Comparative Analysis of Building Insurance Prediction Using Some Machine Learning Algorithms

Chukwuebuka Joseph Ejiyi, Zhen Qin*, Abdulhaq Adetunji Salako, Monday Nkanta Happy, Grace Ugochi Nneji, Chiagoziem Chima Ukwuoma, Ijeoma Amuche Chikwendu, Ji Gen*

University of Electronic Science and Technology of China, Chengdu (China)

Received 8 October 2021 | Accepted 2 February 2022 | Published 17 February 2022

LA UNIVERSIDAD
EN INTERNET

Abstract

In finance and management, insurance is a product that tends to reduce or eliminate in totality or partially the loss caused due to different risks. Various factors affect house insurance claims, some of which contribute to formulating insurance policies including specific features that the house has. Machine Learning (ML) when brought into the field of insurance would enable seamless formulation of insurance policies with a better performance which will also save time. Various classification algorithms have been used since they have a long history and have also got some modifications for optimum functionality. To illustrate the performance of each of the ML algorithms that we used here, we analyzed an insurance dataset drawn from Zindi Africa competition which is said to be from Olusola Insurance Company in Lagos Nigeria. This study therefore, compares the performance of Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Kernel Support Vector Machine (kSVM), Naïve Bayes (NB), and Random Forest (RF) Regressors on a dataset got from Zindi.africa competition and their performances are checked using not only accuracy and precision metrics but also recall, and F1 score metrics, all displayed on the confusion matrix. The accuracy result shows that logistic regression and Kernel SVM both gave 78% but kSVM outperformed LR in precision with a percentage of 70.8% for kSVM and 64.8% for LR showing that kSVM offered the best result.

Keywords

Claim, Insurance, Machine Learning, Prediction, Regressors.

DOI: 10.9781/ijimai.2022.02.005

I. Introduction

INSURANCE is described as a risk management strategy used to hedge against the risk of accidents. Usually, the underwriter provides the insurance while the policyholder buys the said insurance. The policyholder receives an insurance policy with specifications on the circumstances and conditions under which the underwriter will give a stipulated or required compensation to the insured as they continue to be in good standing with the insurance company (that is, they pay their premium). If there is an experience of a loss by the policyholder which is potentially included in the insurance policy, he/she puts forward or files a claim to the underwriter [1], [2]. People have many reasons for insuring their property, but one of the major and basic reasons for insurance is usually because it gives a sense of safety and security

Lagos State, one of the states in Nigeria that is popular for commerce and industry has an insurance company named Olusola Insurance Company (OIC). Being among the most renowned and famous insurance companies available in Lagos state that is into building insurance and not only that but also being one of the oldest. The recent recurrent collapse of buildings in the state has become a big concern to the landlords as well as this insurance company because of their insurance policy. Therefore, the OIC sought a way to be able to theorize by prediction if a policyholder will file a claim in case of the collapse of his building or not and if such an individual will be qualified for the insurance. They made available the dataset collected differently from various sources at different times at the site of Zindi Africa. We analyzed the data for the prediction of claims or no claims with respect to the insurance. We have the task to design and build a model that is applicable for prediction with which it could be determined if a particular building will have or is supposed to have an insurance claim during the specified period or not. Since machine learning algorithms such as LR, RF, kSVM, KNN, NB, and DT can be used, we compared them with the motive of finding out the best predictive algorithm with respect to this data made available by OIC.

Each of the classification algorithms has been used for various predictions. [3], [4], [5] among others used logistic regression only and have seen it to be very efficient. [6], [7], [8] are some of the works that employed the prediction algorithm the Random forest and it was quite good in performance. [9], [10], [11] are some of the researchers that used the Decision tree for their prediction. For example, Amra and Maghari analyzed students' performance using KNN and Naïve Bayesian algorithms and discovered that Naïve Bayes with an accuracy of 93.2% performed much better than KNN with an accuracy of 63.5% [12].

For this comparison, we used Exploratory Data Analysis (EDA) first to help find out the underlying pattern, discover and spot irregularities, frame the possible hypothesis, and check assumptions with the aim to find a good fitting model. The model will then be anchored on the

* Corresponding author.

E-mail addresses: qinzhen@uestc.edu.cn (Z. Qin), jgeng@uestc.edu.cn (J. Gen).

specified building characteristics. After which the target variable, the claim – that is, if the building in question has a minimum of one claim during the said insured period or if it does not have. After the EDA, the data is preprocessed to make the data "parseable" by the machine. Feature selection and then model training and evaluation followed. We used LR, RF, kSVM, KNN, NB, and DT algorithms for the regression analysis, having in mind that the accuracy will be checked, to ascertain the performance of each regressor we used a confusion matrix and determine the best predictive algorithm among them and then compared their performance.

The other parts of this paper have been organized thus: Section II presents the related works for all the models for prediction. The background or what may be called overview was covered in section III. The model design and experimental analysis were the focus of section IV. The result and discussion were covered in section V before the paper was concluded in section VI.

II. Related Works

Bhat and Gandhi reported that the ANN-based framework is gaining popularity since it performs well in prediction [13]. They implemented a normalization technique to improve the performance of ANN and also used the analysis of correlation coefficient to find and determine the best input to the ANN framework and gave a better performance than the statistical model. Various regression techniques have been proposed which can be used for prediction apart from Logistic regression. In [14], Multiple Linear Regression (MLR) is said to be a kind used in which more than one attributes are for prediction. In [15], [16], Ridge regression (RR) and LASSO regressions (LR) were used; here, Ridge regression does the regularization of the regression coefficient while LASSO Regression uses L1 penalty, being the only thing that differentiates it from RR. In [17] a regression form called the Elastic Net (ER) was used as a penalization method. Mislan *et. al.* initiated a Back Propagation Neural Network model with double hidden layers for rainfall prediction [16]. In [18], the NB classification algorithm was used for classification before prediction. In [19] - [20] used Artificial Neural Network theory for prediction. The Linear Regression model was also used in [20], [21].

Some other tools employed for prediction by other authors are as follows; Geographically Weighted Regression. In [22], Bayesian Linear Regression [23]. Support Vector Machine Regression (SVMR) was used by Paniagua Tineo *et. Al.* [19] to predict daily maximum temperature. From their work, the SVMR algorithm was concluded to be able to produce accurate temperature predictions after 24 hours. Onan [24] identified web classification as a research direction with great importance in data science. The author presented different feature selections as well as four different ensemble learning methods on the basis of four different base learners(NB, KNN, C4.5 algorithm together with FURIA algorithm). The author concluded that in web page classification, ensemble learning, as well as feature selection, has the capacity to enhance the predictive ability of classifiers. In another work by Onan *et al* [25], they reported that in order to reduce the training time and also develop a robust and efficient classification model, feature selection is necessary. [26] and [27] achieve good performance of classification on language using various classification models employing feature selection methods that are suitable for such classifications.

Many other algorithms have been employed and several other models proposed and tested for various predictions as reported by [28]. Abhishek *et. Al.* [29] proposed and implemented a model that predicted temperature using a backpropagation neural network. Shobha *et. Al.* [30] used a clustering-based analysis approach to monitor whether based on meteorological data. The authors studied data from agricultural meteorological patterns collected from the Meteorological Centre of Bengaluru district and determined very important agricultural parameters like minimum temperature, maximum temperature, relative humidity, rainfall, and pan evaporation using not only K-Means but also hierarchical clustering which are extremely critical for agriculture. Gill *et. Al.* [31] proposed a model with a backpropagation employing a genetic algorithm. From their work, it became observable that genetic algorithms can be effectively used for prediction alongside backpropagation neural networks. Each work employed EDA to enable better performance of each algorithm that was used.

A lot of things have been predicted, like rainfall, weather, etc. using various algorithms. Paniagua-Tineo *et. Al.* [32] worked on maximum daily temperature prediction by employing Support Vector Machine Regression (SVMR). It was concluded by them, that the SVMR algorithm can produce accurate temperature prediction after 24 hours. Abdel-Aal *et. Al.* proposed an abductive networks approach for the prediction of temperature on an hourly basis [33], which was able to predict the temperature after an hour and also after a day. A modified type of Support Vector Machine (SVM) called Multi-view Least Squares (LS-SVM) regression for black-box temperature prediction was proposed by Houthuys *et. Al.* [34]. The goal of multi-view LS-SVM is to improve the model's performance by taking information from all views into account as there is an appreciable number of observations in black-box weather forecasting.

Prediction of short-term wind power with the aid of empirical mode decomposition-based GA-SYR was experimented by Xie *et. Al.* [35]. First, the wind speed data from NWP is decomposed into the EMD components, including multiple intrinsic mode functions (IMFs) as well as one residue. Thereafter, a Genetic Algorithm Support Vector Regression model (GA-SVR) is used to build models of all components. Similarly, another method referred to as short-term wind power forecasting was implemented by Peng *et. Al.* [36] using numerical weather prediction and error correction methods. [37] Zhang *et. Al.* used SVM by training multiclass predictors and adjusting SVM parameters using Particle Swarm Optimization (PSO). In [38] Papantoniou *et. Al* utilized data obtained from about four European cities to display or introduce the implementation and then evaluation of different neural network-based algorithms used in identification.

Da-Chun Wu *et al* [39] used ANN to forecast or predict air compressor load of different compressors at different times and under different conditions. They also investigated the prediction of the electrical demand peak with ANNs and SVM and discovered that integration of ANNs to SVM gave a significant improvement to the accuracy of the prediction. Priyadarshini Patil *et al* [40] compared the performance of SVM, RF, and ANN in potato blight disease and discovered that the ANN performed better than SVM and RF. From their experiment, ANN gave an accuracy of 92% which was better than 84% and 79% respectively of SVM and RF.

Xiaohu T *et al* [41] used the DT algorithm to predict the winning team in the Chinese super league and got an accuracy of 57.7% when other factors are put into consideration. In another prediction, Nwulu [42] used DT to predict the price of crude oil from data gathered which covered about 24 years. According to his work, DT outperformed other models that he used and had a less computational period. They [43] basically used DT to predict churn as well as KNN and DT gave an accuracy of about 93% which is a good accuracy [43].

Raj *et al* [44] concluded from their comparison that SVM has better performance (accuracy of 82%) when compared to Naïve Bayes (62.5%) when they compared the SVM and NB classifiers used in diabetes prediction. Bayindir *et al* used an NB classifier to predict the daily energy generated from an installed photovoltaic system [45] and an accuracy of 82.2% was obtained.

Salim *et al* predicted the timely graduation of students in Indonesia using KNN because of KNN's robustness on noisy data and its ability to train on a large dataset [46]. With the advent and rise of the technology of machine learning, it has been used in the prediction of flight delays as well [47]. Machine learning algorithms that have been used for this kind of prediction include random forest, decision tree, logistic regression, SVM as well as K-nearest neighbor algorithms [48], [49], [50].

Linear regression has been identified as the basic regression model that has been used for prediction, it takes into account the variation that exists between the variables called independent and those that are dependent also according to [51]. These researchers [51] compared the regression models which will have the capacity to predict graduate admission and found out that linear regression outperformed other models on their dataset. The models compared include Linear regression, Support Vector Machine, Decision Tree Regression, and Random Forest regression. Linear Regression gave an output of 0.00480149 for MSE and 0.72486310 for R2 which is the best among the models [51].

In all these, limited literature was found with respect to insurance prediction. And with the growth in the need for the insurance of properties such as cars, houses, and others, data is needed to be able to build models that will be useful for insurance prediction.

III. Background

Machine Learning (ML) has over time been viewed as a branch or more precisely a subcategory of Artificial Intelligence (AI) that learns computer algorithms and improves via experience [48], [52]. It has grown and become very popular, it has also found usefulness in many fields not occasionally but on daily basis [49], [53]. ML uses training data or sample data to build models which they use to make decisions or predictions in this case without further programming. So ML gives the system all it needs to learn and understand by itself and consequently gives or makes a prediction for the unknown outputs [50]. Machine learning algorithm has its performance depends on the training success, dataset availability, data preprocessing, selection of attributes among others.

Regression analysis is primarily used for two conceptually distinct purposes. First, for prediction as well as forecasting, where its application has a remarkable connection with the field of ML. which second identified usefulness, in some situations, to hypothesize the underlying relationships that exist between the variables that are independent and those that are dependent. It is very important to note that regressions by themselves only bring to light the relationships that are found between a variable that is dependent and another set that is a collection of variables that are independent and in a fixed dataset. To use regressions for prediction or to hypothesize causal relationships, respectively, a researcher must ensure to carefully bring to light the reason why the existing relationships are assumed to have predictive power for a new context or why a relationship between two variables is thought to possess a causal interpretation. The latter is considered crucially important especially when researchers hope to make an estimation of the causal relationships using observational data [54]. The techniques used in the study are introduced in the following subsections.

A. Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) originally inspired by the nervous system of animals, process data in a manner similar to the brain of mammals. They have the capacity to forecast difficult problems. by computationally learning a set of input data and consequently giving out an output that is desirable. ANN is usually defined as a framework instead of an algorithm because it acts as the basis for many machine learning algorithms [55], for complex data (input) processing. It has got many applications in different fields and so we are applying it to this dataset for prediction.

ANN has been identified as a tool that has found usefulness for both regression and classification since it can model systems that have a non-linear relationship [56]. One peculiar thing about ANN is that it is structured such that after training, it has the capacity to give outputs that are reliable and will do that quite fast even if the data is noisy or in cases where some information is missing from the data [57]. ANN in its structure has hidden layers that can grow depending on the network size, this growth helps the network memorize more of the data but it increases the training time [58].

B. Linear Classifier

Linear Regression (LR) is estimated with the Maximum Likelihood Estimation (MLE) approach and provides constant output. The MLE is just“a “likelihood” maximization method which is a function that measures the parameters that seem likely to give the observed data and accepts a joint probability mass function. Statistically speaking, MLE sets the mean and variance as parameters in measuring the particular parametric values for a given model. This set of parameters can be adopted to predict the data required in a normal distribution and also accepts a joint probability mass function [59]. A function known as the logistic function or sigmoid function shown in (1) gives an “S” shaped curve which takes any real number value and maps the number into the interval of 0 and 1. The predicted value becomes 1 (one) in case the predicted values go to positive infinity and 0 (zero) if it is negative infinity. Logistic regression is majorly used for classification but is also useful in solving regression problems since it shows good performance.

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

C. Decision Tree Classifier

Decision Tree (DT) is considered one of the easiest and most popular classification algorithms to learn and interpret. It can be applied in solving classification and regression problems. The decision tree looks more like a flow chart that has the structure as a tree would have in which the features are represented with internal node, while the branches and the leaf node represent the decision rule and the leaf node respectively. But unlike the normal tree where the root is at the base, the uppermost part of the decision tree is where the root node is located. It learns to partition using the attribute values in a recursive manner otherwise called recursive partitioning. The decision tree as the name implies helps in decision making. Its complexity is a function of the number of attributes and records in the dataset [60]. The decision tree makes predictions in a tree-like manner just as people would make decisions when faced with certain challenges especially when there are two options to decide from. When one option is taken, you may have to make other choices based on the one you have decided on and it continues until the result hoped for or expected is got. This is basically the framework of the decision tree [41] and classification is based on characteristics. The algorithm is relatively easy to implement because of how easy it is to understand, and according to Xiaohu T *et al* [41], it is applicable for data analysis and forecasting. Xiaohu T *et al* [41] have also earlier reported that DT is based on instances and that it is referred to as an “inductive learning algorithm.” DT has also been shown to comprise of some major steps which are feature selection, generation of the decision tree and finally pruning of the tree. Some of the identified advantages of DT are easy calculation and workload, simple and easy to understand, interpret, analyze, and a high degree of accuracy [61]. The basic and fundamental idea behind the decision tree algorithm has been identified to be recursive partitioning which is a statistical technique for the analysis of multivariable [42].

Like in the Agricultural point of view where trees are more popular and pruning is done for more productivity, pruning is also usually applied to DT to improve the algorithm's performance [42]. The noise which is one of the characteristics of raw data is said to be easily managed by the DT algorithm [43] because DT has the ability to avoid overfitting by pruning.

D. Random Forest Classifier

Random Forest (RF) on the other hand is also a supervised learning algorithm also used for both regression and classification. Just as the name (forest) suggests, it is supposed to be composed of many trees. The robustness of the forest will then be a function of the number of trees in it. RF functions by creating decision trees on selected data samples on a random basis. Then it usually gets a prediction from every tree and selects the solution that is the best by voting. By this, it provides a very nice indicator of the important features. So, a random forest algorithm is seen as a collection of many decision tree classifiers, each decision tree is got using an appropriate characteristic selection gauge like information gain. Individual trees are dependent on an independent random sample and each tree votes the most frequent class, unlike the regression where the mean of the entire tree results is assumed to be the final output. The random forest has been proven to offer a good feature selection indicator [62]. Its functioning is divided into the following steps; Samples are selected randomly from a given dataset, a decision tree is then constructed for each sample, getting a prediction from individual trees. After that, a vote is performed for each predicted output or result and then the prediction with the highest voted is selected as the final prediction [63].

In all these, Random forest differs from decision tree in the following ways; Random forest keeps from overfitting whereas Decision tree may run into it. It, therefore, follows that RF manages the challenge of overfitting more than DT. Random forest is composed of multiple decision trees, although the Decision tree is faster computationally. It is not difficult to infer that the Decision tree is easier to interpret when compared to the Dandom forest.

E. K-Nearest Neighbor Classifier

Here in K-Nearest Neighbor (KNN), an unknown data point is categorized into its nearest neighbor which is already defined and determined. The nearest neighbor is figured out by k-value which works out the actual neighbors and at the same time the classes that belong to a particular data point [12]. On some occasions, it requires not only one nearest neighbor to ascertain the particular datapoint's class. In KNN it is usually necessary for data points to be in memory at runtime, it is also called "memory-based technique" [12]. There were some improvements proposed by some researchers on the pioneer KNN but the computational complexity and memory requirements have remained unchanged. Nevertheless, the memory requirements can be managed well if there is a reduction in the size of the dataset used, thereby reducing comprehensively the repeated training sample pattern. Some data points that are perceived to have no effects on the result are the ones that are more advisable to remove. The nearest feature line, ball tree, tunable metric, k-d tree, principal axis search tree, and orthogonal search tree are some of the algorithms that have been identified to bring increment to the speed of KNN [12].

F. Naïve Bayes (NB) Classifier

Naïve Bayes (NB) as a machine learning algorithm is designed in such a way that it can accomplish classification tasks. It has been reasoned that its popularity is credited to the fact that it can be written into code quite easily with less time. NB has also been identified as an algorithm that can be implemented in real-time prediction and organizations find it very useful in bringing quick answers to users' request(s). NB classifier is basically anchored on the theorem proposed by Bayes, as such, it is seen as a conditional probabilistic classifier [64]. It is also denoted as the Generative learning model on some occasions [65]. The algorithm is called Naïve Bayes because the existence of a particular feature does not depend on the existence of another feature which is the same principle in conditional probability [44] and it is very helpful for a very large dataset classification. NB performs well and is said to be most suitable for data with high dimensionality [12]. Some of the real-world scenarios where NB has found application are in Recommendation System, Real-Time Prediction, Multiclass prediction, Sentiment Analysis, Text Classification, and the popular Spam Filtering.

The general principle of NB hinges on conditional probability and understanding it is paramount to understanding the NB algorithm.

G. Support Vector Classifier

Support Vector Machine (SVM) is linear naturally [66] and is known to support Linear regression as well as non-linear regression, this feature made it possible for SVM to be referred to as Support Vector Regression as well. As a model which is classified as a supervised machine learning model, it can to make predictions from the earlier learned data. It has been suggested that SVM gives good results when the dataset to analyze is not much. Support Vector Machine a discriminative classifier is expressed by a separating line or hyperplane. Given training data, it outputs a hyperplane that categorizes the new data set. Kernel Support Vector Machine tends to discover the best hyperplane that separates data in a Hilbert space. This best hyperplane is selected to maximize the margins between the classes (usually two, since Kernel SVM is a binary classifier). The kernel functions [66] of the SVM are activated or brought into the light so as to make the linearity of the SVM classifier got by dot product to nonlinearity. The Kernel SVMs allow the hyperplane's extreme margin to adapt in a feature space that has been transformed and has the advantage of taking care of classification difficulties over conventional SVM [67]. SVM aggregates the two classes that are to be classified using support vectors which are the extreme data points separated by hyperplane [44]. In this context, the hyperplane stands as the classification between the objects.

IV. Model Design and Experimental Analysis

A. Model Design

Our work looks at the comparison of the following machine learning algorithms Logistic regression, Decision Tree, Kernel SVM, Random Forest, Naïve Bayes, and K-Nearest neighbor. The basic information of the algorithms and the principles of how they function have been discussed in the previous section. The algorithms were given the same input data which have the same ratio of training and testing data. The data were preprocessed as explained in section IV*B* to reduce noise and undesirable characteristics before they were divided randomly into training and testing sets. The training set was used for the training of the algorithm and the performance of the algorithms tested with the testing sets.

The diagram above Fig. 1 is the flowchart of the model we used for this work.

B. Dataset

The dataset we used was got from zindi.africa [68] the data comprises collected information from 2012 to 2016 which are the years of observation. It was said that for the period of observation, the dataset was collected by various people. The variables in the dataset with their descriptions are shown in Table. I.

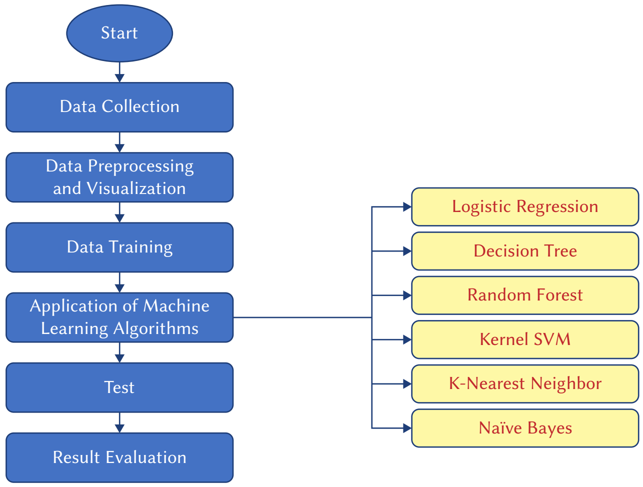


Fig. 1. Block diagram of the model.

TABLE I. Variables and Their Description

No.	Variable	Description
1	Customer ID	Identification number for the Policyholder
2	Year of observation	Year of observation for the insured policy
3	Insured period	Duration of insurance policy in Olusola Insurance (Eg: Full-year insurance, Policy Duration= 1; 6 months = 0.5
4	Residential	Whether the building is residential or not
5	Building painted	Whether the building is painted or not painted (N-Painted, V-Not Painted)
6	Building fenced	Whether the building is fenced or not fenced (N-Fenced, V-Not Fenced)
7	Garden	Whether the building has at least one garden or not (V-has garden; O-no garden).
8	Settlement	The area where the building is located. (R-rural area; U- urban area)
9	Building dimension	Size of the insured building in m^2
10	Building type	The type of building (Type 1, 2, 3, 4)
11	Date of occupancy	Date or Year the building was first occupied
12	Number of windows	Number of windows in the building
13	Geo-code	Geographical location code of the insured building
14	Claim	Target variable. (0: no claim, 1: at least one claim over the insured period).

C. Tools Used

The following tools categorized as python libraries were used for the implementation of the algorithms: The Seaborn that helps with the generation of heatmaps, the Scikit learn/sklearn which helps to ensure that the algorithms are implemented, the Pandas helps in operations that are data-related while the Matplotlib is for plotting of the various required plots. The Jupyter notebook was also used for the writing of the python codes which were consequently executed on Keras/ TensorFlow framework.

D. Data Preprocessing and Visualizations

After data collection, the next step is data preprocessing followed by Data Integration, Data Transformation, and then reduction [69]. Data preprocessing is an essential and basic step in the process of knowledge discovery; because the data obtained from the logs may be incomplete, noisy, or inconsistent [70]. The most promising attributes of quality data include completeness, consistency, and timeliness. The performance of a mining algorithm depends on the quality of the data. But, the real-world data is incomplete and uncertain. The incompleteness of the data can be easily identified and its elimination is sometimes acceptable. But, identification of the inconsistencies in the data is very difficult and even a very negligible amount of inconsistency in data degrades the performance of the mining algorithm at a very high rate. The existence of inconsistencies in the training data affects the performance of the mining algorithm and the removal of such inconsistencies improves the performance of the mining algorithm [70].

Data visualization, on the other hand, aims to transmit the data clearly and effectively via graphical representation. Data visualization has found extensive use in many scenarios like reporting at work, task-progress tracking among others. One of the most popularly used advantages of the visualization technique is in discovering data relationships that may be hard to identify or observe by merely looking at the raw data [71]. These days, people have also used data visualization to design funny and interesting graphics.

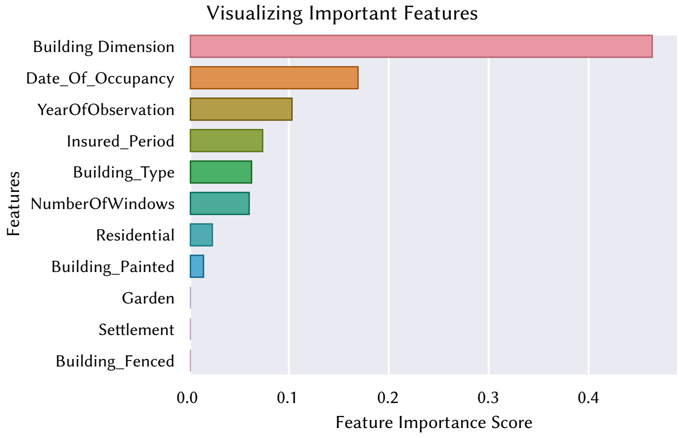


Fig. 2. Feature Importance score.

We generated feature importance of the train data using Gini importance and Fig. 2 above shows the importance from the highest to the lowest. And the correlation matrix is shown in Fig. 3.

Since we have to design a system that will predict the probability of having at least one claim over the period that the building was under insurance. The model will be accessed based on the features of the building. The target variable, in this case, being claim:

- 1 if the building in question has at least one claim over the said period of insurance.
- 0 if the said building does not have a claim over the said period of insurance.

The data collected from the site was then preprocessed. The data were divided into a ratio of 7:3 training (70%) and testing (30%). Only the decision tree and the random forests were tested with 20% of the dataset and trained with 80% of them. The dataset was given a good description of the variables for each of the columns. Selection of the data or division of the dataset was done randomly and the models' performance was checked afterward.

Since the data is noisy, we use methods of dealing with noisy data to fill in the data. From the data set both of the training and testing, 4 of the variables contain noisy data (null values). Those variables are; Garden, Building dimension, Date of occupancy, and Geo-code for both the training and testing dataset. The data were normalized and the vacant data was filled with modal values for the variables Geo-code and Garden, and mean value for the Building dimension and Date of occupancy.

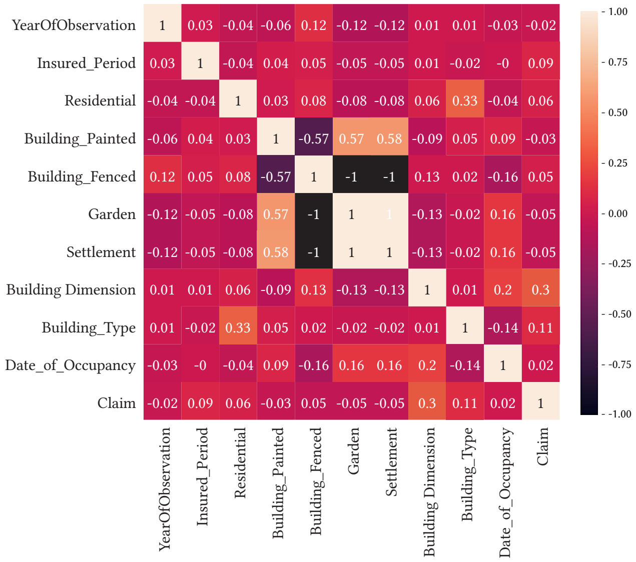


Fig. 3. Figure Correlation matrix.

The input parameters used for the experimental phase and purpose are shown in Table II. We also included the k-fold cross-validation strategy used. We used the k-fold with the best performance for each algorithm. For other models, we used 10 but the decision tree and the random forest seemed to be overfitting with the same value of k-fold, as such we used 5 for them.

TABLE II. Input Parameters

Model name	K-fold	Input parameters

Decision Tree	K = 5	Criterion = entropy
		Splitter = best
		Random_state = 0
		Max_dept = none

KNN	K = 10	Number of neighbours = 5
		Weights = uniform
		Leaf-size = 30
		Algorithm = Kd tree

Logistic Regression	K= 10	Tolerance = 1e-4
		Class weight = balanced
		Solver = Newton cg
		Random state = 0

Kernel SVM	K=10	Kernel = RBF
		Random_state = 0
		Tolerance = 1e-4
		Class weight = balanced

Naïve Bayes'	K=10	Type = Gaussian NB
		Priors = none
		Var_smoothing = 1e-9

Random Forest	K= 5	Number of trees = 10
		Criterion = entropy
		Random_state = 0

E. Performance Evaluation Criteria

The basis for the comparison of these algorithms anchors on some known performance criteria some of which are outlined below with a short description for each of them. We choose these criteria because they are popular and easy to understand in addition to the fact that they can be applied to all the algorithms for easier comparison.

Accuracy: This gives the estimate of the percentage of the actual rate values of all classes. A higher accuracy shows better performance (higher is better). Accuracy simply put is a ratio of appropriately predicted observation to the total observations. The formula for the calculation of accuracy is shown in (2).

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \qquad (2)$$

Precision: This determines the percentage rate of the true positive values for the relevant elements to the irrelevant ones. As with accuracy, higher the precision percentage, higher relevant results were retrieved than the irrelevant ones (higher is better). The formula for the estimation of precision is shown in (3).

$$\text{Precision} = TP / (TP + FP) \qquad (3)$$

Recall: Also referred to as Sensitivity Measure or True Positive Rate is seen as the fraction of a true positive rate of relevant values. A higher ratio means higher retrieved relevant elements (higher is better). The method used for the estimation of recall is shown in (4).

$$\text{Recall} = TP / (TP + FN) \qquad (4)$$

F-Measure (or F1 score): It is a weighted mean of both precision and recall. The upper threshold value of the F1 score is usually 1, which stands for the best score, and the lower threshold value 0, which means the worst score (higher is better). The formula for the estimation of the F1 score is shown in (5).

$$\text{F1Score} = 2^{*} (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \qquad (5)$$

Confusion Matrix

The result from the matrix permits us to have more detailed information about the results for the algorithms used, by making available the number truly predicted positive and negative from the results.

True Positives (TP): These refer to the correctly predicted values that are positive, meaning that the value of the class is actually yes and the model predicted the class also as yes. In this case, the model predicted that a house has a claim and it actually does. The True Positive values estimated from the models we used are shown in Table IV.

True Negatives (TN): These refer to the predicted negative values that were predicted correctly, meaning that the value of the class is no actually and the model predicted the class also as no. In the case of our study, the model predicted no claim when the house does not have any claim. The True negative values estimated from the models we used are shown in Table IV.

False Positives (FP): This is sometimes called Type I error and occurs when the class is no in the real sense while the model predicted it as yes. Taking inference from our study, if the model predicts a claim when it is supposed to be no claim. The False Positive values estimated from the models we used are shown in Table IV.

False Negatives (FN): This one is sometimes called Type II error and occurs when the class is yes in an actual sense but the model predicted it as no. When viewed in the sense of our study, when a model predicts no claim but it is supposed to have a claim. The False Negative values estimated from the models we used are shown in Table IV.

Precision and Recall can all be calculated with the above-got parameters as shown from equations (2) to (5).

V. Result and Discussion

This work looked at the capability of the aforementioned machine learning algorithm's ability to predict claim or no claim on the insurance according to the provided dataset from Olusola Insurance Company. The used input parameters are displayed in Table II and the results in Tables III and IV.

The data was collected by different sources or people according to [68] and at different points. That made the data have some abnormalities, like missing information. This is so because the span of the year of the collection was a bit long and some data may not have been considered important at some point. As time went by, some became relevant and more emphasis was laid on them. Also, some people may not be willing to give some details or may have forgotten some data which might have amounted to the missing of these data. Not minding the cause of the missing data, during the data preprocessing these were taken care of as explained in section IV*B*. The dataset was split into two for training and testing in the ratio of 7:3 for most of the models (KNN, KSVM, LR, and NB) and 8:2 for DT and RF because of the k-fold value used for them.

TABLE III. Model Results

Model name	Accuracy	Precision	Recall	F1
Decision Tree	0.680	0.311	0.329	0.320
KNN	0.757	0.444	0.257	0.325
Kernel SVM	0.788	0.708	0.123	0.210
Logistic Regression	0.788	0.648	0.158	0.254
Naïve Bayes'	0.773	0.507	0.278	0.359
Random Forest	0.756	0.444	0.278	0.342

From Table III above, the Logistic Regression and kSVM both gave an accuracy of 78.8% which is better than others. For a better assessment, the precisions are considered in which kSVM gave a better performance than logistic regression. But both in the recall and F1, LR has relatively higher performance than kSVM. Although NB has a very close accuracy (77.3%) to kSVM and LR and also close precision value to them too, its recall was joint second to the best, and it produced the best F1 score. The model evaluation table shows explicitly that kSVM gave a better performance than the other models.

The logistic regression model performed well possibly because of its ability to learn and perform optimally if the value of the variable is categorical and requests binary output as in this case. The kSVM also performed well too although the data available for learning can be considered small because kSVM gives a good performance and optimally for cases that are linearly separable.

We will take a good look at the confusion matrix which is a convenient presentation of the measured accuracy of a model with two or more classes. Table III presents the values for the components of the confusion matrix. For each of the models.

The kSVM with a true positive value of 1243 gave the highest number of true positives, representing about 80% of the data. It is quite closely followed by LR with 1230 representing about 76% of the data. Although their true negative values were quite low when compared to others, they are seen to have outperformed the other algorithms with kSVM giving the best prediction which is meticulously followed by LR. We have noted above the possible reason(s) for the high performances of the kSVM and LR.

TABLE IV. Confusion Matrix Table

Model name	True Positive (TP)	False Positive (FP)	False Negative (FN)	True Negative (TN)
Decision Tree	990	272	251	123
KNN	1142	120	278	96
Kernel SVM	1243	19	251	46
Logistic Regression	1230	32	315	59
Naïve Bayes'	1161	101	270	104
Random Forest	1132	130	270	130

In this section too, we tend to give a broader view of the models used using SHAP (SHapley Additive exPlanations) [72] value is an emergence from the concept of Shapeley and it works at increasing the transparency of models. The Shapely value tends to construct an additive explanation model considered as "contributors." [47] the predictor or model generates a value of prediction for every sample while the SHAP value is described as the value given to every feature in the sample. In any case, where the SHAP value indicates negative, it is a pointer that the feature influences the prediction by lowering the predicted value; but in the case where it is positive, it shows that the influence of that feature will bring an increase to the predicted value in the prediction. When the baseline or the expected value which is the mean of the output of the model over the training data is estimated, the predicted value becomes the sum of this obtained base line and the contribution value of every feature available in the sample.

Using the SHAP, the random forest predictor was plotted in Fig. 6. From the SHAP value plot, we decipher the positive as well as the negative relationships that exist between the predictor and the target variables. From the figure, the target variables are on the y-axis and the impact on the x-axis shows that the building dimension is the variable with the highest impact on the predictor. Other information that can be drawn from the above plot are

- Feature importance which is placed in the descending order of importance in the figure, the least important variable feature with respect to the predictor is the building painted.
- The impact is displayed by the horizontal location. This is correlative with the feature importance and the location signifies whether the effect produced by the value has the association with the prediction in a higher or lower form.
- Original value depicted by the colors on the plot. The variables with high values are shown in red and the ones with low values are shown in blue for this observation.
- Correlation: from the plot, it is observed that the building dimension has a level of correlation rated as high and has both positive and negative impact on the predictor as shown by the red and blue colors. But whereas the building dimension is leading in the positive impact, the insured period is taking the lead on the negative impact. This will make us agree that the building dimension has a positive correlation with the target variable while the insured period has a negative correlation with the same. For random forest model.

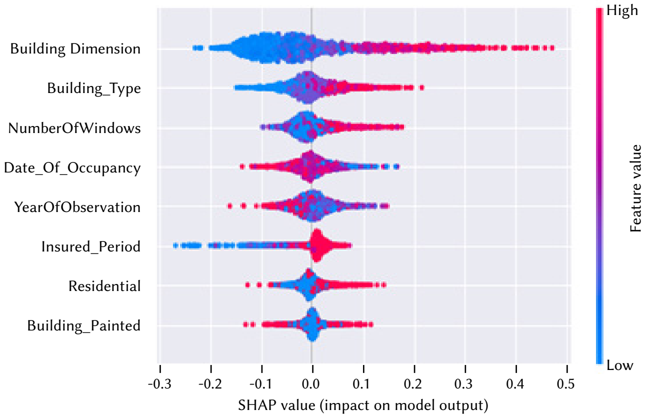


Fig. 4. SHAP value plot for Random Forest.

From Fig. 4 it is observable intuitively that the building dimension is considered to have the most important impact on the model output which is followed by the building type. Although the building type has the highest positive impact and second only to insured_period in the negative impact. The ones with no or negligible impacts were not covered in the plot. From the foregoing, it is easy to agree that the type of building will go a long way to determine whether a client will file for a claim or not in case of unforeseen contingencies when the house is under insurance.

To estimate how important each feature is to the predictor or model as shown in Fig 5 the SHAP estimates the mean absolute value got from the SHAP values from every feature presenting it in a form that the horizontal or y-axis stands for the feature importance while the row or x-axis stands for a feature. The plot in Fig 5. shows variable importance giving a list of the variables in the descending order of their importance to the predictor or model. The variables located at the top have more contribution to the model than those located at the lower part of the plot. From our context, the building dimension has the greatest impact on the predictor while the building painted has the least impact. Those with the greatest impact and as such have a higher power of predictivity. The plot was obtained from the random forest algorithm.

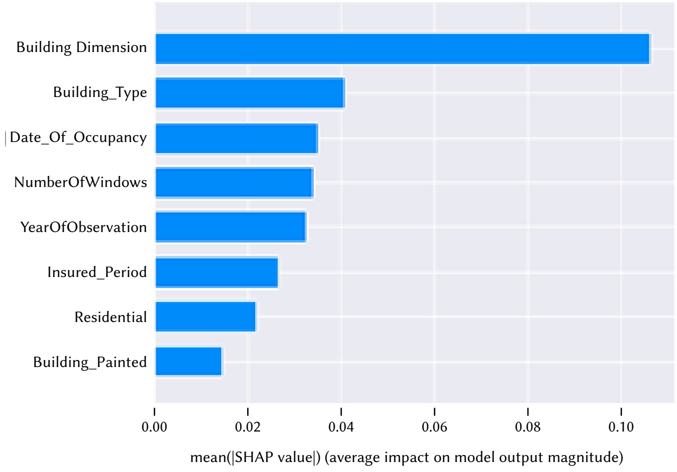


Fig. 5. SHAP value plot for variable importance.

Individual SHAP value plot was made as shown in Fig. 6 to show local interpretability which helps to enhance the transparency of the model – Random forest. When some of the variables were taken at random from the table of values of the variables and plotted using SHAP, the plot obtained is shown in fig. 6.

Any of the rows chosen at random is plotted and the output value is the prediction obtained from that particular row. The base value represents the average of the output of the model over the training data. The different colors (Red/Blue) represents features that influence the prediction to the positive side that is giving a higher prediction (to the right) and influencing the prediction to the negative side giving a lower prediction on the model. The red color with the arrow pointing to the right shows a higher prediction while the blue color with the arrow that points to the left shows a lower prediction.

Other models that we used for the prediction gave a similar result when analyzed used SHAP value showing that the building dimension contributed more to the predictability of the model followed by the building type. It follows the same trend in the variable importance.

Limitations of the models we compared are that their activities are only constrained within the dataset they were trained on like most machine learning algorithms. And the strength is that it gives an opportunity to people who may be confused on which model to employ to know the model that may be able to perform better in a particular scenario. Another thing we can consider as the study's strength is that it was employed in a scenario of a dataset that is considered small with lots of abnormalities. This shows that even in cases of a relatively small dataset with abnormalities, the dataset can be preprocessed and be used for prediction.

VI. Conclusion

This study was carried out on different machine learning algorithms on the insurance dataset from Zindi.africa [68] to get a prediction of whether a customer will have claims or apply for claims over his/her property or not based on the attributes of the building as explained earlier. We preprocessed the data, carried out all the necessary data engineering, implemented the algorithm on python, and obtained the results as shown in the previous section. Some of the algorithms have special qualities and situations for optimum performance, NB for instance is to be chosen in real-time prediction situations and where there are multi classes to be predicted. The DT has special abilities in handling noisy data as well as being able to avoid overfitting when pruning is applied in DT, apart from it being easy to implement and interpret. SVM has shown to be appropriate in a situation where the data available is really small for any reason. The results were analyzed in light of the GINI index and the SHAP values.

The result from the GINI index showed that the Kernel Support Vector Machine outperformed the other algorithms and its performance was followed by that of Logistic regression. Although the results from the algorithms are closely related perhaps because of the amount of data provided but in doubt of the result got from one, one can use the next one with higher with better accuracy, thereby reducing the human effort and time to do the task manually. When analyzed using the SHAP values, it was discovered that the feature with the highest influence on the predictors was the building dimension. This was shown in Fig 4. Consequently, the plots from the SHAP analysis showed that for each row in the distribution, the

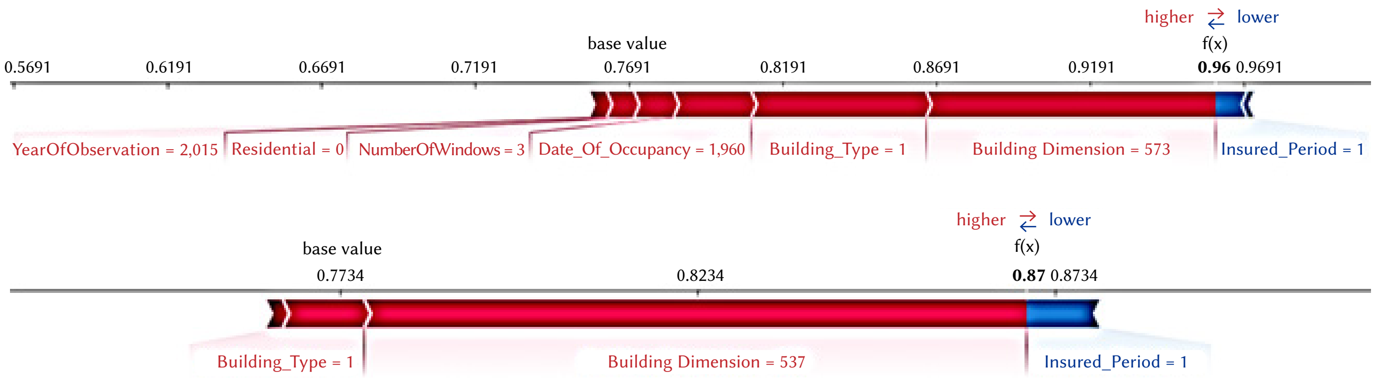


Fig. 6. Individual SHAP value plot for Random Forest.

building dimension has a great deal of influence t the prediction followed by the building type.

Some other algorithms have been theorized to have shown good performance on certain conditions of the dataset for example large dataset, noisy data, etc, it is, therefore, important to put into consideration the amount of dataset available and some other things surrounding it before one concludes on the classification algorithm to use.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No.62072074, No.62076054, No.62027827, No.61902054, No.62002047), the Frontier Science and Technology Innovation Projects of National Key R&D Program (No.2019QY1405), the Sichuan Science and Technology Innovation Platform and Talent Plan (No.2020JDJQ0020) and the Sichuan Science and Technology Support Plan (No. 2020YFSY0010).

References

[1] H. Sufriyana, Y. W. Wu, and E. C. Y. Su, “Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia,” *EBioMedicine*, vol. 54, 2020, doi: 10.1016/j.ebiom.2020.102710.

[2] Y. Huang and S. Meng, “A Bayesian nonparametric model and its application in insurance loss prediction,” *Insurance: Mathematics and Economics*, vol. 93, pp. 84–94, 2020, doi: 10.1016/j.insmatheco.2020.04.010.

[3] P. Li, S. Li, T. Bi, and Y. Liu, “Telecom customer churn prediction method based on cluster stratified sampling logistic regression,” *IET Conference Publications*, vol. 2014, no. CP660, pp. 282–287, 2014, doi: 10.1049/CP.2014.1576.

[4] Z. Kai-Hui, L. Lei, and L. Peng, “Customer churn prediction based on cluster stratified sampling logistic regression,” *International Journal of Digital Content Technology and its Applications*, 2011, doi: 10.4156/jdcta.vol5.issue10.45.

[5] L. Tao, D. Zhu, L. Yan, and P. Zhang, “The traffic accident hotspot prediction: Based on the logistic regression method,” *ICTIS 2015 - 3rd International Conference on Transportation Information and Safety, Proceedings*, pp. 107–110, Aug. 2015, doi: 10.1109/ICTIS.2015.7232194.

[6] H. Lan and Y. Pan, “A crowdsourcing quality prediction model based on random forests,” *Proceedings - 18th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2019*, pp. 315–319, Jun. 2019, doi: 10.1109/ICIS46139.2019.8940306.

[7] X. Ye, X. Wu, and Y. Guo, “Real-time Quality Prediction of Casting Billet Based on Random Forest Algorithm,” *Proceedings of the 2018 IEEE International Conference on Progress in Informatics and Computing, PIC 2018*, pp. 140–143, Jul. 2018, doi: 10.1109/PIC.2018.8706306.

[8] Y. Liu and H. Wu, “Prediction of road traffic congestion based on random forest,” *Proceedings - 2017 10th International Symposium on Computational Intelligence and Design, ISCID 2017*, vol. 2, pp. 361–364, Feb. 2018, doi: 10.1109/ISCID.2017.216.

[9] J. Guo, H. Liu, Y. Luan, and Y. Wu, “Application of birth defect prediction model based on c5.0 decision tree algorithm,” *Proceedings - IEEE 2018 International Congress on Cybermatics: 2018 IEEE Conferences on Internet of Things, Green Computing and Communications, Cyber, Physical and Social Computing, Smart Data, Blockchain, Computer and Information Technology, iThings/Gree*, 2018, doi: 10.1109/Cybermatics_2018.2018.00310.

[10] X. Hu, Y. Yang, L. Chen, and S. Zhu, “Research on a Customer Churn Combination Prediction Model Based on Decision Tree and Neural Network,” *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2020*, pp. 129–132, 2020, doi: 10.1109/ICCCBDA49378.2020.9095611.

[11] R. K. Gupta, S. S. Lathwal, A. P. Ruhil, T. K. Mohanty, and Y. Singh, “Lameness prediction in Karan fries cross-bred cows using decision tree models,” *2015 International Conference on Computing for Sustainable Global Development, INDIACom 2015*, 2015.

[12] I. A. A. Amra and A. Y. A. Maghari, “Students performance prediction using KNN and Naïve Bayesian,” *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, 2017, doi: 10.1109/ICITECH.2017.8079967.

[13] G. A. Bhatt and P. R. Gandhi, “Statistical and ANN based prediction of wind power with uncertainty,” *Proceedings of the International Conference on Trends in Electronics and Informatics, ICOEI 2019*, vol. 2019-April, no. Icoei, pp. 622–627, 2019, doi: 10.1109/icoei.2019.8862551.

[14] A. Yusof and S. Ismail, “Multiple Regressions in Analysing House Price Variations,” *Communications of the IBIMA*, 2012, doi: 10.5171/2012.383101.

[15] M. A. Babyak, “What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models,” *Psychosomatic Medicine*, 2004, doi: 10.1097/00006842-200405000-00021.

[16] Mislan, Haviluddin, S. Hardwinarto, Sumaryono, and M. Aipassa, “Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan - Indonesia,” *Procedia Computer Science*, 2015, doi: 10.1016/j.procs.2015.07.528.

[17] A. Chogle, P. Khaire, A. Gaud, and J. Jain, “House Price Forecasting using Data Mining Techniques,” *House Price Forecasting using Data Mining Techniques*, 2017, doi: 10.17148/IJARCCE.2017.61216.

[18] D. Sangani, K. Erickson, and M. Al Hasan, “Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting,” *Proceedings - 14th IEEE International Conference on Mobile Ad Hoc and Sensor Systems, MASS 2017*, 2017, doi: 10.1109/MASS.2017.88.

[19] A. Nur, R. Ema, H. Taufiq, and W. Firdaus, “Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study : Malang, East Java, Indonesia,” *International Journal of Advanced Computer Science and Applications*, 2017, doi: 10.14569/ijacsa.2017.081042.

[20] A. Khalafallah, “Neural Network Based Model for Predicting Housing Market Performance,” *Tsinghua Science and Technology*, 2008, doi: 10.1016/S1007-0214(08)70169-X.

[21] N. Bhagat, A. Mohokar, and S. Mane, “House Price Forecasting using Data Mining,” *International Journal of Computer Applications*, 2016, doi: 10.5120/ijca2016911775.

[22] S. C. Bourassa, E. Cantoni, and M. Hoesli, “Spatial dependence, housing submarkets, and house price prediction,” *Journal of Real Estate Finance and Economics*, 2007, doi: 10.1007/s11146-007-9036-8.

[23] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, “Geographically weighted regression: a method for exploring spatial nonstationarity,” *Geographical Analysis*, 1996, doi: 10.1111/j.1538-4632.1996.tb00936.x.

[24] A. Onan, “Classifier and feature set ensembles for web page classification,” *Journal of Information Science*, 2016, doi: 10.1177/0165551515591724.

[25] A. Onan and S. KorukoGlu, “A feature selection model based on genetic rank aggregation for text sentiment classification,” *Journal of Information Science*, 2017, doi: 10.1177/0165551515613226.

[26] A. Onan, “An ensemble scheme based on language function analysis and feature engineering for text genre classification,” *Journal of Information Science*, 2018, doi: 10.1177/0165551516677911.

[27] A. Onan, S. Korukoğlu, and H. Bulut, “Ensemble of keyword extraction methods and classifiers in text classification,” *Expert Systems with Applications*, 2016, doi: 10.1016/j.eswa.2016.03.045.

[28] A. Sharaff and S. R. Roy, “Comparative Analysis of Temperature Prediction Using Regression Methods and Back Propagation Neural Network,” *Proceedings of the 2nd International Conference on Trends in Electronics and Informatics, ICOEI 2018*, no. Icoei, pp. 739–742, 2018, doi: 10.1109/ICOEI.2018.8553803.

[29] K. Abhishek, M. P. Singh, S. Ghosh, and A. Anand, “Weather Forecasting Model using Artificial Neural Network,” *Procedia Technology*, 2012, doi: 10.1016/j.protcy.2012.05.047.

[30] N. Shobha and T. Asha, “Monitoring weather based meteorological data: Clustering approach for analysis,” *Proceedings-IEEE International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2017* -, 2017, doi: 10.1109/ICIMIA.2017.7975575.

[31] J. Gill, B. Singh, and S. Singh, “Training back propagation neural networks with genetic algorithm for weather forecasting,” *SIISY 2010 - 8th IEEE International Symposium on Intelligent Systems and Informatics*, 2010, doi: 10.1109/SISY.2010.5647319.

[32] A. Paniagua-Tineo, S. Salcedo-Sanz, C. Casanova-Mateo, E. G. Ortiz-García, M. A. Cony, and E. Hernández-Martín, “Prediction of daily maximum temperature using a support vector regression algorithm,”

Renewable Energy, 2011, doi: 10.1016/j.renene.2011.03.030.

[33] R. E. Abdel-Aal, "Hourly temperature forecasting using abductive networks," *Engineering Applications of Artificial Intelligence*, 2004, doi: 10.1016/j.engappai.2004.04.002.

[34] L. Houthuys, Z. Karevan, and J. A. K. Suykens, "Multi-view LS-SVM regression for black-box temperature prediction in weather forecasting," *Proceedings of the International Joint Conference on Neural Networks*, 2017, doi: 10.1109/IJCNN.2017.7965975.

[35] H. Xie, M. Ding, L. Chen, J. An, Z. Chen, and M. Wu, "Short-term wind power prediction by using empirical mode decomposition based GA-SYR," *Chinese Control Conference, CCC*, 2017, doi: 10.23919/ChiCC.2017.8028818.

[36] X. Peng, D. Deng, J. Wen, L. Xiong, S. Feng, and B. Wang, "A very short term wind power forecasting approach based on numerical weather prediction and error correction method," *China International Conference on Electricity Distribution, CICED*, 2016, doi: 10.1109/CICED.2016.7576362.

[37] W. Zhang, H. Zhang, J. Liu, K. Li, D. Yang, and H. Tian, "Weather prediction with multiclass support vector machines in the fault detection of photovoltaic system," *IEEE/CAA Journal of Automatica Sinica*, 2017, doi: 10.1109/JAS.2017.7510562.

[38] S. Papantoniou and D. D. Kolokotsa, "Prediction of outdoor air temperature using neural networks: Application in 4 European cities," *Energy and Buildings*, 2016, doi: 10.1016/j.enbuild.2015.06.054.

[39] D. C. Wu, B. Bahrami Asl, A. Razban, and J. Chen, "Air compressor load forecasting using artificial neural network," *Expert Systems with Applications*, no. October, p. 114209, 2020, doi: 10.1016/j.eswa.2020.114209.

[40] P. Patil, N. Yaligar, and S. Meena, "Comparision of Performance of Classifiers - SVM, RF and ANN in Potato Blight Disease Detection Using Leaf Images," *2017 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2017*, pp. 3–7, 2018, doi: 10.1109/ICCIC.2017.8524301.

[41] X. Tang, Z. Liu, T. Li, W. Wu, and Z. Wei, "The application of decision tree in the prediction of winning team," *Proceedings - 2018 International Conference on Virtual Reality and Intelligent Systems, ICVRIS 2018*, 2018, doi: 10.1109/ICVRIS.2018.00065.

[42] N. I. Nwulu, "A decision trees approach to oil price prediction," *IDAP 2017 - International Artificial Intelligence and Data Processing Symposium*, pp. 0–4, 2017, doi: 10.1109/IDAP.2017.8090313.

[43] M. A. Hassonah, A. Rodan, A. K. Al-Tamimi, and J. Alsakran, "Churn Prediction: A Comparative Study Using KNN and Decision Trees," *ITT 2019 - Information Technology Trends: Emerging Technologies Blockchain and IoT*, 2019, doi: 10.1109/ITT48889.2019.9075077.

[44] R. S. Raj, D. S. Sanjay, M. Kusuma, and S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," *1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering, ICATIECE 2019*, pp. 41–45, 2019, doi: 10.1109/ICATIECE45860.2019.9063792.

[45] R. Bayindir, M. Yesilbudak, M. Colak, and N. Genc, "A novel application of naive bayes classifier in photovoltaic energy prediction," *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, vol. 2017-Decem, pp. 523–527, 2017, doi: 10.1109/ICMLA.2017.0-108.

[46] A. P. Salim, K. A. Laksitowening, and I. Asror, "Time Series Prediction on College Graduation Using KNN Algorithm," *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166238.

[47] B. Zhang and D. Ma, "Flight delay prediciton at an airport using maching learning," *Proceedings - 2020 5th International Conference on Electromechanical Control Technology and Transportation, ICECTT 2020*, 2020, doi: 10.1109/ICECTT50890.2020.00128.

[48] H. Khaksar and A. Sheikholeslami, "Airline delay prediction by machine learning algorithms," *Scientia Iranica*, 2019, doi: 10.24200/sci.2017.20020.

[49] L. Belcastro, F. Marozzo, D. Talia, and P. Trunfio, "Using scalable data mining for predicting flight delays," *ACM Transactions on Intelligent Systems and Technology*, 2016, doi: 10.1145/2888402.

[50] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, 2016, doi: 10.1109/DASC.2016.7777956.

[51] M. S. Acharya, A. Armaan, and A. S. Antony, "A comparison of regression models for prediction of graduate admissions," *ICCIDS 2019 - 2nd International Conference on Computational Intelligence in Data Science, Proceedings*, 2019, doi: 10.1109/ICCIDS.2019.8862140.

[52] C. J. Ejiyi, O. Bamisile, N. Ugochi, Q. Zhen, N. Ilakoze, and C. Ijeoma, "Systematic Advancement of Yolo Object Detector For Real-Time Detection of Objects," *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 279–284, Dec. 2021, doi: 10.1109/ICCWAMTIP53232.2021.9674163.

[53] C. J. Ejiyi, J. Deng, T. U. Ejiyi, A. A. Salako, M. B. Ejiyi, and C. G. Anomihe, "Design and Development of Android Application for Educational Institutes," *Journal of Physics: Conference Series*, 2021, doi: 10.1088/1742-6596/1769/1/012066.

[54] R. D. Cook and S. Weisberg, "Criticism and Influence Analysis in Regression," *Sociological Methodology*, 1982, doi: 10.2307/270724.

[55] S. O. Bamisile Olusola, Ariyo Oluwasanmi, Chukwuebuka Joseph Ejiyi, Nasser Yimen, "Comparison of machine learning and deep learning algorithms for hourly global / diffuse solar radiation predictions," *International Journal of Energy Research*, no. January, pp. 1–22, 2021, doi: 10.1002/er.6529.

[56] K. Methaprayoon, C. Yingvivatanapong, W. J. Lee, and J. R. Liao, "An integration of ANN wind power estimation into unit commitment considering the forecasting uncertainty," *IEEE Transactions on Industry Applications*, 2007, doi: 10.1109/TIA.2007.908203.

[57] M. A. F. Azlah, L. S. Chua, F. R. Rahmad, F. I. Abdullah, and S. R. W. Alwi, "Review on techniques for plant leaf classification and recognition," *Computers*. 2019, doi: 10.3390/computers8040077.

[58] A. Ramil, A. J. López, J. S. Pozo-Antonio, and T. Rivas, "A computer vision system for identification of granite-forming minerals based on RGB data and artificial neural networks," *Measurement: Journal of the International Measurement Confederation*, 2018, doi: 10.1016/j.measurement.2017.12.006.

[59] S. Sperandei, "Understanding logistic regression analysis," *Biochemia Medica*, 2014, doi: 10.11613/BM.2014.003.

[60] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Archives of Psychiatry*, 2015, doi: 10.11919/j.issn.1002-0829.215044.

[61] L. Song, "Research on the application of data mining algorithm based on decision tree," *Metallurgical and Mining Industry*, 2015.

[62] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, "Predicting reaction performance in C–N cross-coupling using machine learning," *Science*, 2018, doi: 10.1126/science.aar5169.

[63] Y. L. Pavlov, "Random forests," *De Gruyter*, 2019, doi: https://doi.org/10.1515/9783110941975.

[64] M. Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms: Second Edition," *John Wiley & Sons, Inc., Hoboken, New Jersey*, 2011, doi: 10.1002/9781118029145.

[65] K. L. Priya, M. S. Charan Reddy Kypa, M. M. Sudhan Reddy, and G. R. Mohan Reddy, "A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier," *Proceedings of the 4th International Conference on Trends in Electronics and Informatics, ICOEI 2020*, no. Icoei, pp. 603–607, 2020, doi: 10.1109/ICOEI48184.2020.9142959.

[66] Mahima and N. B. Padmavathi, "Comparative study of kernel SVM and ANN classifiers for brain neoplasm classification," *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT 2017*, 2018, doi: 10.1109/ICICICT1.2017.8342608.

[67] Y. Zhang and L. Wu, "An MR brain images classifier via principal component analysis and kernel support vector machine," *Progress in Electromagnetics Research*, 2012, doi: 10.2528/PIER12061410.

[68] "Competitions - Zindi." https://zindi.africa/competitions (accessed Jul. 17, 2020).

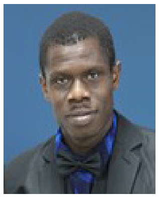
[69] S. Sharma and A. Bhagat, "Data preprocessing algorithm for Web Structure Mining," *Proceedings on 5th International Conference on Eco-Friendly Computing and Communication Systems, ICECCS 2016*, 2017, doi: 10.1109/Eco-friendly.2016.7893249.

[70] S. Samsani, "An RST based efficient preprocessing technique for handling inconsistent data," *2016 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2016*, 2017, doi: 10.1109/ICCIC.2016.7919591.

[71] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," *3rd Edition Morgan Kaufmann Publishers, Waltham.*, 2012, doi: 10.1016/

C2009-0-61819-5.

[72] S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, 2017.



Chukwuebuka Joseph Ejiyi

Chukwuebuka Joseph Ejiyi received his Bachelor’s Degree in 2014 from the Federal University of Technology Owerri (FUTO) Nigeria. He went on to obtain a master’s degree in Software Engineering at the University of Electronic Science and Technology of China (UESTC) in 2021. . He is currently pursuing a Ph.D. degree with the Schoool of Information and Software Engineering at UESTC Chengdu China. His research interest is in Artificial intelligence, Deep Learning and he is currently working on Object detection using a single-stage neural network as well as Object classification. He also has a strong interest in image analysis especially with regards to medical images.

Zhen Qin

Zhen Qin is currently a professor in the School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC). He received his Ph.D. degree from UESTC in 2012. He was a visiting scholar in the Department of Electrical Engineering and Computer Science at Northwestern University. His research interests include data fusion analysis, mobile social networks, wireless sensor networks, and image processing.

Abdulhaq Adetunji Salako

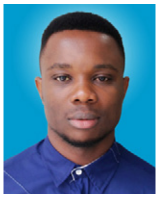
Abdulhaq Adetunji Salako received a bachelor’s degree in information technology from the Valley View University (VVU), Accra- Ghana, West Africa, in 2017. He is currently pursuing an M.Sc. degree in computer science and engineering with the University of Electronic Science and Technology of China (UESTC). From 2015 to 2019, he worked under sub-contracts for Wigal Solutions, Ghana, and Teachers Fund of GNAT, Ghana. He is a member of the Data Visualization Research Team - UESTC. His research interests include information visualization, visual analytics, data mining, explainable AI, and NLP.

Happy Nkanta Monday

Happy Nkanta Monday received the B.Tech. degree in agricultural and environmental engineering from the Federal University of Technology, Akure, Nigeria, in 2013 and the M.Eng. degree in electronic science and technology from the University of Electronic Science and Technology of China, in 2018. He is currently pursuing a Ph.D. degree with the school of computer science and engineering, University of Electronic Science and Technology of China. His research interests include computer vision, wavelet, super-resolution, and deep learning.

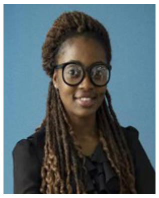
Grace Ugochi Nneji

Grace Ugochi Nneji received the B.Tech. degree in computer science from the Federal University of Technology, Owerri, Nigeria, in 2014 and the M.Eng. degree in software engineering from the University of Electronic Science and Technology of China, in 2019. She is currently pursuing a Ph.D. degree with the school of software engineering, University of Electronic Science and Technology of China. Her research interests include computer vision, re-identification, super-resolution, and deep learning.



Chiagoziem C. Ukwuoma

Chiagoziem C. Ukwuoma received the B.Eng. degree (Mechanical Engineering-Automobile Technology) from the Federal University of Technology Owerri in 2014 and his MSc. degree (Software Engineering) from the University of Electronic Science and Technology of China (UESTC) in 2020. He is currently a Ph.D. student at the University of Electronic Science and Technology of China (UESTC). His research interests include Object Detection and Object Classification.



Ijeoma Amuche Chikwendu

Ijeoma Amuche Chikwendu Received a B.Sc degree in Information management technology at the Federal University of Technology Owerri in 2014 and a Masters degree in Information and Communication Engineering at the University of Electronic Science and Technology of China (UESTC) in 2021. She is currently pursuing a Ph.D. degree at the same University where she obtained her master’s degree. Her research interest is Statistical signal processing, Distributed estimation, target tracking, and localization.

Ji Gen

Ji Gen is currently a professor in the School of Information and Software Engineering, University of Electronic Science and Technology of China (UESTC). He received his Ph.D. degree from UESTC in 2012. His research interests include system software, information processing. He is a communication evaluation expert of information Department of National Natural Science Foundation of China.