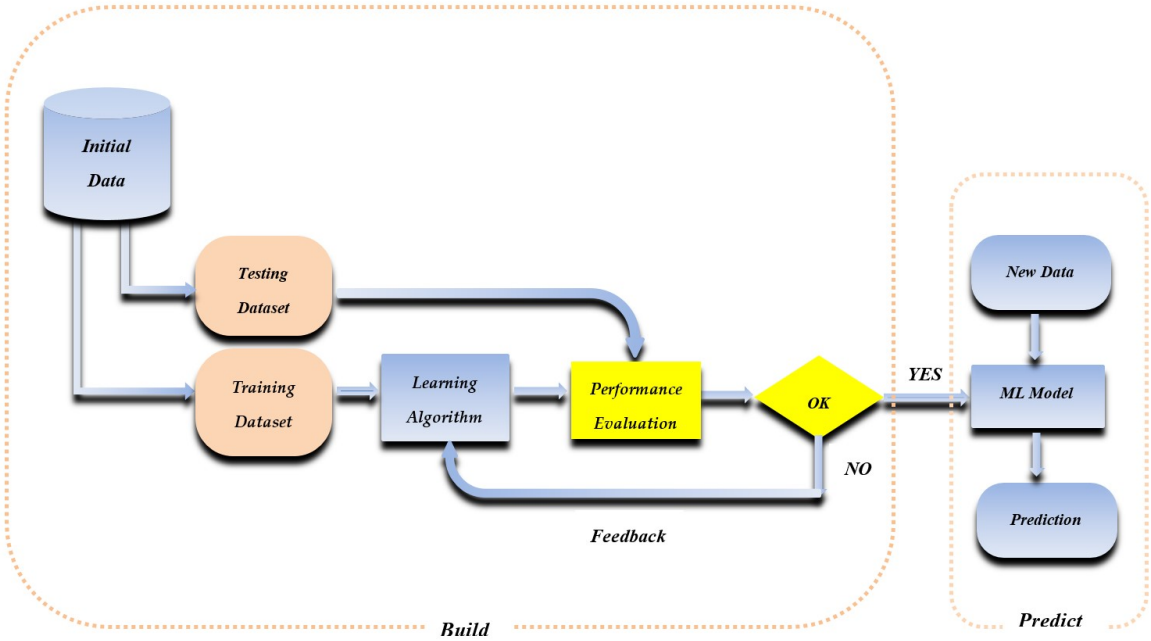


Figure 1. Typical Workflow of ML

In supervised learning algorithms, by using previous knowledge as well as newly labeled data, better predictions can be made for the future. But in unsupervised learning algorithms, classification and labeling have not been done on the information required for training. In this method, the computer system must infer the hidden structure from unlabeled data. Semi-supervised machine learning algorithms are between the two previous types, in other words, it uses both labeled and unlabeled types for training. Such systems can significantly improve learning accuracy.

In the reinforcement learning algorithm, a computer program related to the dynamic environment must achieve a specific goal, in this method the machine learns to make its decisions in a space that is always in trial and error conditions. Figure 1 shows a typical machine learning workflow used in machine modeling. By using machine learning algorithms, computer systems are trained in such a way that they can continue to learn and progress until their goals are met. In machine learning algorithms, there are three main stages: collecting the required data, learning and finally evaluating the efficiency of the desired model.

In the first step, input and output data (labeled data) are usually presented. Of course, in some machine learning algorithms, for better quality, all input data are placed in the range of [0,1] [6]. To test any model, the initial data is randomly divided into two categories, training (usually 70%) and testing (usually 30%) are divided. As it is shown, more percentage of data is used for educational purposes. In the second step, one of the selected algorithms for better learning is taught. The selection of the desired algorithm should be based on the specific applications of that algorithm in solving various problems. Therefore, different algorithms should be examined and the algorithm that is suitable for a specific issue It is more suited to be chosen. In the third step, when a machine learning model has found the required training, its performance is processed and evaluated for the training data, and the correctness of its output according to various functions such as mean absolute error (MAE) and mean square error (MSE), etc. are evaluated.

3.1. Comparative algorithms

In machine learning, there is a set of classification algorithms such as Logistic regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest (RF). To create machine learning models, the choice the algorithms for getting good results is very important. The selection of model depends on many variables in the problem, such as the amount of data, the dimensions of the data, and the distribution hypothesis. A model with low bias and variance (although these two are in the opposite direction in most cases) are two essential and desirable features. In collective machine learning methods, basic models are combined as building blocks to create more complex models. Most of the time these models alone do not perform as well because of having high bias or variance [7, 8]. Ensemble algorithms usually have more stability than single models and are more successful in overfitting management. Of course, these models require more calculations due to their structure. In general, experimentally, ensemble models have better results than individual models [9]. In this article, the probabilistic random forest algorithm method is used to predict the seismic vulnerability model.

3.1.1. Logistic Regression

Logistic Regression is a supervised machine learning algorithm that is used for classification problems [10] and one of its main goals is to predict the probability that a sample belongs to a main class. This algorithm receives a linear function as input and maps it to an interval between 0 and 1 using a function known as sigmoid function, that shows in Equation 5. The values obtained from this function show the probability of belonging to a main class, and the higher the probability, the higher the probability that the

sample is in that class. Like linear regression, logistic regression estimates the relationship between a dependent variable and one or more independent variables. The unit of measurement of the prediction result in linear regression is usually continuous (such as numerical values) and requires less sample for prediction, but in logistic regression The unit of measurement of the prediction result is probabilistic and it is between 0 and 1, and it needs larger samples to have sufficient statistical power to detect the appropriate effect. In this algorithm, by using L1 and L2 regularization techniques [11], it is possible to avoid over-matching in different scenarios.

$$g(z) = \frac{1}{1+\exp(-z)} \qquad (5)$$

3.1.2. Decision Tree (DT)

DT is a machine learning algorithm uses in classification and prediction problems. In this method, a decision tree is built for all the input data as shown in Figure 2, which includes the root node, two or more branches, decision nodes and leaf nodes. The root node is the first decision node. And according to the characteristics of each data, the decision nodes direct the navigation path to a specific branch, the leaf nodes contain the final label that shows a special prediction. The decision tree has the power to solve problems with discrete and continuous variables, and for this purpose it is known as classification and regression tree (CART) and can to work with huge data. One of the advantages of this method is the ability to make better decisions for it is when the increase of input data occurs. By dividing the root node into smaller subsets, the decision tree can achieve the objectives of the problem. For this division, it can use various criteria such as entropy, Gini index, regression, etc. This partitioning process continues until it cannot find another partition that matches the criterion of the problem or reaches the maximum depth of the tree [12].

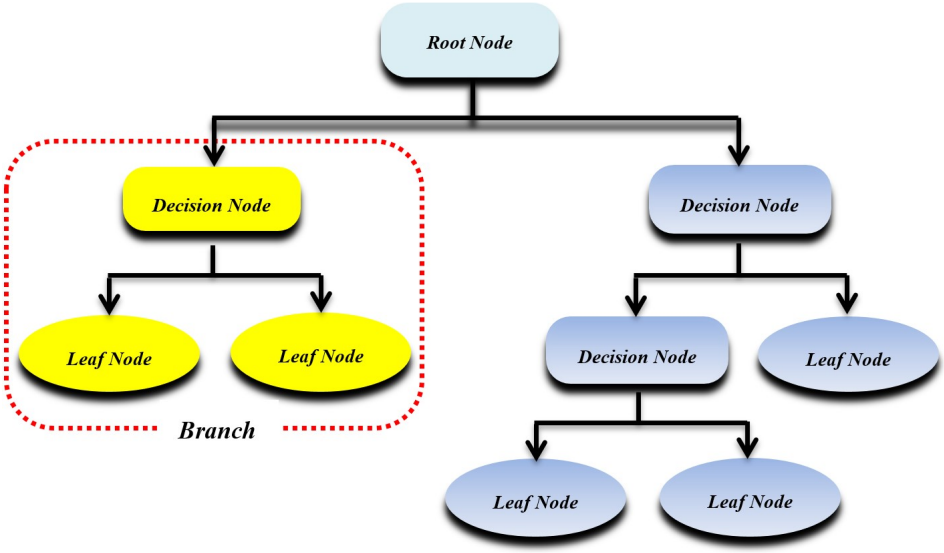


Figure 2. An example of a decision tree structure

3.1.3. K-Nearest Neighbor (KNN)

KNN classification is one of the oldest method and at the same time most effective supervised learning algorithms for data set classification. The KNN algorithm is designed based on the assumption that similar objects exist near each other. Compared to other classification algorithms, KNN uses a

lazy learning approach. In other words, it simply stores the samples in the training phase and does nothing until the test samples are received. Figure 3 presents a diagram of the K-Nearest Neighbor classification. As can be seen, the parameter k plays such an important role in the classification of new samples. Because different k values may lead to very different classification results, in addition, different distance calculations may also lead to different neighborhood, and thus to different classification results. Hence, the given value of K determines the accuracy of the predictions and the number of errors, therefore choosing the suitable K is of fundamental importance in this algorithm. The choice of ideal K depends on the data, but large values of K reduce the impact of noise on classification, by making boundaries and groupings less distinct.

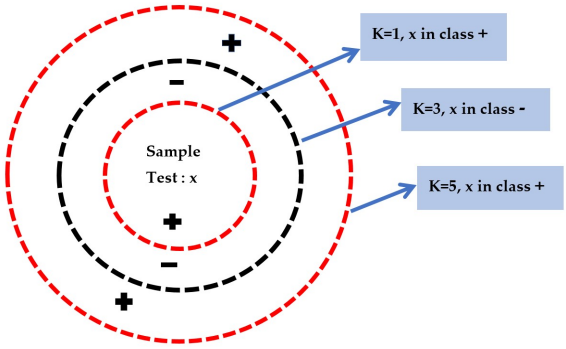


Figure 3. Diagram of the K-Nearest Neighbor classification

The nearest neighbor classification should be based on learning through analogy. The training samples are described by p features. Each sample represents a point in the p-dimensional space. This way, all training samples are stored in a p-dimensional pattern space. Hence, when given an unknown sample, the KNN classifier searches the pattern space for K training samples are closest to the unknown sample. These K training samples are the KNN of the unknown sample. To determine the KNN of the data point, we must use the criteria of analogy or disanalogy between the data points. There are many criterion of analogy or disanalogy, including Euclidean distance, Minkowski distance, Hamming distance, Pearson correlation coefficient, and cosine similarity. KNN algorithm is used for both classification and regression. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Usually, for classification problems, voting and regression problems, averaging can be used to predict the test sample.

The following algorithm shows how KNN works:

Step 1. Select the number K of the neighbor (the number k is determined by the user)
Step 2. Calculate the Euclidean distance (or other k distance criterion) of the number of neighbors.
Step 3. Sort the distance and determine the K nearest neighbors based on the minimum Euclidean distance calculated.
Step 4. Among these K neighbors, count the number of data points in each class.
Step 5. Assign the new data points to the class with the maximum number of neighbors.

3.1.4. Random Forest

Another machine learning algorithm is random forest, which includes a group of easy-to-use decision trees and often provides good results, even without adjusting its meta-parameters. Due to its simplicity, this method can be used for classification and regression problems. As a supervised machine learning model, random forest learns to map data to outputs in the training phase. During training, the model is given data that is relevant to the problem domain and the correct value that the model must learn, to make a correct prediction. The model learns the relationships between the data and the values we want to predict. As shown in Figure 4, the final output of the random forest is obtained based on the majority vote or the average of the results of the decision trees. Therefore, the overfitting problem that Eliminates existing in decision trees. In addition to inheriting the advantages

of the decision tree, random forest can work easily with extensive data with thousands of input variables [13]. Random forest and its improved methods have been used in various civil engineering fields due to the good results [14-17]. Advantages and disadvantages of random forest:

• It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
• Random forests perform better than discrete decision trees for many of data elements.
• Random forest has less variance than discrete decision trees.
• Random forests are flexible and have very high accuracy.
• In random forest algorithm there is no need for data scaling, even without data scaling, good accuracy will remain. Even if a large part of the data is missing, random forest algorithms will be highly accurate.

Complexity is the main disadvantage of random forest algorithms. Constructing random forests is much more hard and time-consuming than decision trees. To implement the random forest algorithm, more computing resources are needed. A random forest will be less intuitive if there is a large set of decision trees. The prediction process using random forests is much time superior compared to other algorithms.

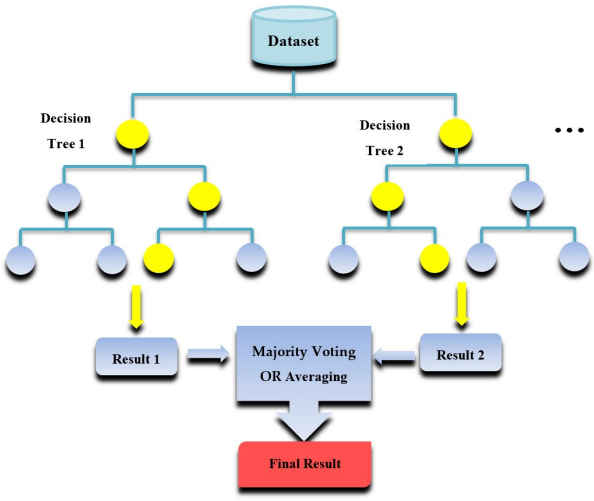
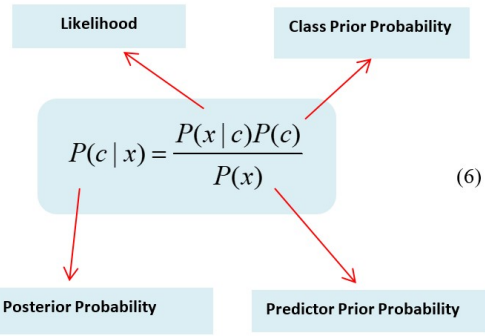


Figure 4. General structure of random forest method

3.1.5. Naive Bayes

Naïve bayes is a simple and very useful classification algorithm in solving problems related to machine learning. This algorithm is classified as a supervised learning algorithm [18]. .Naïve bayes is a probability-based algorithm and uses bayes theory for classification. This algorithm assumes that all the variables of the data set are naive, which it means, there is no relationship between them, and it evaluates the probability of an event occurring based on past knowledge. [19]. In this algorithm, we consider the independent input data, and this simplifies the possible calculations. When new data is presented to this algorithm, it examines the probability of each class and selects the class that has a higher probability as the predicted class for the selected data [11]. However, its performance may affect due to its strong assumptions on features independence. Gaussian, Multinomial, Complement, Bernoulli, and Categorical are the common variants of NB classifier, this algorithm is very widely used and works well when the amount of training data is small compared to the number of features [20].

The Bayes theorem describes the probability of the occurrence of an event based on the knowledge obtained in the past relative to the conditions of occurrence of that event. Simple Bayesian classification assumes that the effect of the value of a predictor (x) on a data class (c) is independent of the values of other predictors. In Bayesian classification, the main goal is to find posterior probabilities, such as probability of a data class as for some observed characteristics, which is shown in equations 6 and 7.



(6)

$$P(c|x) = P(x_1|c) \times P(x_2|c) \times,,,\times P(x_n|c) \times P(c) \quad (7)$$

In the above equation, c represents the desired class and x represents the features that each must be calculated separately. **P(c|x)** is the posterior probability of class c with predictor (feature x), **P(c)** The probability of the desired class and **P(x|c)** is the likelihood criterion that represents the probability of predictor x having class c and **p(x)** is the prior probability of predictor x.

3.1.6. Simple Bayes Tree

Creating a Bayes Tree is the same as the other methods. At each stage, the features that have the greatest impact on separating of categories are selected. The selected feature continues until the usefulness of the new node in separation from the parent node is greater. If this condition is not met and the degree of impurity of the parent node is greater, the new node becomes a leaf and the samples that belong to this leaf reached are categorized by Bayes learning algorithm and the target sample category is recognized. Unlike other trees, the simple Bayes tree does not assign a specific category to the samples that reach leaves. Thus, it uses Bayes algorithm to categorize these samples [20].

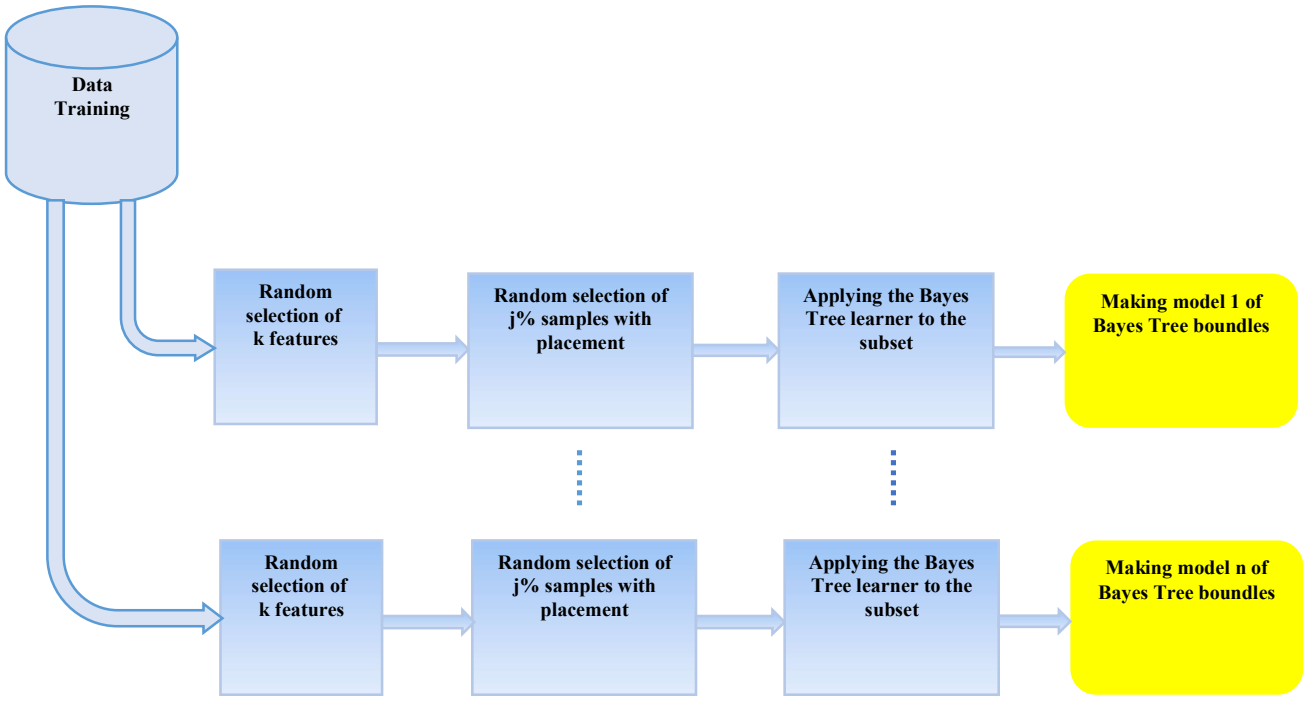
3.1.7. Random Forest Bayes

Random forest is an easy-to-use machine learning algorithm, which usually provides excellent results even without adjusting its meta-parameters. Due to its simplicity, this algorithm can be used for classification and regression. The random forest algorithm is an ensemble classifier that includes several decision trees in different subsets of the data set. Instead of a decision tree, it makes predictions from each tree based on majority votes or averaging. The more number of trees in the forest leads to higher accuracy and avoids the problem of overfitting. Bayes random forest is one of ensemble algorithms and it is a classification algorithm that consists of a combination of decision tree and simple Bayes. To create a model based on the Bayes decision tree, the following steps are performed on the training data. In the first step, a Bayes random forest is created with the following steps:

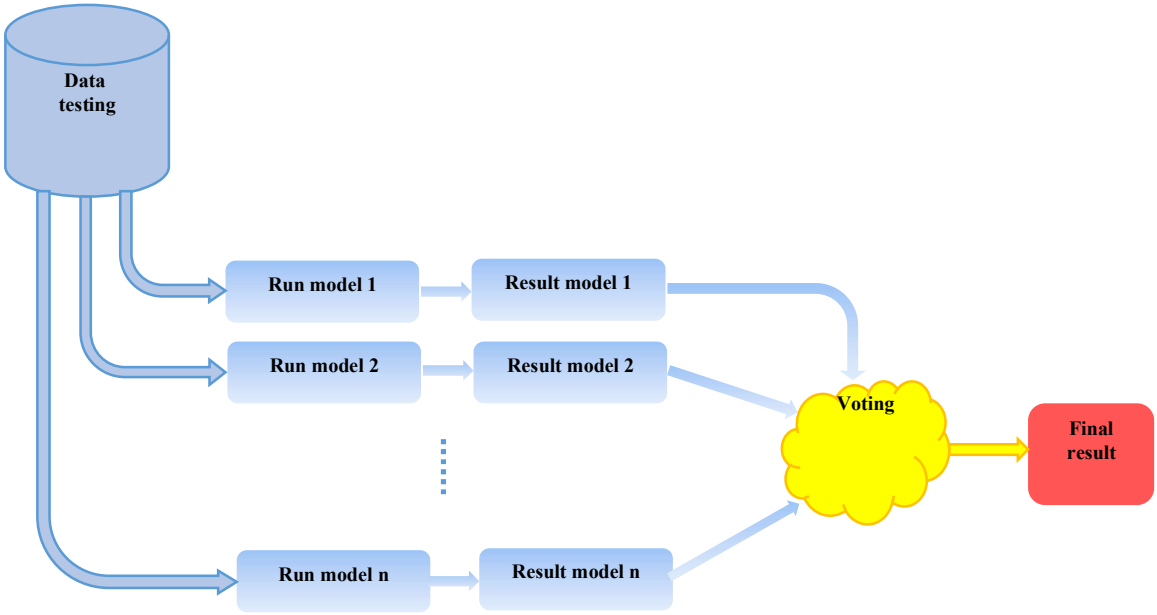
1. The number of k features is randomly selected.
2. From the obtained subset, j percent of the samples are selected using the placement sampling method.
3. Bayes tree learning algorithm is applied on the obtained subset in two steps.
4. A classification model is created and saved according to the selected features and samples.
5. Steps 1 to 4; It is repeated n times and n classification models are formed.

In the second step, the built models are applied to the test samples according to step 6:

6. All test samples are given to the classifier model, *1* to *n* that we have created, and this model specifies m categories for each sample, which determines the final category of each sample by considering the majority vote. The process of the proposed method, which has two stages of training and testing, is shown in Figure 5.



(A)



(B)

Figure 5. A. training step and B. testing step of the proposed method

3.2. Evaluation Criteria

Evaluation criteria are selected based on the type of problem in machine learning. Different criteria are used in regression and classification problems. In this research, which is a multi-class classification problem, the accuracy rate is an important measure to evaluate the performance of the model, but it is not enough. Therefore, other criteria have been introduced to analyze model performance and classification issues, with the help of which we can have a more comprehensive and general understanding of model performance. These measures include sensitivity and F1 values, which are described below. To calculate the values of the criteria, after applying the training data on the created model, the results are displayed in the form of a confusion matrix, in which the results matrix has the following four modes:

True - Positive (TP), is when both the actual and predicted classes of data points are 1.

True-Negative (TN), is when the actual and predicted classes of data points are zero.

False- positive (FP), is when the actual class of the data point is zero and the predicted class is 1.

False-Negative (FN), is when the actual class of the data point is one and the predicted class is zero.

Based on the values of the above states, the evaluation criteria will be as follows:

3.2.1. Accuracy

Accuracy score in machine learning is an imperial assessment that shows the measurement of the number for correct predictions (i.e., the number of true positives add the number of true negatives) which is a model to make the total number of predictions. We achieve the number of correct predictions divide total number of predictions

$$Accuracy = \frac{true\ positives + true\ negatives}{total\ examples} \quad (8)$$

As can be seen in equation 8, this criterion determines whether the model is properly trained or not and how is its efficiency in general. But this measure does not provide detailed information about the model's performance.

3.2.2. Precision

Precision score is one indicator in machine learning model's performance, the quality of a positive prediction has been taken by the model. Precision achieve the number of true positives divided by the total number of positive predictions (i.e., the number of true positives add the number of false positives). When the value of false positives is high, precision will be a suitable measure estimate as shown in equation 9.

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \quad (9)$$

3.2.3. Recall

Recall, which is used in machine learning know in true positive rate (TPR), the percentage of data samples that a machine learning model correctly recognize as belonging to a class of interest, the “positive class”, out of the total samples for that class. When the value of false negatives is high, Recall will be a suitable measure as shown in equation 10.

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives} \quad (10)$$

3.2.4. F1 Score

F1 score is a machine learning imperial assessment that measures a model's accuracy. It consists of precision and recall scores of a model. The accuracy metric calculate which how a model has been made a correct prediction across the entire dataset. The F1 criterion is a suitable criterion for evaluating the Accuracy of an experiment, The F1 criterion is one at best and zero at worst. F1

score will be a suitable measurement as shown in equation 11.

$$Fl = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (11)$$

4. Results and discussion

In Table 2, we have shown a sample of the features used in this research. The dataset in this table had six geographical regions and each of them has nine features. 8 features are considered for modeling and the 9th feature is considered such as the target feature. The methods used in all seven regions have been implemented separately.

At this step, all the features in the data set have been examined and analyzed to extract and fix possible data problems, an example of which is shown in Figure 6.

As can be seen in Figure 6, the statistical information of the MWNS feature has been extracted, in which the number, in addition to the same statistical parameters such as maximum, minimum and similar values, the number of null values, duplicate values, zero values, as well as quartiles and skewness for Outliers are identified. Based on the statistical data of each feature, sufficient knowledge is obtained from it to pre-process the desired data if needed. To implement this research, the scikit-learn libraries in Python language have been used. To divide the data, the data separation method uses in two groups of training and test data, which is a method for measuring the performance quality of a machine learning algorithm. In general, this approach is used in any type of supervised learning algorithm. To divide the data, two methods are more common. The first method is simple division, which is usually divided into two groups of training and test data sets in the ratio of 70 to 30. This method is not suitable for small data sets and is unstable. The second method is cross validation. In this method, the k-fold technique is usually used to divide the data. In the second method, the data are first randomly divided into k groups with the same size. One of the groups is then reserved for testing and used to measure and evaluate the performance of the remaining k-1 groups. The cross-validation process will be repeated k times. In each fold, one data group remains intact to test the model trained with the help of other groups. This process will be repeated until all the groups are used as training and testing groups, and then the results are combined. By using all the data in both the training and testing sections in the k-fold validation technique, the overfitting problem of the model is reduced. In this research, the second method is used. For the initial evaluation and selection of the appropriate algorithm for prediction, learning algorithms were first implemented on the entire data and without regionalization, and their accuracy criteria obtained, the results of which are shown in Figure 7.

Table 2. Sample of features

	EQID	NUMS	FLOA	COLA	CWNS	CWEW	MWNS	MWEW	SDMG	*MMI*	*PGA*
0	1999_Duzce	*0.3*	*105.0*	*1.62*	*0.12*	*0.11*	*0.0*	*0.0*	*M*	*8.42*	*0.46*
1	1999_Duzce	*5.0*	*267.0*	*2.80*	*3.33*	*4.35*	*0.0*	*0.0*	*M*	*7.73*	*0.48*
2	1999_Duzce	*6.0*	*246.0*	*3.21*	*2.10*	*0.79*	*0.0*	*0.0*	*S*	*7.73*	*0.48*
3	1999_Duzce	*5.0*	*250.0*	*4.00*	*0.00*	*1.34*	*0.0*	*0.0*	*M*	*7.73*	*0.48*
4	1999_Duzce	*1.0*	*124.0*	*2.91*	*0.00*	*0.00*	*0.0*	*0.0*	*L*	*7.73*	*0.48*
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
657	.	.	.	.	.	.	.	.	.	.	.

Figure 6. Checking the statistical characteristics of the features

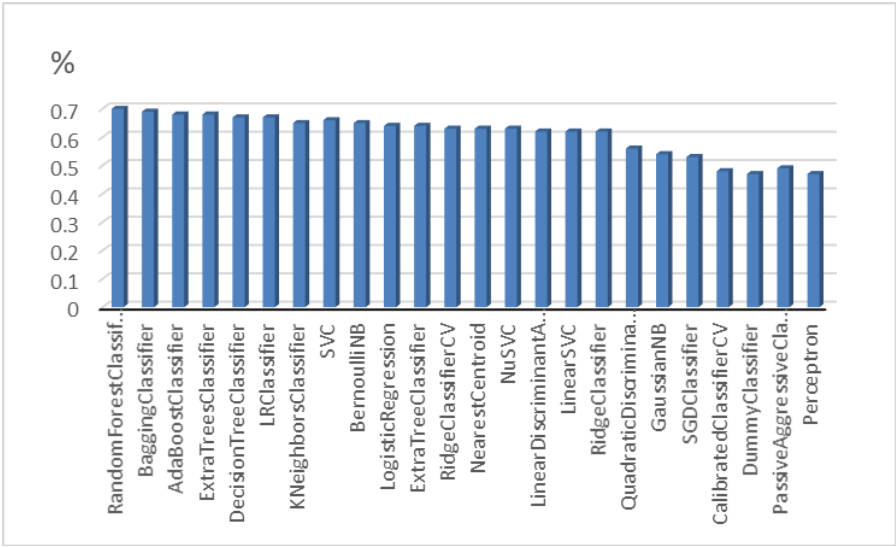


Figure 7. Accuracy criteria for ML algorithms

Based on the results of this stage, it can be seen that collective algorithms have performed better. Based on this, the random forest method, which has the highest accuracy value in collective algorithms, along with three other methods with the highest accuracy value among other method, were selected so that the following steps continue with these four methods.

In other words, the chosen methods are random forest, logistic, KNN and decision tree. In the next step, we improved the classic random forest based on the simple Bayes tree and created a probabilistic random forest, the hyper parameters of Bayes Random Tree Algorithm are listed in Table 3.

At this step, the data of all six stages were evaluated separately with four available methods, which include LR, KNN, DT, RF and the proposed modeling method. With the discussed criteria, the results are shown in Table 4. The following table shows the results based on six different geographical regions in the dataset. It can be seen that ensemble methods such as random forest have generally provided better results.

Table 3. Hyper parameters of the PRF

Parameter	Value
Feature selection method	Random
Number of features (k)	40
Method of selecting samples	Placement
Ratio of the number of samples (j)	90% of the training set
number of repetitions	N
learning algorithm	Simple Bayes tree
Data partition method	5 - Fold

Table 4. Results of different algorithms

Algorithms	EQ.	Accuracy	Precision	Recall	F1
Logistic Regression	Duzce	**0.73**	**0.69**	**0.72**	**0.69**
	Bingol	**0.63**	**0.64**	**0.63**	**0.63**
	Nepal	**0.69**	**0.67**	**0.70**	**0.66**
	Taiwan	**0.76**	**0.70**	**0.74**	**0.71**
	Ecuador	**0.67**	**0.68**	**0.72**	**0.64**
	Pohang	**0.55**	**0.54**	**0.54**	**0.54**
KNN	Duzce	**0.57**	**0.61**	**0.79**	**0.69**
	Bingol	**0.67**	**0.67**	**0.62**	**0.64**
	Nepal	**0.65**	**0.71**	**0.77**	**0.74**
	Taiwan	**0.67**	**0.86**	**0.67**	**0.75**
	Ecuador	**0.67**	**0.76**	**0.81**	**0.79**
	Pohang	**0.67**	**0.50**	**0.60**	**0.55**
Decision Tree	Duzce	**0.61**	**0.63**	**0.86**	**0.73**
	Bingol	**0.63**	**0.67**	**0.57**	**0.62**
	Nepal	**0.67**	**0.79**	**0.70**	**0.75**
	Taiwan	**0.73**	**0.88**	**0.78**	**0.82**
	Ecuador	**0.71**	**0.88**	**0.79**	**0.83**
	Pohang	**0.67**	**0.40**	**0.50**	**0.44**
Random Forest	Duzce	**0.72**	**0.63**	**0.67**	**0.63**
	Bingol	**0.70**	**0.70**	**0.73**	**0.70**
	Nepal	**0.71**	**0.70**	**0.82**	**0.73**
	Taiwan	**0.78**	**0.61**	**0.64**	**0.64**
	Ecuador	**0.71**	**0.68**	**0.59**	**0.71**
	Pohang	**0.60**	**0.45**	**0.45**	**0.45**
Statistical Random Forest	Duzce	**0.83**	**0.67**	**0.76**	**0.89**
	Bingol	**0.70**	**0.75**	**0.75**	**0.75**
	Nepal	**0.69**	**0.75**	**0.85**	**0.80**
	Taiwan	**0.73**	**0.78**	**0.88**	**0.82**
	Ecuador	**0.73**	**0.74**	**0.93**	**0.83**
	Pohang	**0.60**	**0.55**	**0.55**	**0.55**

By examining the values in Table 4, it can be concluded that among the classical methods of machine learning, the random forest algorithm has better results for all six regions. The reason is that in random forest, instead of one model, several models are created and classification is done based on them. The final result is based on the voting of the results of the models, so the classification error is reduced. In other words, random forest generally provides better results than single methods such as trees. The proposed method has a slight increase in three types such as accuracy criteria, different criteria and especially F1. there has been a good improvement to show the process of creating tree and evaluating them based on the Bayes rule is suitable for this problem. For a better comparison of the proposed method with other methods, the average values of random forest evaluation parameters that have the best results among other machine learning methods and the proposed method are shown in Figure 8.

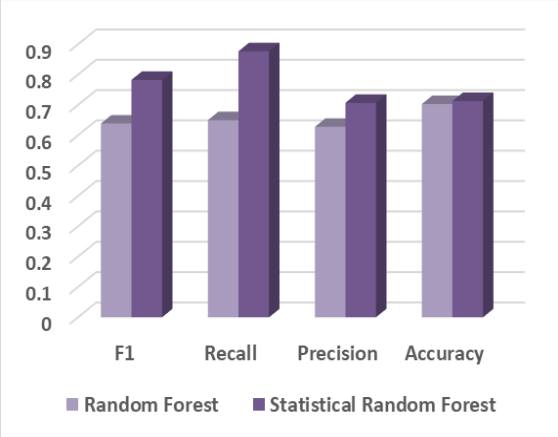


Figure 8. Comparison of random forest and the proposed method

It can be seen in Figure 8 that the proposed method for predicting the damage or non-damage of low-rise RC buildings performed well compared to other methods.

5. Conclusions

In this research, not only it has data with suitable geographical distribution, but also has a different classification. In the recommended manner with title in probabilistic random forest has been used the combination of decision tree and simple Bayes method. In the used manner of having smaller issues instead of significant issues caused classification algorithm makes better learning for users and caused better recognition in every sample by Bayes algorithm. In the other hand, using ensemble algorithm decrease errors classification. The collection of data in this article use in different places that probabilistic random forest performs the better results. According to caught results be understood the ensemble algorithm and correct distribution not only causes increase prediction accuracy, but also decreases errors.

Improved machine learning algorithms implemented in the RVS prediction techniques can help to achieve sustainable cities and promotion of more resilience cities. In such a way that irreparable loss of life and money can be prevented by identifying vulnerable areas, allocating budget and prioritizing the costs of rehabilitation and retrofitting. For future research, the method presented in this article can be used for other types of buildings.

6. References

[1]. F. F. E. M. A. (US), Rapid Visual Screening of Buildings for Potential Seismic Hazards: A Handbook (FEMA P-154), *Applied Technological Council (ATC)*.

[2]. Harirchian, E.; Aghakouchaki Hosseini, S.E.; Jadhav, K.; Kumari, V.; Rasulzade, S.; Işık, E.; Wasif, M. & Lahmer. T. A. (2021). Review on application of soft computing techniques for the rapid visual safety evaluation and damage classification of existing buildings. *Journal of Building Engineering*. 43, 102536. https://doi.org/10.1016/j.jobe.2021.102536

[3]. Shaheryar, A.; Abarca, A.; Perrone, D. & Monteiro, R. (2022). Large-scale seismic assessment of RC buildings through rapid visual screening. *International Journal of Disaster Risk Reduction*. 80, 103219. https://doi.org/10.1016/j.ijdrr.2022.103219

[4]. Hassan, Ahmed F., & Mete A. Sozen. (1997). Seismic vulnerability assessment of low-rise buildings in regions with infrequent earthquakes. *ACI structural journal*, 94(1). pp31-39

[5]. Elyasi, N.; Kim, E. & Yeum. C.M. (2023). A Machine-Learning-Based Seismic Vulnerability Assessment Approach for Low-Rise RC Buildings. *Journal of Earthquake Engineering*, 1-17. https://doi.org/10.1080/13632469.2023.2220033

[6]. Chollet F. (2017). Deep learning with Python, *Manning Publications*, ISBN 9781617296864

[7]. Kunapuli, G. (2023). Ensemble Methods for Machine Learning, *Manning Publications,* Shelter Island. ISBN 1617297135.

[8]. D. Mienye & Y. Sun, (2022) A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, 10, pp. 99129-99149. https://doi.org/10.1109/ACCESS.2022.3207287

[9]. Kuncheva, L. and Whitaker, C., (2003), Measures of diversity in classifier ensembles. *Machine Learning*, 51, pp. 181-207. https://doi.org/10.1023/A:1022859003006

[10]. Le Cessie S. & Van Houwelingen JC. (1992), Ridge estimators in logistic regression. *J R Stat Soc Ser C (Appl Stat),*, 41(1), 191–201. https://doi.org/10.2307/2347628

[11]. Pedregosa F.; Varoquaux G.; Gramfort A.; Michel V.; Thirion B.; Grisel O.; Blondel M.; Prettenhofer P.; Weiss R.; Dubourg V.; et al. (2011), Scikit-learn: machine learning in python. *J Mach Learn Res*., 12, 2825–30.

[12]. G´eron A. (2019), Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems, *O'Reilly Media*.

[13]. Breiman L. (2001). Random forests, *Machine Learning*, 45, pp5–32. https://doi.org/10.1023/A:1010933404324

[14]. Guo, X. and Hao, P. (2021), Using a Random Forest Model to Predict the Location of Potential Damage on Asphalt Pavement. *Appl. Sci.,* 11, 10396. https://doi.org/10.3390/app112110396

[15]. Belmokre, A., Mihoubi, M.K. & Santillán, D. (2019), Analysis of Dam Behavior by Statistical Models: Application of the Random Forest Approach. *KSCE J Civ Eng*, 23, 4800–4811. https://doi.org/10.1007/s12205-019-0339-0

[16]. Qinghua, H., Ma, Q., Dang, D.& Xu, J. (2023), Modal Parameters Prediction and Damage Detection of Space Grid Structure under Environmental Effects Using Stacked Ensemble Learning, *Structural Control and Health Monitoring*, pp1-24. https://doi.org/10.1155/2023/5687265

[17]. Khodaparasti, M.; Alijamaat, A; & Pouraminian, M. (2023). Prediction of the concrete compressive strength using improved random forest algorithm, *J Build Rehabil,* 8(92). https://doi.org/10.1007/s41024-023-00337-8

[18]. John GH., and Langley P. (1995), Estimating continuous distributions in Bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, *Morgan Kaufmann Publishers Inc*. pp338–345.

[19]. Sarker IH. (2019), A machine learning based robust prediction model for real-life mobile phone data. *Internet Things*, 5, pp180–93. https://doi.org/10.1016/j.iot.2019.01.007

[20]. Kohavi, R. (1996). Scaling Up the Accuracy of NaïveBayes Classifiers a Decision Tree Hybrid. *Second International Conference on Knowledge Discovery and Data Mining*, pp. 202-207