Technology in Society 63 (2020) 101413

Contents lists available at ScienceDirect

Technology in Society

journal homepage: http://www.elsevier.com/locate/techsoc

A study on credit scoring modeling with different feature selection and machine learning approaches

Shrawan Kumar Trivedi

Department of Management Studies, Indian Institute of Technology (ISM), Dhanbad, India

ARTICLE INFO

Keywords:
Credit scoring
Machine learning classifier
Random forest
C5.0
SVM
Naïve bayes
Bayesian
Feature selection technique
Information-gain
Gain-ratio
Chi-square

ABSTRACT

A bit hurdle for financial institutions is to decide potential candidates to give a line of credit identifying the right people without any credit risk. For such a crucial decision, past demographic and financial data of debtors is important to build an automated artificial intelligence credit score prediction model based on machine learning classifier. In addition, for building robust and accurate machine learning models, important input predictors (debtor's information) must be selected. The present computational work focuses on building a credit scoring prediction model. A publicly available German credit data is incorporated in this study. An improvement in the credit scoring prediction has been shown with the use of different feature selection techniques (such as Information-gain, Gain-Ratio and Chi-Square) and machine learning classifiers (Bayesian, Naïve Bayes, Random Forest, Decision Tree (C5.0) and SVM (support Vector Machine)). Further, a comparative analysis is performed between different machine learning classifiers and between different feature selection techniques. Different evaluation metrics are considered for analyzing performance of the models (such as accuracy, F-measure, false positive rate, false negative rate and training time). After analysis, a best combination of machine learning classifier and feature selection technique are identified. In this study, a combination of random forest (RF) and Chi-Square (CS) is found good, among other combinations, with respect to good performance accuracy, F-measure and low false positive and false negative rates. However, training time for this particular combination was found to be slightly higher. Result of C5.0 with chi-square was comparable with the best one. This study provides an opportunity to financial institutions to build an automated model for better credit scoring.

1. Introduction

Lending money is a traditional process. It increases potential of the debtor but the creditor lends money based on several factors. Lenders can get information about the individual's behaviour on repayment of money through the design of a credit scoring process. This process was developed on 1960 where percentage risk involved in lending money is calculated that is, likelihood of the customers to repay the money to lenders within a particular period of time. Credit Scores are given to individuals before lending money, to observe the risk involved.

In the retail credit sector, credit scoring is used for developing empirical models to support decisions [1]. This sector has substantial economic importance. For example, in 2013, in the US, consumer loans worth $1,132bn were held by banks, compared to $1,541bn in corporate business. In 2012, individuals' loans and mortgages were higher than those of corporates in UK. These numbers indicate that financial institutions require quantitative tools to take credit decisions. Credit scoring models capture probability of the unpaid behaviour of borrowers in future. Lenders use predictive models called scorecards in application scoring, to evaluate the probability of an applicant being in default.

Classification algorithms are developed using probability of default [2]. The study of Baesens et al., [3]; is the most used and benchmarking classifier till date. It is seen in literature that only a few studies discuss more recent advancement of predictive learning. The important trend currently in machine learning (e.g. Ref. [4], is the development of different algorithms that are then improved weighing wise, using exploratory search and developing selective multiple classifier systems. Only a few attempts are made to verify the potential of systematic credit scoring. Recent advancements in the field focus on three areas namely, i) developing of scorecards using novel classification (examples of some such classifications are: extreme learning machines, rotation forest etc.); ii) the scorecards are observed by measuring novel performance (e.g., H-measure or partial Gini coefficient); and, iii) checking and comparing performance of the scorecards with the help of hypothesis tests [5]. These developments generate small interest in analysis of prediction

E-mail addresses: shrawan@iitism.ac.in, f10shrawank@iimidr.ac.in.

https://doi.org/10.1016/j.techsoc.2020.101413
Received 9 October 2017; Received in revised form 7 September 2020; Accepted 17 September 2020
Available online 28 September 2020

modeling, with some limitations of previous studies which can be listed as: i) cannot be used for small datasets, ii) different state-of-the-art classifiers are not comparable, and iii) can use only small sets of indicators leading to less accuracy. The research gaps mentioned above need attention [3]. Therefore, progress in terms of development, application of the same, can be used in credit industry for preparing predictive decision support models that can evaluate and help in uncovering the expansion in terms of additional literature and improvements in the field. (For reviews: [1]; Anil Kumar & Ravi, 2007). Credit worthiness of customers is established on exploratory variables, and can be estimated by the use of support models. Data about a particular customer can be extracted from past transactions. Additionally, application forms filled at the start of the process can give a lot of information about the customer and her/his demographics, which would help in building up Retail models. Moreover, when it comes to data collection of a corporate, financial statements such as balance sheets, ratios, indicators of macro-economics, are used in building up a corporate risk model. Certain challenges are found when credit scoring is done for individual consumers as compared to corporate because in the latter, different variables are used. Therefore, studies mostly focus on retail and corporate business.

In consumer credit risk modeling, a variety of prediction tasks occur. As per Basel II Capital accord, it is necessary for banks and other financial institutions to estimate: i) probability of default (PD); ii) exposure at default (EAD); and, iii) loss given default (LGD). The recent trending models for research topic are EAD and LGD models [6,7].

Machine learning models are considered important tools for building predictive models. Several researches in literature discuss credit scoring models that use machine learning classifiers. However, building an optimum credit score prediction model is a potential area of research. To build a robust, accurate and sensitive machine learning prediction model, the information of input predictors is important. Feature selection are methods to evaluate the informative features and reduction of dimension of data. In literature, many feature selection techniques are tested that showed improvement in credit score prediction.

[8]; did research on credit scoring and tested machine learning models using four feature selection techniques i.e., genetic algorithm, information gain ratio, relief F, and principal component analysis. After tests on various parameter settings, they arrived at principal component analysis as the best feature selection technique. Another research done by Ref. [9]; develops credit scoring prediction model using four feature selection methods i.e., correlation matrix, Classification and Regression Tree (CART), and Principle Component Analysis (PCA). PCA is an identified good feature selection technique in this research. [10]; incorporated feature selection techniques with deep learning to build a credit scoring model. [11]; experimented wrapper feature selection method with the use of genetic algorithms, multiple population genetic algorithms (MPGA) and hybrid multiple population genetic algorithms (HMPGA). This research concluded that HMPGA is better than GA and MPGA feature selection methods. The research done by Ref. [12]; used information gain directed feature selection method to build a credit scoring prediction model. [13]; used genetic algorithm (GA) based feature selection method and found it efficient for machine learning in building credit score prediction model. [14]; proposed a hybrid variable neighborhood search and estimation of distribution technique with the elitist population strategy for selecting important features to build credit scoring prediction model. Another research done by Ref. [15]; constructed a hybrid data mining prediction model using feature selection and machine learning models. In this research, five feature selection (classifier feature selection, correlation feature selection, gain ratio, information gain and relief F) are combined to get the informative predictor to build effective machine learning credit scoring model.

On the other hand, various machine learning models have been tested in research or credit score prediction. A research done by Ref. [16]; incorporated support vector machine (SVM) and Artificial Neural Network (ANN), for building credit scoring models. They concluded that ANN outperforms SVM with significant high accuracy. Another research done by Ref. [17]; used artificial neural network (ANN) to build extreme learning machine (ELM) credit score predication model. Weidong, 2019, constructed credit scoring prediction model using random forest (RF) and XGboost machine learning models. In the study of [18]; random forest (RF) and statistical methods (logit and ordinary least square (OLS)) were experimented. This research suggests that random forest is a better performer than the logit and OLS technique. Another research done by Ref. [19]; did a brief literature survey of machine learning models on credit scoring application and also touched the problem of imbalance of the data in review. A study done by Ref. [20]; tested fuzzy logic for credit scoring and found it accurate in uncertainty modeling.

From the above discussion, it is found that several feature selection and machine learning models have been tested on credit scoring. An in-depth analysis is required to get the best predictor of credit scoring and compatible machine learning algorithms. In literature, many feature selection and machine learning models are tested but yet, an optimum credit scoring prediction model remains a potential area of research. The above discussion leads to the following research questions.

RQ1. What are the most informative predictors of the credit scoring models?

RQ2. Which feature selection method is capable to evaluate best credit score predictors?

RQ3. Which is the accurate and robust machine learning classifier for predicting credit scoring?

RQ4. Which combinations of feature selection and machine learning models are best suited in developing a credit scoring prediction model?

To address the above research questions, this study tested three feature selection methods (Chi-Square, Information gain and Gain Ration) and five machine learning models (Bayesian, Naïve Bayes, support vector machine (SVM), Decision Tree (C5.0) and Random Forest (RF)), on publicly available German Credit Scoring data. All the models were evaluated using various metrics such as performance accuracy, f-measure, true positive rate, true negative rate and training time with 66–34% split method and 10-fold cross validation. The framework of this research is depicted in Fig. 1.

2. Testing corpora

German Credit scoring data set,[1] publicly available, is considered for developing credit scoring model. This data consists of 21 Variables (20 input predictors and 1 outcome variable) which describe different characteristics of the respondents. This data contains a total of 1000 instances of people who have taken credit earlier. Target variable is the deflators and not-defaulter people for credit where '1' represents 'they are not-defaulters', with 70% cases, and '0'represents 'defaulter', with 30% cases (Table 1). New applicants for credit are evaluated based on historical data of the debtors in terms of different input predictors. At first, machine learning classifiers with feature selection techniques were trained on 66% of the dataset and 34% data was used for testing the classifiers. Further, 10-fold cross validation was used to evaluate the models.

2.1. Pre-processing of corpora

In Pre-processing, transformation of data into a particular format is done, which is effective and efficient for user of the data. Data pre-processing [21] is one of the most important steps in classification. There are various methods to collect data. The data which is collected may not be accurate all the time due to missing values that give misleading results. Hence, it is necessary to clean the data before analysis. Missing and noisy data is removed in this process [22]. Cleaning,

[1] https://archive.ics.uci.edu/ml/datasets/South+German+Credit.

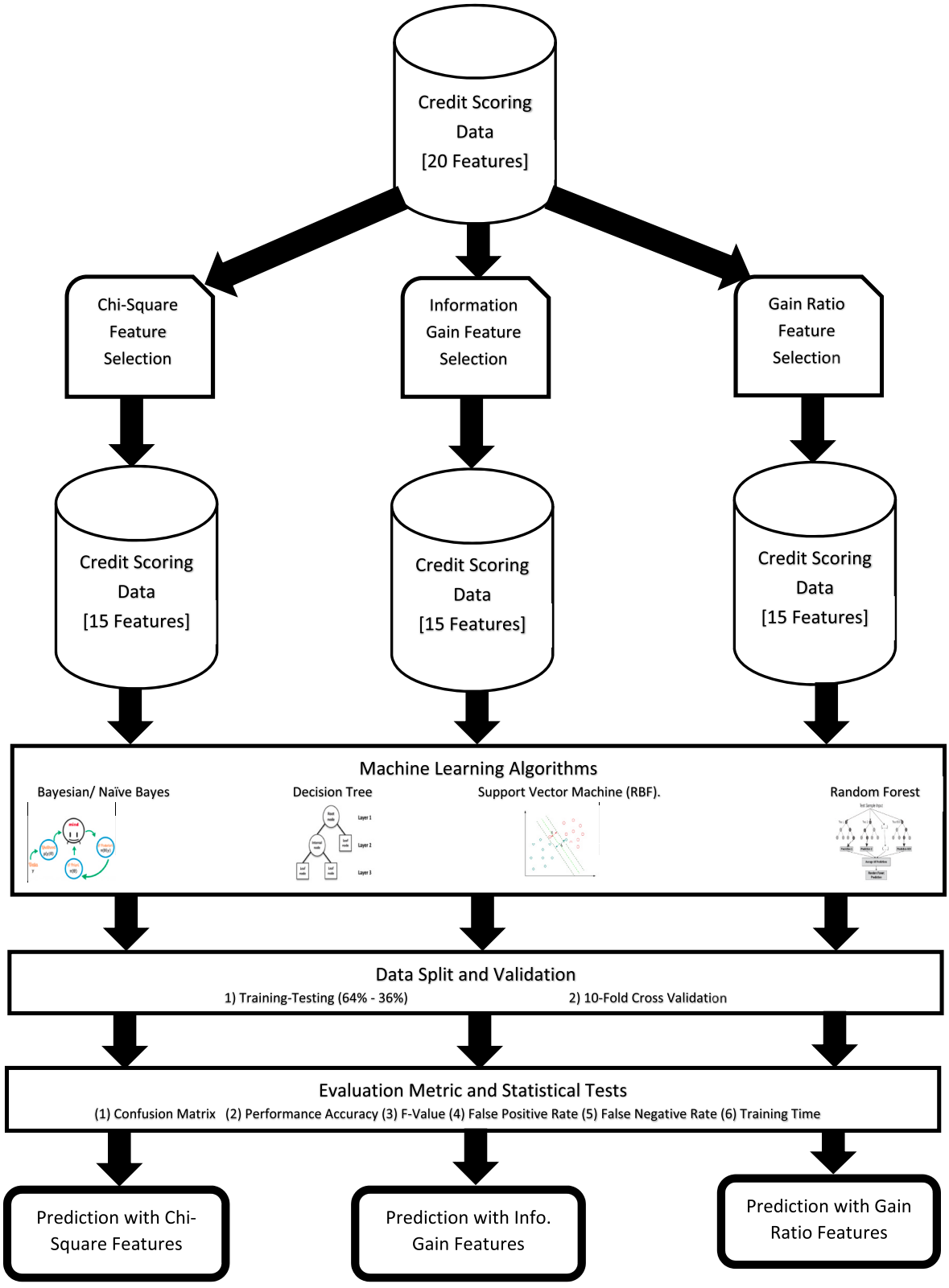

Fig. 1. Credit scoring model framework.

filtering, integration, discretization and normalization of data are performed in pre-processing of dataset. Since the data taken in this study is secondary and publicly available, it is already pre-processed and structured in nature. However, the credit data contains 20 input predictors for classifying defaulter and not-defaulter persons, which can be reduced using feature selection techniques, to get the most informative features.

3. Feature selection techniques

After Pre-processing, feature selection methods are performed. It is not necessary that all the 20 input features are most informative in nature to predict the output classes. To get this most informative features set, feature selection techniques are incorporated. In this section, all the

Table 1
Data description.

S/N	Feature Name	Mean	Standard Deviation
1	Account Balance	2.57	1.25
2	Duration of Credit (month)	20.90	12.05
3	Previous Credit Status	NA	NA
4	Purpose		
5	Credit Amount	3271.24	2821.34
6	Value Savings/Stocks	2.10	1.57
7	Length of current employment (Years)	3.38	1.20
8	Instalment per cent	2.97	1.12
9	Sex & Marital Status	NA	NA
10	Guarantors	NA	NA
11	Duration in Current address (Years)	2.845	1.10
12	Most valuable available asset	NA	NA
13	Age (years)	35.54	11.35
14	Concurrent Credits	2.68	0.71
15	Type of apartment	NA	NA
16	No of Credits at this Bank	1.41	0.58
17	Occupation	NA	NA
18	No of dependents	1.15	0.36
19	Telephone	NA	NA
20	Foreign Worker	NA	NA

three feature selection techniques taken in this research (Such as, Information Gain, Gain Ratio and Chi-Square) have been described.

3.1. Information gain (IG)

This method works to identify informative predictors of the classes, to build an accurate machine learning application [23,24]. It calculates information gain of the input features and provides those features to machine learning for predicting accurate class of the new instance, to take credit decision. These classes are decided by matching the patterns of input predictor from trained machine learning model. The value of information gain is measured by computing the fall in overall entropy after introducing a new feature. Entropy is the expected value of a particular feature required for the classification of an instance. Let us assume x and y are two variables where x is input feature for output y; entropy of y is computed as Equation (1).

$$S(Y) = -\sum_{y \in Y} P'(y)\log_2 P'(y) \tag{1}$$

The change in entropy after introducing input predictor x is given in Equation (2).

$$S\left(\frac{Y}{X}\right) = -\sum_{x \in X}(x)\sum_{y \in Y} P'\left(\frac{y}{x}\right)\log_2 P'\left(\frac{y}{x}\right) \tag{2}$$

Information gain is the entropy difference of predictor y, and entropy of predictor y after introducing input predictor x (Equations (3)–(5)).

$$IG = S(Y) - S\left(\frac{Y}{X}\right) \tag{3}$$

$$IG = S(X) - S\left(\frac{X}{Y}\right) \tag{4}$$

$$IG = S(Y) + S(X) - S(Y,X) \tag{5}$$

Information gain (IG) is proportioned balanced measurement so that value of this for y after observation of x is the same as the value for x after observing y.

3.2. Gain ratio (GRGRGR)

An extended version of Information Gain (IG) is called Gain ratio (GR) [25,26]. Information gain (IG) shows biasness towards choice of feature with higher numerical values, even in the condition where there is less information. This reveals the weakness of IG.

Gain ratio is used to compensate biasness of IG which is a non-symmetrical measure. From Equations (3) and (4), the GR is the ratio of IG and overall entropy of x as given in Equation (6).

$$GR = \frac{IG}{S(X)} \tag{6}$$

From Equation (6), when variable y is predicted, information gain is normalized by dividing the overall entropy of x. Due to normalization, it gives value of GR between 0 and 1. If it is 1, the information in x will predict y, and if it is 0, then x and y will have no relation among themselves. Gain ratio is different from IG because it favors features with lower numerical value.

3.3. Chi-square (χ^2)

Chi square [27,28,29] is used for observing a model's best features. By completely analyzing the chi-squared statistics, it provides appreciated features from the feature space for the particular class. In this method, testing of initial hypothesis is done, with an assumption that the two features are different,

$$\chi^2 = \Sigma_{i-1}^{r}\Sigma_{j-1}^{c}\left(\frac{O^{ij} - E^{ij}}{E^{ij}}\right) \tag{7}$$

where O is termed as frequency that is observed, E is termed as frequency that is expected. Larger value of x implies considerable confirmation against the acceptance of initial hypothesis.

4. Machine learning classifiers

Various machine learning classifiers are famous in the research where technological impact on the society has been seen significantly. Many businesses and societal applications are well tackled by machine learning classifier [30]. A work done by Ref. [31] explored the use of artificial intelligence (AI) and machine learning (ML) for sustainable growth of live streaming music industry. Also the technological automations are employed in the cognitive behaviour of human being [32]. The machine learning application are also popular in health care industry where many disease prediction research incorporates such technologies [33]. E-commerce sector has also well utilized artificial intelligence technology in many business applications [34]. Recently, the use of machine learning has gained attention from financial institutions [35] for making prediction applications such as credit scoring. This study also concentrates to build an effective credit scoring model. This section describes all state-of-art machine learning model to identify a best model for credit scoring.

4.1. Bayesian model

Bayesian model [36] is a probabilistic classifier with the nature of white box. It is used to predict particular class of membership samples. This model works on Bayesian theory and is explained by the following model. Let us consider a training sample set, $D = \{u_1 \ldots\ldots u_n\}$, where mission of the classifier is to evaluate and determine the training sample and its function $f : (x_1 \ldots . x_n) \rightarrow C$. This function helps to decide the label for the sample $x = (x_1 \ldots . x_n)$with respect to the highest probability of the class as per the label $P(c_j / x_1, \ldots . x_n)$. According to minimum error probability criterion:

If $p\left(c_i/x\right) = \max_{j=1,\ldots i} P\left(c_j/x\right)$ then we can determine that $x \in c_i$.

The two commonly used models are Naïve Bayes [37] and Bayesian belief. Naïve Bayesian classifier assumes that independent samples are used. Even if the calculation is simplified in this model, the variables are correlated really. Bayesian model is a graphical model where conditional independencies are characterized between variable subsets. Two sections of Bayesian network are namely, cyclic graph and conditional

probability tables.

4.2. Support vector machine (SVM)

In recent years, SVM is considered the best classifier developed for Pattern Classification [38,39]. It does not limit distribution of data and is mostly used for small samples. This model which is based on structural risk, also achieves good robustness. Let us consider S is a dataset and M with observations set defined as $\{(x_t, y_t) : x_t \in R^N; y_t \in \{-1, +1\}, t = 1, 2 \ldots M\}$ where $\{x_t, z, y_t \in \{-1, +1\}\}$ which denotes equivalent labels of binary class, suggesting whether the client or customer is at default. Main purpose of this categorization is to find a maximal hyper plane by which examples of opposite labels are separated. This constraint is written as:

$$y_i((w, x_t) + b) - 1 \geq 0, \; t = 1, 2 \ldots M \tag{8}$$

where w is defined as the normal of the plane and b is defined as intercept. (w,b) denotes a linear set. $^2/_{\|w\|}$ is margin of separation. The optimal hyper plane is the point where margin $^2/_{\|w\|}$ is maximum. Subject to constraints of $y_t((w, x_t) + b) - 1 \geq 0, \; t = 1, 2 \ldots M$.Then solving the quadratic equation $\min_{w,b} {}^1/_2\|w\|^2$ is the classification problem.

$$y_i((w, x_t) + b) - 1 \geq 0, \; t = 1, 2 \ldots M \tag{9}$$

By bringing in langrage multipliers, $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_m)$ the problem is changed to solve the dual program as follows:

$$\max Q(\alpha) = \sum_{t=1}^{M} \alpha_t - \frac{1}{2} \sum_{s=1}^{n} \alpha_t \alpha_s y_t y_s (x_t, x_s) \tag{10}$$

$$s\,t \sum_{t=1}^{M} y_t \alpha_t = 0, \alpha \geq 0, t = 1, 2 \ldots M \tag{11}$$

If $\alpha > 0$ then x_t is termed as support vector. From above formulation, classification decision function is written as Equations (12) and (13).

$$f(x) = \mathrm{sgn}((w, x) + b) \tag{12}$$

$$\varepsilon^t = sgn\left\{ \sum_{t=1}^{M} \alpha_t y_t (x_t, x) + b \right\} \tag{13}$$

The decision function obtained above defines that examples are classified as class +1 when $(w, x) + b > 0$ and class −1 when $(w, x) + b < 0$. The mapping of input vectors in the form of high-dimensional feature space via an inferred chosen mapping function ϕ is to be done if the mapping is non-separable. By means of a Kernel function $k(x_t, x_s) = \varnothing(x_t)\varnothing(x_s)$, mapping can be done implicitly. There are four kinds of kernels like Linear, Polynomial with degree d, sigmoid and RBF kernels. In linear non separable case, training errors are allowed. So called slack variables ε^t are thus introduced in order to be tolerant of classification error.

4.3. Decision tree (C5.0)

C5.0 is a decision tree classifier and it is an extension of the earlier version i.e., C4.5 [40,41] and ID3 [42] decision tree classifier. C4.5 algorithms was suggested and implemented by Ross Quinlan and extended to C5.0 by involving efficient pruning of the decision tree construction. This classifier divides the dataset based on the attributes taken from training data. The idea of entropy is to train the data to develop the decision tree. C5.0 neglects values that are missing, and uses smart pruning to reduce size of the tree. It first grows the tree which over fits the data, and further, by using pruning method, it removes nodes and branches which are inefficient, and over fits the model on training (Rajeswari et al., 2017).

Algorithm for C5.0.
Checking of base cases.

- It normalizes information gain for every individual feature x^t by fitting the splitting strength of x^t.
- After evaluation, suppose x_b^t feature carries important information, a decision tree node is constructed which splits on xbt.
- The above steps are iterated for intermediate sub lists formed by splitting x_b^t.
- Pre pruning is involved to remove inefficient nodes to remove overfitting of data.

4.4. Random forest (RF)

Random Forest is a classification method which is an ensemble of classifiers approach [43,44,45,46]. Ensemble of decision tree is done using bagging method and decision dump is used as a base decision tree classifier. It combines decisions of several decision dump classifiers to generate a suitable result. Bagging ensemble approach modifies samples of data and randomly selects feature input; further, each sample is taken to train the decision dump classifier. Finally, aggregate result is computed by voting of all the classifiers taken for ensemble.

Algorithm of Random Forest.

Given: n^T- number of training examples, x^i-number of all features, x^e-number of features selected for Ensembles, m^i -number of all Ensemble members.

The Random forest is constructed using m^i trees.

For m^i classifiers, the below steps are performed -

a. Bagging method is performed to create n^T number of samples and each sample is used as training data for decision dump classifier.
b. For constructing trees for random forest, random features are selected using Gini index and the tree grows without trimming.
c. Training samples from bagging methods are applied on decision trees m^i to generate trained models. Further, trained models are evaluated using voting mechanism and a final classification decision is taken.

5. Experimental design

5.1. Software and hardware

Classification models and filtering techniques were performed in R-Studio and JAVA based environment (WEKA 3.8). All the computations were performed on a computer system with Window 8 operating system, 8 GB RAM and Intel CORE i3 processor. Libraries of R such as library (caret), library (e1071), library (C50), were used to test the classification models. Features were selected using three techniques namely, Information Gain, Gain Ratio and Chi-Square, using WEKA 3.8 software package.

5.2. Data splitting

A requisite for robust and accurate classifiers is good choice of training and testing split of the data set [47]. In literature, it has been suggested that if data instances are more, data should be split in 66% and 34% proportion for training and testing respectively. In this study, data was split in 66%–34% proportion for training and testing. In addition, 10-fold cross validation was also used to validate the results.

In the 10-fold cross validation [48], entire data is split in 10 parts with random sampling. At first, classifiers are trained with 10% data and 90% data is taken for testing the models. This process is repeated for 20–80, 30–70 … ….70-30, 80–20 and 90–10% for training set and testing set respectively. Finally, results are evaluated for each setting and aggregated with average for final outcome.

Out of 20 features, total 15 most informative features were selected

and machine learning classifiers were tested.

5.3. Evaluation metrics

To evaluate the feature selection mechanism and machine learning classifiers, five metrics [49] were used namely, Performance Accuracy, F-Value, False Positive, False Negative and Training Time (Table 2).

Performance accuracy of any classification model is calculated using confusion matrix. This is the ratio of all the accurately classified instances and all the instances.

F-Value is an effective way to calculate classification accuracy. It uses precision and recall harmonic mean to calculate the value. It evaluates percentage of defaulter and not-defaulter cases from the testing set after prediction.

False Positive Rate (FP Rate) is also known as misclassification rate of positive cases, which is considered good if the value is low or close to zero. It is also a good measure of sensitivity of a classifier where low FP indicates high sensitivity of the classifier.

False Negative Rate (FN Rate) is applied for negative instances which are misclassified. For a good classifier, False Negative value is preferred to be as low as possible. It also indicates specificity of a classifier where low FN means high specificity of the classifier.

Training time of the classifier was also calculated in this study. This metric is used for making a rapid model. Stop watch is used to compute the classifier training time which is measured in seconds.

6. Cost sensitive evaluation and discussion

In this study, five classifiers were used namely, Bayesian classifier, Naive Bayes (NB), Decision Tree (C5.0), Random Forest (RF), and Support Vector Machine with Radial Basis Kernel (SVM-RBF). Three feature/word selection techniques were used i.e., Information gain, Gain-Ratio and Chi-Square. Performance of classifiers was measured by important evaluation metrics i.e., Accuracy, F-Value, False-Positive rate, False Negative Rate and training time in seconds.

After testing of the classifier on the trained credit scoring machine learning models, the following observations were derived.

6.1. Observation 1

Results of Accuracy and F-Measure for 66%–34% training-testing split and 10-fold cross validation are demonstrated in Tables 3 and 4 where test results for both metrics are more or less the same for all the classifiers. Value of these metrics should be high for accurate prediction results. Analysis was done for all three feature selection techniques separately and the results noted.

In Chi-square feature selection technique, Accuracy and F-measure for RF Classifier was highest i.e., 76.20% and 76.18% respectively. In addition, C5.0 and SVM-RBF were the second and third best performers with accuracy 75.31% and 74.71% & F-Value 75.29% and 74.71% respectively. Results of probabilistic classifiers (Bayesian and NB) were the fourth and fifth best performers of the study with accuracy 71.01% and 69.74% respectively, and F-Value 71.00% and 69.71% respectively.

In Gain-ratio feature selection method, Accuracy and F-measure for RF and SVM-RBF classifiers were the highest i.e., with accuracy and F-value 77.40% for both classifiers. In addition, C5.0 was the second best performer with accuracy 74.15% and F-Value 74.12%. Results of probabilistic classifiers (Bayesian and NB) were the third best performers of the study with accuracy and F-Value 71.01% for both classifiers.

In Info-Gain feature selection technique, Accuracy and F-measure for SVM-RBF Classifier was the highest i.e., 78.24% for both metrics. In addition, RF and C5.0 were the second best performers with accuracy 75.31% and F-Value 75.29% for both metrics. Results of probabilistic classifiers (Naïve Bayes and Bayesian) were the third and fourth best performers of the study with accuracy 70.61% and 64.12% respectively, and F-Value 70.59% and 64.10% respectively.

Results of 10-fold cross validation, strongly supporting the results, came from 66 to 35% splitting method for Chi-square and Gain Ratio methods. However, results of Info-Gain were different. In addition, results of 10-fold cross validation showed better training than 66-34% splitting method. Chi-square and Gain-Ratio showcased the best prediction results for RF classifier. However, accuracy and F-Value were maximum for Chi-Square features i.e., 93.12% and 93.10% respectively. Results of Accuracy and F-Value of RF for Gain-Ratio features were 91.90% for both metrics. C5.0 came out as the second best classifier in 10-fold cross validation, with 92.21% accuracy and 92.20% F-measure for Chi-square feature selection. However, results for gain ratio and Info-gain were slightly less than the best one. SVM, NB and Bayesian techniques were under performers, with accuracy range between 70.12% and 77.70 within the range of all the feature selectors.

Table 2
Instruments for performance measure.

Instruments	Related Formulas
Performance Accuracy	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
F-Value	$F-Value = \frac{2*Precision*Recall}{Precision+Recall}$
False Positive	$FalsePositive = \frac{FP}{FP+TN}$
False Negative	$FalseNegative = \frac{FN}{FN+TP}$
Training Time	Using stopwatch (In Second)

TP = True Positive; TN = True Negative; FP=False Positive; FN=False Negative; Precision = TP/(TP + FP); Recall = TP/TP + FN.

Table 3
F-Measure and Accuracy of all classifiers (66–34).

Accuracy, F-Measure (in %)	Chi-square	Gain-Ratio	Info-Gain
Bayesian	71.01, 71.00	71.00, 71.00	64.12, 64.10
NB	69.74, 69.71	71.00, 71.00	70.61, 70.59
SVM	74.71, 74.71	77.40, 77.40	78.24, 78.24
C5.0	75.31, 75.29	74, 15, 74.12	75.31, 75.29
RF	76.20, 76.18	77.40, 77.40	75.31, 75.29

Table 4
F-Measure and Accuracy of all classifiers (10-Fold).

Accuracy, F-Measure (in %)	Chi-square	Gain-Ratio	Info-Gain
Bayesian	71.01, 71.00	71.01, 71.00	70.91, 70.90
NB	70.12, 70.10	70.43, 70.40	70.51, 70.50
SVM	77.02, 77.00	77.40, 77.40	77.70, 77.70
C5.0	92.21, 92.20	89.71, 89.70	89.21, 89.20
RF	93.12, 93.10	91.20, 91.20	90.90, 90.90

6.2. Observation 2

Tables 5 and 6 compare False Positive Rate and False Negative rate of different classifiers and classification techniques for 66-34% split and 10-fold cross validation respectively. For an efficient classifier, False Positive Rate and False Negative rate should be as low as possible. First, results of 66–34% split methods are discussed.

Table 5
False Positive and False Negative Rates of all classifiers (66–34).

FP Rate (in %)	Chi-square	Gain-Ratio	Info-Gain
Bayesian	23.80, 39.30	23.80, 39.30	08.40, 77.00
NB	29.41, 32.35	28.99, 30.69	29.41, 29.41
SVM	12.18, 55.88	02.57, 70.66	11.76, 45.09
C5.0	15.96, 45.00	20.20, 47.05	14.70, 48.03
RF	00.00, 23.00	09.24, 53.92	11.76, 59.90

Table 6
False Positive and False Negative Rate of all classifiers (10-Fold).

FP Rate (in %)	Chi-square	Gain-Ratio	Info-Gain
Bayesian	12.90, 61.0	12.90, 61.00	14.00, 59.70
NB	00.20, 99.00	00.40, 97.66	00.00, 98.00
SVM	02.50, 69.33	02.57, 70.66	02.50, 70.66
C5.0	0.00, 26.00	00.00, 34.00	00.04, 35.00
RF	0.00, 23.00	00.00, 29.00	00.00, 30.33

If we consider Chi-Square feature selection technique, Random Forest classifier showed the lowest FP and FN rate i.e., 00.00% and 23.00%, while C5.0, SVM, NB and Bayesian were the under performers in this study, with respective FP and FN rates - C5.0 (15.96%, 45.00%), SVM (12.18%, 55.88%), NB (29.41%, 32.35%) and Bayesian (23.80%, 39.30%).

In the Gain ratio feature selection technique, SVM classifier showed the lowest FP rate i.e., 02.57%; NB classifier had the lowest FN rate i.e., 30.69%. However, for FP rate, other classifiers were under performers with values RF (09.24%), C5.0 (20.20%), NB (28.99%) and Bayesian (23.80%). For FN rate, other classifiers were under performers with values RF (53.92%), C5.0 (47.05%), SVM (70.66%) and Bayesian (39.30%).

In the Information Gain technique, Bayesian classifier showed the lowest FP rate i.e., 08.40%; NB classifier had the lowest FN rate i.e., 29.41%. However, for FP rate, other classifiers were under performers with values RF (11.74%), C5.0 (14.70%), SVM (11.76%) and NB (29.41%). For FN rate, other classifiers were under performers with values RF (59.90%), C5.0 (48.03%), SVM (45.09%) and Bayesian (77.00%).

Results of 10-fold cross validation that strongly support the results came from 66 to 35% splitting method for Chi-square where RF classifier showed lowest FP and FN rate; however, the result of Gain ratio and Info-Gain were different. In addition, results of 10-fold cross validation showed better training than 66-34% splitting method. For all feature selection methods, RF classifier was found more sensitive in reducing False Positive Rate (FPR) and False Negative Rate (FNR), compared to other classifiers, with FPR range 00.00% for all feature selection methods and FNR ranging between 23.00% and 30.33% for all feature selection. In addition, C5.0 classifier was found second best with FPR 00.00%–00.04% within the range of all feature selection methods and FNR 26.00%–35.00% for all the feature selection methods. The remaining classifiers i.e., SVM, NB and Bayesian, were under performers, with FPR 00.00%–14.00% and FNR 59.70%–99.00%, within the range of all the feature selection methods.

6.3. Observation 3

Tables 7 and 8 compare the time taken for training of different classifiers and filtering techniques, for 66-34% splitting and 10-fold cross validation. As stated earlier, time taken for training of classifiers should be as low as possible. For 66%–34% splitting method, Bayesian classifier and C5.0 were found with rapid training time for Gain ratio and Information gain methods, and Naïve Bayes for Information gain method, with 0.01 s and for gain ratio 0.02 s. RF was found second best rapid classifier of this study with training time 0.18 s for gain ratio and 0.18 s for Information gain. However, for Chi-square feature selection method, RF was found an under performer with 12.91 s training time.

Table 7
Training Time of all classifiers.

Time taken (Sec)	Chi-square	Gain-Ratio	Info-Gain
Bayesian	03.01	00.01	00.01
NB	03.02	00.02	00.01
SVM	04.80	00.29	00.33
C5.0	06.00	00.01	00.01
RF	12.91	00.18	00.18

Table 8
Training Time of all classifiers.

Time taken (Sec)	Chi-square	Gain-Ratio	Info-Gain
Bayesian	03.04	00.01	00.01
NB	03.18	00.02	00.01
SVM	04.66	00.29	00.33
C5.0	08.22	00.01	00.01
RF	16.20	18.00	17.00

All the results from 66 to 34% are supported by 10-fold cross validation results. Again for this method, Bayesian classifier and C5.0 are found with rapid training time for Gain ratio and Information gain methods, and Naïve Bayes for Information gain method with 0.01 s, and for gain ratio 0.02 s. RF is found second best rapid classifier of this study with training time 0.18 s for gain ratio and 0.17 s for Information gain. However, for Chi-square feature selection method, RF is found to be under performer with 16.20 s training time.

7. Discussion

On comparing all the required metrics, accuracy is high and false positive rate is low for Random forest classifier and Chi-square feature selector. Even though training and testing time is high for RF classifier but as indicated this classifier is considered god due to high Accuracy and F-measure, and low False positive rate and False negative rate. Finally, after analysis, it can be said that Random Forest Classifier is good compared to other classifiers for predicting the credit score. In addition, Chi-square is found better in selecting important predictors to predict defaulter and Not-defaulter of credit.

Chi-square feature selection suggests that “Type of apartment” is the most informative predictor to decide whether a person can be defaulter or not. It clearly depicts that apartment type of the debtor where they live, is important to be considered and this predictor should be given a good score. In addition, other predictors such as “Concurrent Credits”, “Gender and Marital Status”, “Account Balance” are also found potential predictors in this study. In addition, “employment experience”, “other credits” and “current occupations” are found important predictors in calculating credit score. Out of 20 predictors, the top 15 predictors are selected. Credit score generators and financial institutions are advised that all the 15 top predictors reported in this study should be considered for calculating credit score to decide defaulters and not-defaulters of credit. Important predictors retained from best feature selection (Chi-Square) of this study are mentioned in Table 9.

In the case of German Credit, ratio of defaulter and not-defaulter debtors is 70% Not defaulter and 30% Defaulters. The data is considered imbalanced and a strong possibility is misclassification of cases that are defaulters because of less training instances. This study included FPR

Table 9
Chi Square Top 15 features.

S/N	Top 15 Informative features
1	Type of apartment
2	Concurrent Credits
3	Sex and Marital Status
4	Foreign Worker
5	Account Balance
6	Payment Status of previous credits
7	Value Savings/Stocks
8	Most Valuable available assets
9	Length of Current Employment
10	Guarantors
11	No of credits at this bank
12	Occupation
13	Telephone
14	No of Dependent
15	Duration of credits

and FNR also to tackle this problem. Specificity of the prediction models is found good due to low FPR (almost Zero for most of the classifiers). The main concern of this study is specificity (Low FNR). All the classifiers were evaluated with FNR and it is found that the proposed combination (Chi-Square & RF) is good in minimizing FNR. The proposed model is found good to tackle imbalanced data.

The proposed model only underperforms for training time which is 12–17 s. Since the proposed machine learning RF is an ensemble of classifiers techniques, it may take more training because multiple classifiers are involved. Training time is not very important if accuracy and false positive is good for the classifier. Once the classifier is trained, it is found to be the best prediction model.

8. Conclusion, limitations and future work

Credit Scoring is important for identifying defaulters of credit for which accurate information for prediction is required. At the start of this research, the objective was to identify good input predictors for building machine learning models, and the best feature selection and machine learning model. This study has successfully achieved its aims by comparing three feature selection techniques (Chi-Square, Information-gain and Gain-Ratio) and five Machine Learning Classifiers (Bayesian, Naïve Bayes, SVM (support Vector Machine), Decision Tree (C5.0), and Random Forest). At first, feature selection techniques were compared where Chi-square feature selection method was found suitable with most informative predictors for all machine learning models. For machine learning models, overall, Random forest was found best amongst other machine learning classifiers and gives up to 93% accuracy. The RF model was also capable of reducing False Positive (Type 1 error) and False Negative (Type 2 error). Finally, this research is able to find an appropriate combination of Random Forest machine learning model. Chi-square feature selection method is found to be a good choice to build a robust, accurate and sensitive credit scoring model. However, training time of this classifier was found slightly high, but in consideration of other metrics, training time is not a big concern. Additionally, decision tree (C5.0) showed comparable results and was the second best performer of this study.

As with other studies, this study also has certain limitations. Only German Credit data was taken in this research. However, a broad scope is open to test other credit data on the identified prediction model. Three popular feature selection methods and five machine learning classifiers have been tested in this research. In future studies, other feature selection methods and machine learning classifiers may also be incorporated for identifying credit scoring predictors. Further, other splitting methods may also be incorporated to check credibility of the machine learning models.

Author statement

I am the sole author of this manuscript and I certify that I have done all the work of this research and I am taking public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

Acknowledgements

The author thank the anonymous referees, and the editor for their valuable feedback, which significantly improved the positioning and presentation of this paper.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.techsoc.2020.101413.

References

[1] J.N. Crook, D.B. Edelman, L.C. Thomas, Recent developments in consumer credit risk assessment, Eur. J. Oper. Res. 183 (3) (2007) 1447–1465.
[2] D.J. Hand, W.E. Henley, Statistical classification methods in consumer credit scoring: a review, J. Roy. Stat. Soc. 160 (3) (1997) 523–541.
[3] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, J. Vanthienen, Benchmarking state-of-the-art classification algorithms for credit scoring, J. Oper. Res. Soc. 54 (6) (2003) 627–635.
[4] I. Partalas, G. Tsoumakas, I. Vlahavas, An ensemble uncertainty aware measure for directed hill climbing ensemble pruning, Mach. Learn. 81 (3) (2010) 257–282.
[5] P.J. García-Teruel, P. Martinez-Solano, Determinants of trade credit: a comparative study of European SMEs, Int. Small Bus. J. 28 (3) (2010) 215–233.
[6] R. Calabrese, Predicting bank loan recovery rates with a mixed continuous-discrete model, Appl. Stoch Model Bus. Ind. 30 (2) (2014) 99–114.
[7] X. Yao, J. Crook, G. Andreeva, Support vector regression for loss given default modelling. European Journal of Operational Research, 240(2), 2015.
[8] F.N. Koutanaei, H. Sajedi, M. Khanbabaei, A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring, J. Retailing Consum. Serv. 27 (2015) 11–23.
[9] S. Sadatrasoul, M. Gholamian, K. Shahanaghi, Combination of feature selection and optimized fuzzy apriori rules: the case of credit scoring. International arab journal of information technology (IAJIT), 12(2), 2015.
[10] V.S. Ha, H.N. Nguyen, Credit scoring with a feature selection approach based deep learning, in: MATEC Web of Conferences, vol. 54, EDP Sciences, 2016, p. 5004, 0.
[11] S. Maldonado, J. Pérez, C. Bravo, Cost-based feature selection for Support Vector Machines: an application in credit scoring, Eur. J. Oper. Res. 261 (2) (2017) 656–665.
[12] S. Jadhav, H. He, K. Jenkins, Information gain directed genetic algorithm wrapper feature selection for credit rating, Appl. Soft Comput. 69 (2018) 541–553.
[13] Y. Liu, A. Ghandar, G. Theodoropoulos, Island model genetic algorithm for feature selection in non-traditional credit risk evaluation, June, in: 2019 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2019, pp. 2771–2778.
[14] Wei Chen, Zhongfei Li, Jinchao Guo, A VNS-eda algorithm-based feature selection for credit risk classification, Math. Probl Eng. 2020 (2020).
[15] J. Nalić, G. Martinović, D. Žagar, New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers, Adv. Eng. Inf. 45 (2020) 101130.
[16] N.I. Nwulu, S.G. Oroja, M. Ilkan, A comparative analysis of machine learning techniques for credit scoring. International Information Institute (Tokyo), Information 15 (10) (2012) 4129.
[17] A. Bequé, S. Lessmann, Extreme learning machines for credit scoring: an empirical evaluation, Expert Syst. Appl. 86 (2017) 42–53.
[18] D.A.V. de Paula, R. Artes, F. Ayres, A.M.A.F. Minardi, Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques, RAUSP Management Journal (2019).
[19] X. Dastile, T. Celik, M. Potsane, Statistical and machine learning models in credit scoring: a systematic literature survey. Applied Soft Computing, 106263, 2020.
[20] G. Teles, J.J. Rodrigues, K. Saleem, S. Kozlov, R.A. Rabêlo, Machine learning and decision support system on credit scoring, Neural Comput. Appl. 32 (14) (2020) 9809–9826.
[21] B. Fallah, K.T.W. Ng, H.L. Vu, F. Torabi, Application of a multi-stage neural network approach for time-series landfill gas modeling with missing data imputation, Waste Manag. 116 (2020) 66–78.
[22] R.S. Somasundaram, R. Nedunchezhian, Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values, Int. J. Comput. Appl. 21 (10) (2011) 14–19.
[23] B. Azhagusundari, A.S. Thanamani, Feature selection based on information gain, Int. J. Innovative Technol. Explor. Eng. 2 (2) (2013) 18–21.
[24] S.K. Trivedi, S. Dey, A modified content-based evolutionary approach to identify unsolicited emails, Knowl. Inf. Syst. 60 (3) (2019) 1427–1451.
[25] J. Dai, Q. Xu, Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification, Appl. Soft Comput. 13 (1) (2013) 211–221.
[26] S.K. Trivedi, S. Dey, A comparative study of various supervised feature selection methods for spam classification, March, in: Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, 2016, pp. 1–6.
[27] L. Ali, C. Zhu, M. Zhou, Y. Liu, Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection, Expert Syst. Appl. 137 (2019) 22–28.
[28] S. Bahassine, A. Madani, M. Al-Sarem, M. Kissi, Feature Selection Using an Improved Chi-Square for Arabic Text Classification, vol. 32, Journal of King Saud University-Computer and Information Sciences, 2020, pp. 225–231, 2.
[29] S.K. Trivedi, S. Dey, A. Kumar, Capturing user sentiments for online Indian movie reviews. The Electronic Library, 2018.
[30] M. Cubric, Drivers, barriers and social considerations for AI adoption in business and management: a tertiary study. Technology in Society, 101257, 2020.
[31] K. Naveed, C. Watanabe, P. Neittaanmäki, Co-evolution between streaming and live music leads a way to the sustainable growth of music industry–Lessons from the US experiences, Technol. Soc. 50 (2017) 1–19.
[32] S. Fox, Mass imagineering: combining human imagination and automated engineering from early education to digital afterlife, Technol. Soc. 51 (2017) 163–171.

[33] M. Coccia, Deep learning technology for improving cancer care in society: new directions in cancer imaging driven by artificial intelligence, Technol. Soc. 60 (2020) 101198.
[34] M. Al-Emran, V. Mezhuyev, A. Kamaludin, Towards a conceptual model for examining the impact of knowledge management factors on mobile learning acceptance. Technology in Society, 101247, 2020.
[35] C.A. Wongnaa, S. Babu, Building resilience to shocks of climate change in Ghana's cocoa production and its effect on productivity and incomes, Technol. Soc. 62 (2020) 101288.
[36] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2–3) (1997) 131–163.
[37] D.D. Lewis, Naive (Bayes) at forty: the independence assumption in information retrieval, April, in: European Conference on Machine Learning, Springer, Berlin, Heidelberg, 1998, pp. 4–15.
[38] O. Chapelle, V. Vapnik, Model selection for support vector machines. In Advances in neural information processing systems (pp. 230–236, 2000.
[39] V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[40] S.K. Trivedi, P.K. Panigrahi, Spam classification: a comparative analysis of different boosted decision tree approaches, J. Syst. Inf. Technol. (2018).
[41] S.K. Trivedi, S. Dey, Analysing user sentiment of Indian movie reviews. The Electronic Library, 2018.
[42] M. Shaheen, T. Zafar, S. Ali Khan, Decision tree classification: ranking journals using IGIDI, J. Inf. Sci. 46 (3) (2020) 325–339.
[43] M. Pal, Random forest classifier for remote sensing classification, Int. J. Rem. Sens. 26 (1) (2005) 217–222.
[44] Tripathi, A., Sharma, R. D., & Trivedi, S. K. Identification of plant species using supervised machine learning. Int. J. Comput. Appl., 975, 8887.
[45] S.K. Trivedi, S. Dey, An enhanced genetic programming approach for detecting unsolicited emails, December, in: 2013 IEEE 16th International Conference on Computational Science and Engineering, IEEE, 2013, pp. 1153–1160.
[46] L.V. Utkin, An imprecise deep forest for classification, Expert Syst. Appl. 141 (2020) 112978.
[47] R.R. Picard, K.N. Berk, Data splitting, Am. Statistician 44 (2) (1990) 140–147.
[48] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat. Comput. 21 (2) (2011) 137–146.
[49] G.S. Handelman, H.K. Kok, R.V. Chandra, A.H. Razavi, S. Huang, M. Brooks, H. Asadi, Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods, Am. J. Roentgenol. 212 (1) (2019) 38–43.