

COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHMS FOR DETECTING BANKRUPTCY

BY

Maria Sultana Keya

ID: 171-15-1215

Himu Akter

ID: 171-15-1212

&

Md. Atiqur Rahman Sozib

ID: 171-15-1283

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Sabab Zulfiker

Lecturer

Department of CSE

Daffodil International University

Co-Supervised by:

Amit Chakraborty Chhoton

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

APPROVAL

This Project titled “**COMPARISON OF DIFFERENT MACHINE LEARNING ALGORITHM FOR DETECTING BANKRUPTCY**”, submitted by **Maria Sultana Keya, Himu Akter, and Md. Atiqur Rahman Sozib** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 15-01-2021.

BOARD OF EXAMINERS

Chairman

Professor Dr. Touhid Bhuiyan

Professor & Head

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Dr. S. M. Aminul Haque

Associate Professor & Associate Head

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner

Ohidujjaman

Senior Lecturer

Department of CSE

Faculty of Science & Information Technology

Daffodil International University

External Examiner

Dr. Mohammad Sorif Uddin

Professor

Department of Computer Science & Engineering

Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sabab Zulfiker, Lecturer, Department of CSE**, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Md. Sabab Zulfiker

Lecturer

Department of CSE

Daffodil International University

Co-Supervised by:

Amit Chakraborty Chhoton

Lecturer

Department of CSE

Daffodil International University

Submitted by:

Maria Sultana Keya

ID: 171-15-1215

Department of CSE

Daffodil International University

Himu Akter

ID: 171-15-1212

Department of CSE

Daffodil International University

Md. Atiqur Rahman Sozib

ID: 171-15-1283

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty **Allah** for His divine blessing that makes it possible to complete the final year project/internship successfully.

We are really grateful and wish our profound indebtedness to **Md. Sabab Zulfiker**, Lecturer, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Comparison of Different Machine Learning Algorithms for Detecting Bankruptcy” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate at Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

There have been severe experiments from academics and merchandisers concerning models for Predicting bankruptcy. The paper propounds an extensive rethink of work done during 5 years in the petition of intellectual strategy to accomplish bankruptcy prediction problems. Several machine learning directions are being used in this paper for Predicting bankruptcy. Some algorithms: AdaBoost, Decision tree, J48, Bagging, Random Forest are used in this paper. By traditional models, machine learning models offer enhancing bankruptcy prediction accuracy. Different types of models are tested using several evaluation metrics. The five years Bagging accuracy range is 95% within 97% among another model. Here we include k-fold cross-validation($k=10$) to measure our accuracy. Bagging accuracy is high in this paper. A confusion matrix is used to recount the perfection of a classification model that gives true values for knowing.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of Examiners.....	ii
DECLARATION.....	iii
ACKNOWLEDGEMENT.....	v
ABSTRACT.....	vi
CHAPTER 1: INTRODUCTION.....	1
1.1 Background.....	3
1.1.1 How bankruptcy works.....	3
1.1.2 Types of bankruptcy.....	3
1.1.3 Causes of bankruptcy.....	4
1.2 Motivation of the research.....	4
1.3 Problem Statement.....	4
1.4 Research questions.....	5
1.5 Research objectives.....	5
1.6 Research scope.....	5
1.7 Research organization.....	5
CHAPTER 2: LITERATURE REVIEW	6
CHAPTER 3: RESEARCH METHODOLOGY	9
3.1 Introduction.....	9
3.2 Data description.....	9
3.3 Algorithm description.....	10
3.3.1 Adaboost.....	10
3.3.2 Decision Tree.....	10
3.3.3 J48.....	11
3.3.4 Bagging.....	11
3.3.5 Random forest.....	12
3.4 Implementation process.....	13
3.4.1 Data Input.....	14

3.4.2 Data Handling.....	14
3.4.3 Split Data.....	14
3.4.4 Recursive Feature Elimination (RFE).....	14
3.4.5 Synthetic Minority Oversampling (SMOTE).....	15
3.4.6 Machine Learning Classifier.....	15
3.4.7 K-fold Cross-Validation.....	16
3.4.8 Classifier Result.....	16
3.4.9 Best Model.....	16
3.5 Data visualization	17
CHAPTER 4: RESULTS AND DISCUSSION.....	19
CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS.....	27
5.1 Conclusion.....	27
5.2 Future Work.....	27
CONTRIBUTION FROM THIS R&D PROJECT.....	28
REFERENCES.....	28
APPENDICES.....	31
PLAGIARISM REPORT	32

LIST OF TABLES

Tables	PAGE NO
Table: 1.1 1st year: Evaluation Outcome Machine Learning Algorithm To Measure Bankruptcy.....	20
Table: 1.2 2nd year: Evaluation Outcome Machine Learning Algorithm To Measure Bankruptcy.....	21
Table: 1.3 3rd year: Evaluation Outcome Machine Learning Algorithm To Measure Bankruptcy.....	22
Table: 1.4 4th year: Evaluation Outcome Machine Learning Algorithm to Measure Bankruptcy.....	23
Table: 1.5 5th year: Evaluation Outcome Machine Learning Algorithm To Measure Bankruptcy.....	24
Table: 2 Confusion Matrix for Machine Learning Techniques bankruptcy prediction.....	25
Table: 3 Accuracy from different kinds of algorithms in five.....	26
Table: 4 Proposed model and Accuracy.....	26

LIST OF FIGURES

FIGURES	PAGE NO
Fig.1: The graphical illustration of the method for projecting the bankruptcy of various algorithms step by step.....	13
Fig.2: Recursive Feature Elimination.....	17
Fig.3: RFE-Feature Importance.....	18

CHAPTER 1

INTRODUCTION

When a company or industry or individual person can't payment debit that is considered as Bankruptcy. Bankruptcy statistics are also a tracing guide. Bankruptcy comes from Italian Banca rotta, that verbally objective of "broken bench" Bankruptcy did not attend in archaic Greece. The Premier act of Parliament under English regulations conducting with bankruptcy according to the act of bankrupts of 1542. Bankruptcy is recorded in East Asia. A law, which ordered just dying castigation for any man who became bankrupt 3 times in accordance with the Maqrizi the Yassa that included Genghis Khan. To meet debenture repayments has been watched for many purposes and considering the defect of a nation. A borrower nation for defeating to pay the debt to impound assets to insist its creditors right. For example, Egypt was being attacked by the United Kingdom (UK) in 1882 and the United States (US) 'gunboat diplomacy' in Venezuela in 1890, and starting in 1951 the United States (US) control of Haiti[1].

The progress of international capital markets was fully restricted in accordance with Kenneth's Rogoff that we nevertheless list numerous faults of France, Portugal, Russia, Spain, and the Italian city States. That is significant to count the crux and find out the hazard in monetary fatalities[2].

Bankruptcy hype is a White-collar crime that is defined by the sociologist Edwin Sutherland in 1939. It carried wage theft, fraud, bribery, cybercrime, labor racketeering[3]. In the United States (US) bankruptcy hype is a federative crime. Bankruptcy laws are different in several countries like bankruptcy is restricted to individual's are used to collect in the United Kingdom (UK). In the United States (US) bankruptcy is used to formal solvency, bankruptcy is restricted to industries and individuals in Finland. Bankruptcy used to individuals and governed by the federative bankruptcy Act 1966 in Australia law. In Australia 'Corporate law Economic reform program Act 2004' simplified the law 3354 pages pygmy of other nations like Sweden that incorporation law is less than 200 pages long[4].

The superintendent bankruptcy indicated that executors in Canada. The client section 2011 is separated into 77993, bankruptcies, and 45006 client proposals. From 2010 the simplification is described as 8.9%[5].

Bankruptcy is used in Ireland only to be natural persons. In the first week of May 2016 'The Parliament of India' obtained solvency and bankruptcy code 2016. In 1920 the earlier law that is called the 'Provincial Insolvency Act'. The Insolvency and Bankruptcy Code (IBC) contains 255 sections and 11 schedules in India[6].

Individuals and partnerships bankruptcy described in the United Kingdom. An intimate in Scotland should have more than £1500 debt. Bankruptcy can be resultant of Insolvency according to Swiss law. This is a court order from debt incitement to the progressing that is used in general planted commercial products. In Sweden bankruptcy is an external method that may include an institute or personal and not similar to solvency. In France, more than 40000 insolvencies coming were unveiled in Spain all together approximately bad due to debit value in France was 1.3%, Spain 2.6% by comparison.

Every year many industries fall into bankruptcy. The reserved fair is actually essential and investors are expounded to high dimensions in Bangladesh. In Bangladesh the economic sections are small and sub improvement and that section stays an outgoing and banking principle Fair. The Bangladesh law was invoked with the bankruptcy of several only. Bankruptcy is the resolution to debt problems. Particular income is unaffected by bankruptcy. Bankruptcy doesn't get into social security, veteran benefits, payments, and payments from pensions. Various samples can be accomplished that will be beneficial to commit judgment.

Machine learning algorithms is an unbelievable achievement in the bounds of Artificial Intelligence. Technology can enhance our lives with the help of machine learning approaches. In modern times machine learning algorithms are being used in multiple bounds and industries. For sampling as medical diagnosis, image processing, marketing, prediction, weather forecasting, socio-economic behavior analysis, classification, learning association regression, and many other things. As machine learning algorithms use prediction, so we use machine learning algorithms for bankruptcy prediction. Many types of knowledge can be found and this knowledge can be used to predict the assignment status by using various machine learning and data mining techniques. After the performance of

this paper, it will help to search for various approaches to predict bankruptcy and keep down the best approach for prediction. The main purpose of this paper is to predict bankruptcy by using machine learning algorithms, overcome risk, and balance the consequences of various machine learning algorithms for recognizing which algorithm gives the best performance.

1.1 Background

1.1.1 How bankruptcy work:

Bankruptcy works before, after, and during the time. Before time bankruptcy is an alternative to investigate assets debts and individual budgets to realize that bankruptcy will price. Bankruptcy is requesting any information bankruptcy testator asks for debts and assets wants to finish bankruptcy to collect all receipts that have been accepted last month. During most lawsuits the bankruptcy stage is humdrum. During the bankruptcy, requisites to complete the menstruous budget, collect information about money, taxes, monthly payments, and many other things. After emission, all debts are formally dispelled because it contains too much debt.

1.1.2 Types of bankruptcy:

There are mainly two types of bankruptcy.

Liquidation

Restructuring

Liquidation: Liquidation is the only way to fully dead wind up an industry and shut down. For the business, it mentions the industry administrator stop all measures employed are removed and bank accounts are outstanding. It offers bankruptcy advancing the area.

Restructuring: Restructuring is the collective administration for the regulation of reorganization of the legitimate proprietor operational or another form of an industry for the objective of building that more beneficially and currently exigent. For reasoning restructuring to alleviate price, to enhance the competitive benefit and to reduce debt.

1.1.3 Causes of bankruptcy:

When industries are not economical for elongated times proprietors may be labored to go information bankruptcy to attend the market. There are many causes of bankruptcy.

Building agreement: Building agreement in the overall financial and inelastic market in that a business handler are general causes of bankruptcy.

Impoverished judgment making: Lack of planning and inches heading deciding can lead to a speedy decision and business defeat.

Other causes: It can outcome a host of other redialing which suppress profitability bankruptcy. In order, causes of bankruptcy include poor business position, damage to the original employer, personal issues, unwanted inversion likened storms, fraud.

1.2 Motivation of the research:

We use machine learning algorithms to predict bankruptcy. Several reasons to motivate bankruptcy prediction. Most important factors, bankruptcy is motivating current economic problems. An unexpected expense, bad budgets, bad loans become another reason that bankruptcy is motivated.

1.3 Problem Statement:

We want to generate excellent prediction methods concerning accuracy time and learning defeat into elected methods by applying several machines learning Classifiers and assimilate the outcomes after data preprocessing and disposal methods.

1.4 Research questions:

Can we allow a better specialty viewpoint to predict bankruptcy more correctly to the dataset on machine learning repertory?

Can we find better predictive machine learning methods that can predict the appearance of bankruptcy to reduce from statistical experiment more correctly with lower FN(False Negative) count?

1.5 Research objectives:

To allow a better alleviate property viewpoint on the dataset used.

To find a better predictive machine-learning algorithm to find more correctly with a Lower FN count.

1.6 Research scope:

Classification based machine learning algorithms can predict bankruptcy congeniality or not. We find better accuracy using machine learning approaches. In the future, we work in deep learning to make decisions as human brains.

1.7 Research organization:

The rest of the report is organized as follows in chapter 2 we discuss related work. In chapter 3, we discuss methodology. In chapter 4, we describe the result and performance analysis. And after all, in chapter 5, we conclude this report.

CHAPTER 2

LITERATURE REVIEW

Zelenkov et al. [7] proposed a two-step classification method (TSCM) based on genetic algorithms. It selects both the pertinent reasons and the model converts to the application. It included 912 observations and 55 features. The proposed method showed the best correctness (0.934) between the checked models and as well showed the maximum balance accuracy-withdrawal ratio.

Loukeris et al. [8] Corporate accounting statements provide tax services and financial markets including valuable data on the financial health of companies. Helpful implementation in the processing is enhanced economical and accounting data procedure of artificial intelligence. The limitation is that high-precision analysis is limited for each individual organization.

Tomczak [9] used this integrated model in corporate bankruptcy assessments to ensure the company's survival in the face of increasing competition in the market. The five liquidity ratios of bankruptcy and operating were analyzed here. In order to change the financial situation, the indicators were reflected in the previous year of bankruptcy. They have limitations in different sectors of manufacturing and analysis of insolvent companies.

Devi et al. [10] support statistics and machine learning strategies in reviews based on numerous statistics and machine learning to solve the problem of bankruptcy forecasting. Also, optimization strategies, E.g. Genetic Algorithms (GA) and Particle Swarm Optimization (PSO), were desegregated to further improve predictions. Limited to machine learning techniques using the Apache Mahout tool.

Fernández et al. [11] used focused and unfocused models for bankruptcy forecasting developed using financial data for covering various economic sectors. Which in turn determines excellence by comparing models and helps make decisions about using a particular model.

Chen [12] Collected data for compared some traditional statistical methods and it has made four important contributions: (1) Almost 80% less financial ratio, (2) Traditional statistical methods are better handler, (3) Given empirical results is C5.0 and CART supply the greatest prediction, and (4) Support Vector Machines. Limited in long-term predictions.

Erdogan's [13] goal was to analyze bank bankruptcy using practical measures using support vector machines (SVMs) to analyze financial ratios. For this, he used data sets. Restricted to build the best results for SVMs.

Kim et al. [14] proposed to discover the consequences of bankruptcy decisions using data mining methods based on genetic algorithms. The results showed that the rules derived from genetic algorithms have higher accuracy and greater coverage than early learning. Lacking in the GA method. They have limitations in redundant and overlapping structure.

Alaminos et al. [15] goal was to create a worldwide model with a high volume to predict bankruptcy all over the world. They used a systematic framework of logistics that filled this gap in the empirical literature of this study. Restricted in the availability of farm data. In the future, a methodology needs to be set up to investigate what affects macro conditions.

Laitinen et al. [16] used Taylor's expansion to solve problems related to the functional form of bankruptcy forecasting models. The functionality of Taylor's model in the case of bankruptcy is assessed by putting it in the logistic regression model. They have some limitations to predict bankruptcy.

Martin et al. [17] used an ant-miner algorithm to predict qualitative bankruptcy. Qualitative bankruptcy forecast the rules have been persuaded by classifying bankruptcy different classes. The correctness of the regulations can be advanced by experimenting with large data sets in Ant-Miner.

Chen et al. [18] used a feasible and transfigure of eXtreme Gradient Boosting (xgboost) which is accomplished by Gradient Boosting structures. It is designed to be extensible so that users can easily define their own objectives. Limited in xgboost.

Shumway [19] performed the Hazard methods are more suitable to unit part of methods to count bankruptcy. He mentioned an easy method for estimating a separated-period hazard model and he proposed a model that was more accurate than those of alternative models. The limitation of the hazard model, the coefficient is quite small.

Zięba et al. [20] proposed a fiction model for forecasting bankruptcy which takes advantage of the Extreme Gradient Boosting for learning a gift of the decision tree. They used synthetic features like a sample regress method as it is exhibited in an evolutionary process.

Martin et al. [21] proposed a model by using two algorithms to improve the prediction accuracy. Fuzzy ID3 Algorithm and Ant Colony Optimization Algorithm (ACO) are provided for more accurate prediction. Restricted in Ant Colony Algorithm. Needed to improve for better prediction.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction:

This section is split into 3 subsections: data description, algorithms description, implementation procedures. The subsections are concisely delineated below:

3.2 Data description:

Dataset is on bankruptcy predictions of the Polish association. The information had gathered from Emerging Markets Information Service (EMIS, [Web Link]), which is an information base bearing data on developing business sectors over the globe. Slave association disbanded during the period 2000-2012, while the existing organizations are still being evaluated from 2007 to 2013.

Forming on the gathered information 5 characterization event had recognized, that relies upon the determining time frame:

- 1stYear the information including money related division from the first year of the gauging time frame and comparing group names shows chapter 11 status following 5 years. The information-carrying 7027 occurrences (fiscal summaries), 271 speaks to the bankrupted association, 6756 firms that didn't impoverish in the determining time frame.
- 2ndYear the information including monetary division off the second year of the gauging time frame and relating class name that demonstrates insolvency status following 4 years. The information contains 10173 occasions (budget reports), 400 speaks to the bankrupted association, in the gauging time frame 9773 firms that didn't go bankrupt.
- 3rdYear the information including budgetary division out of this third year of the gauging time frame and relating class name that demonstrates liquidation status following 3 years. The information including 10503 occasions (budget summaries), 495 speaks to bankrupted organizations, 10008 firms didn't bankrupt in the estimating time frame.
- 4thYear the information-carrying money related division out of the fourth year of the estimating time frame and comparing class marks that demonstrate chapter 11 status following 2 years. The information including 9792 events (fiscal reports), 515 speaks to the bankrupted association, 9277 firms that didn't impoverish in the anticipating time frame.
- In 5th Year information including budgetary events from the fifth year of the gauging time frame and relating class mark that demonstrates liquidation status following 1 year.

The information includes 5910 occasions (budget reports), 410 speaks to the bankrupted association, 5500 firms didn't impoverish in the anticipated time frame.

This research has been performed using 64 attributes and one class for predicting the company's future which will show the result as bankrupt or not.

3.3 Algorithm description:

3.3.1 Adaboost:

AdaBoost stands for adaptational Boosting classifier. A group of weak classifiers is combined into a robust one using this approach. Here, the subsequent equation represents the classification exploitation of the AdaBoost algorithmic rule.

$$F(x) = \text{sign}\left(\sum_{m=1}^M \theta_m f_m(x)\right),$$

Here, the f_m weak classifier is portrayed by f_m and θ_m represents the corresponding weight.

3.3.2 Decision Tree:

A Decision Tree may be a tabular illustration of inputs versus rules/cases/test conditions. It's an awfully effective tool used for each complicated software package testing and necessities management. multidimensional language helps to see all potential combos of conditions for testing and testers may determine incomprehensible conditions simply. The conditions are indicated as True(T) and False(F) values.

The decision of cacophonous the data is controlled by entropy and which may be outlined by the equation below, wherever p_j is that the likelihood of the j th ca

$$E(S) = -\sum_{j=1}^c p_j \log_2 p_j$$

3.3.3 J48:

The J48 formula is employed to classify totally different applications and perform correct results of the classification. J48 formula is one of the simplest machine learning algorithms to look at the information flatly and unendingly.

3.3.4 Bagging:

The Bootstrap integration, in addition to what is perceived to be content, can be a machine learning tool that incorporates a meta-algorithm designed to increase the soundness and accuracy of machine learning algorithms used in phases and retrieval. Adding reduces variability and helps to avoid overdoing it. Bootstrap predictive models (bagging) can be a common way to incorporate many types of predictive models and incorporate them (or incorporate them) into bulk predictions. the asset may be a clear rule of thumb for all copies of the bootstrap measurement unit of the main training activity performed, the retrospective or subdivision rule (often referred to as a lower student result) is applied to each bootstrap sample and, within the context of retrospect, a new predictive measurement unit predictions from readers of each foundation. Once the problem of segregation is addressed, the lowest student prediction rating scale includes a bullying vote in bulk or by measuring class-potential opportunities that may be open. this will be drawn to Equation where X is a record that we have a tendency to predict $\hat{f}_{bag}$ is the bagged prediction, and $\hat{f}_1(\square), \hat{f}_2(\square), \dots, \hat{f}_b(\square)$ are predictions from readers of each foundation

$$\hat{f}_{bag} = \hat{f}_1(X) + \hat{f}_2(X) + \dots + \hat{f}_b(X)$$

Because of the aggregation method, bagging effectively reduces the variance of a private base learner (i.e., averaging reduces variance); but bagging doesn't perpetually improve upon a private base learner. Bagging works particularly well for unstable, high variance base learners—algorithms whose foretold output undergoes major changes in response to little changes within the coaching knowledge. This includes algorithms like call trees and KNN (when k is sufficiently small). However, for algorithms that area unit a lot stable or have a high bias, sacking offers less improvement on foretold outputs since there's less variability.

3.3.5 Random forest:

The random forest has nearly identical hyperparameters as a call tree or a textile classifier. Random forest adds extra randomness to the model, whereas growing the trees. rather than finding out the foremost necessary feature whereas ripping a node, it searches for the simplest feature among a random set of options. Working of Random Forest algorithmic program. We can perceive the operating of the Random Forest algorithmic program with the assistance of the following steps –

Step one 1st, begin with the choice of random samples from a given dataset.

Step a pair of Next, this algorithmic program can create a wire tree for each sample. After that, it will get the prediction effect on each call tree.

Step three during this step, options are performed for each foretold result.

Step four finally, choose the foremost voted prediction result because of the final prediction result. When victimizing the Random Forest rule to resolve regression issues, you're victimizing the mean square error (MSE) to however your information branches from every node.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^N (x_i - y_i)^2$$

3.4 Implementation process:

The implementation procedures are illustrated in this part.

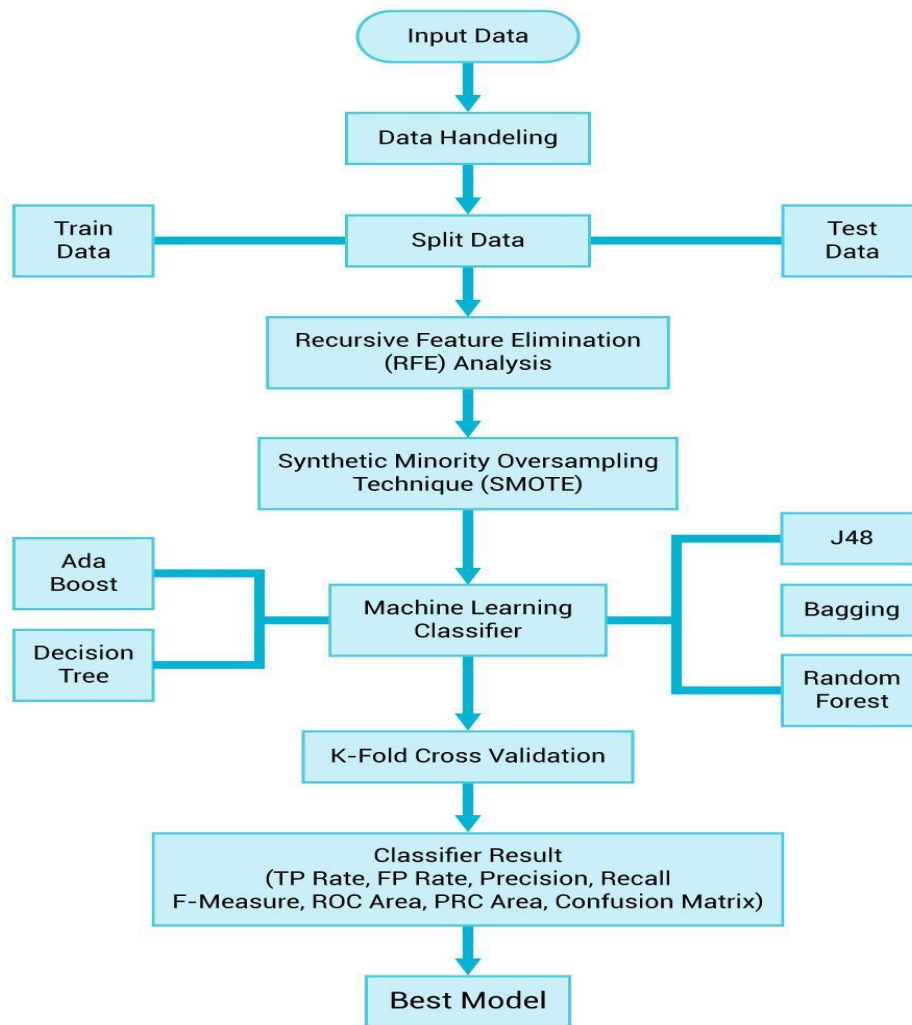


Fig. 1: The graphical illustration of the method for projecting the bankruptcy of various algorithms step by step.

3.4.1 Data Input:

After collecting the data for 5 years we are implementing that data in the proposed system has been performed in this step. For this research purpose, we have 64 attributes and 1 class as input data.

3.4.2 Data Handling:

Data Handling refers to the way toward social events, recording and introducing data in a way that is useful to other people - for example, in diagrams or graphs. In this part of handling data, we gathered 5 years of data and all the missing values in that data table. We gather data from different types of attributes and record the data with accuracy and precision into our data table so that the data can be useful to predict bankruptcy. We analyzed all the data to draw the best conclusion on which data will be useful to others.

3.4.3 Split Data:

Splitting a data-set in splitting the data training or testing. Split in this report the methodology is used for training or testing.

3.4.4 Recursive Feature Elimination (RFE):

Recursive feature elimination (RFE) is a method of eliminating a feature that compatibles the pattern and shifts a weak feature (or features) until a specified number of features are reached. RFE is popular because it is easy to configure and use and because it is useful in selecting those elements (columns) in the training database that is most important or most relevant to predicting the volatility of the target.

There are two important configuration options when using RFE: the selection of the number of features to select and the selection of the algorithm used to help select the features. Both of these hyper-parameters can be tested, although the method of operation is not very dependent on these well-prepared hyper-parameters. Applying Recursive feature elimination (RFE) we gain the 12 best attributes in our data set. So Recursive feature elimination (RFE) is an effective way to remove features in the training database of feature selection. And the use of RFE is the selection of the separation feature and the predictive modeling problems. Though the 5 years datasets quantity is the same so in this paper, we only set one-year features.

3.4.5 Synthetic Minority Oversampling (SMOTE):

Using an out-of-box machine algorithm is a problem where one class in the training set dominates the other. The Synthetic Minority Over-Sampling Technique (SMOTE) solves this problem.

SMOT (the process for the sampling of a few samples) is one of the most widely used methods of congestion to solve the problem of inequality. It aims to balance the distribution of a segment by periodically increasing the examples of the minority category by repeating them. SMOT incorporates a few new developments among a few existing ones.

Unequal data handling:

- * Change the test matrix. If we use an incorrect test matrix on unequal data, it can give us misleading results.
- * An example of a database. Sample means changing the unequal distribution of classes across the database.
- * Change the algorithm and the path to the problem.

Strong evidence has shown that SMOTE tends to perform better than the sub-sample or sample. And fraud detection has proven to be very important.

3.It checks the number of selected features and the algorithm wrapped in the RFE process.

3.4.6 Machine Learning Classifier:

In this part, we use five different types of machine learning classifiers to predict the accuracy of precision in bankruptcy prediction. Type of algorithms used for prediction in this project is Adaboost, Decision Tree, J48, Bagging, Random Forest.

3.4.7 K-fold Cross-Validation:

To appreciate machine learning patterns on a restricted data swatch that resample the sequence of cross-validation. Cross-validation is generally used to sacrifice machine learning algorithms to count its ability of machine learning methods on invisible data. This contains individual parameters mentioned K and data sample is to intersect that mention the number of classes. A sequence is aged to count the ability on new data of the patterns in K-fold cross-validation. K is selected by the reference to the mode like K=10 contains 10fold. K-fold cross-validation is considered the easiest model and easily works on large data. It is also likened the first method to another method. It is simple to comprehend and it generally results in a few interested or few upbeat counts of the model.

3.4.8 Classifier Result:

For Predicting the class of given data points classification method is being used. In the flowchart find out FP rate, TP, rate, F-Measures, precision, recall, ROC area, PRC area, and confusion Matrix. Classification associated the class of supervised learning here the destination also shifted with input data. It is the most significant portion to appreciate the classifier to witness its pertinency after training the methods. The quantity of the topicality is being used as precision and recall. The recall is considered as True positive. For testing false positive is the quantity of accuracy. To evaluate the information recovery system used F-Measure contains

$(2 * \text{precision} * \text{recall}) / (\text{precision} + \text{recall})$. Recover Operating Characteristic Curve contains a section of true positive rate with false-positive rate. Precision-Recall Curves are conscious of the idea of True Positive, True Negative, False Positive, False Negative. Evaluating this execution of classification confusion matrix is used that's a table to calculate the execution of a classification.

3.4.9 Best Model:

Best Model contains the model which is the best among all the models to predict anything from a paper perspective. In the proposed paper five algorithms are used and the best algorithm is Bagging for detecting accuracy.

3.5 Data visualization:

In this section, we attach Recursive Feature Elimination because Recursive Feature Elimination is successful in determining those features (graphs) in a training dataset that are more or perhaps more effective in measuring the target variable. Besides fewer features can make machine learning algorithms run more smoothly (less resource or training time) and be more powerful.

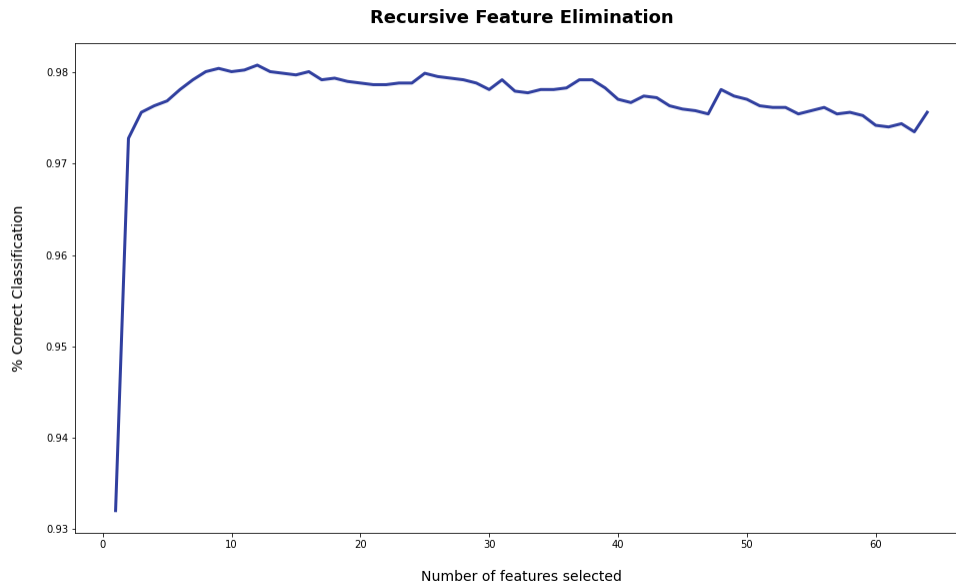


Fig 2: Recursive Feature Elimination

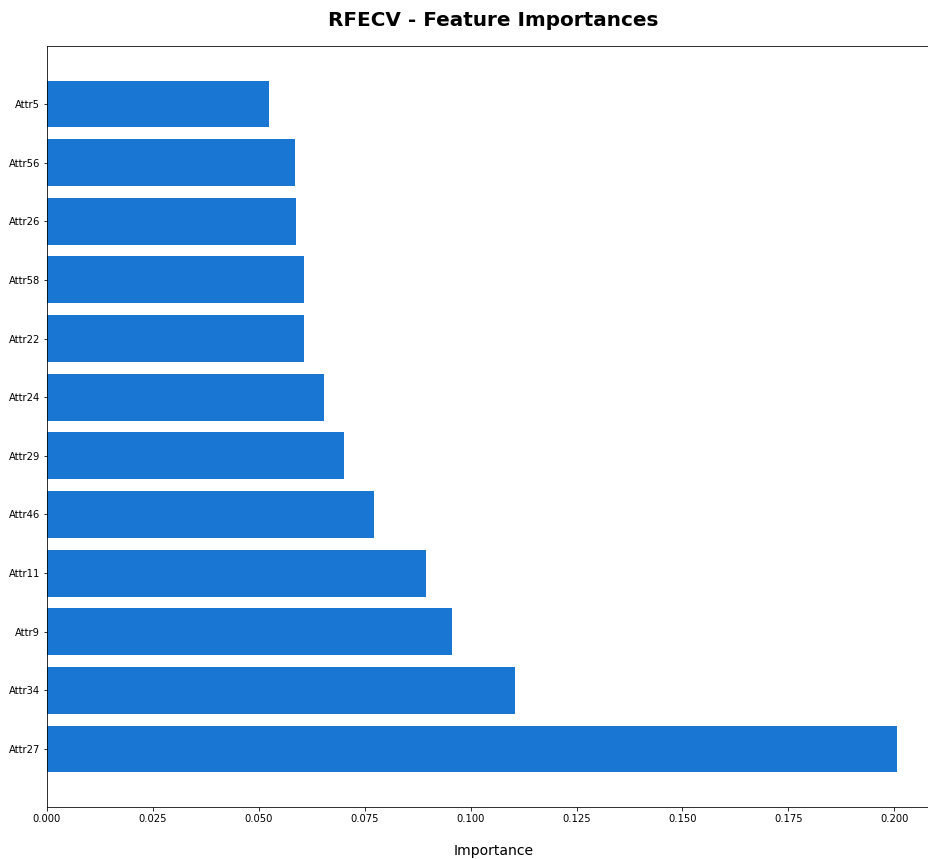


Fig 3: RFECV- Feature Importance

Applying Recursive feature elimination (RFE) we gain 12 importance attributes Such as Attr5, Attr9, Attr11, Attr22, Attr24, Attr26, Attr27, Attr29, Attr46, Attr56, Attr58 in our data set which are included for feature selection and its show fig:3. We also add curves to our dataset by applying Recursive feature elimination (RFE) which shows in fig:2.

CHAPTER 4

RESULTS AND DISCUSSION

In the result section Table- 1.1, 1.2, 1.3, 1.4, 1.5 reflects the various output values of all classification techniques determined on the basis of different measurements. Here five classifiers are used for predicting bankruptcy to see the output of the respective classifiers on the basis of TP Rate, FP Rate, Precision, F-Measure, Recall, ROC, and PRC values. In our study, we use k-fold cross-validation(k=10) to measure our accuracy.

Table-2 displays the confusion matrix for five years of predicting bankruptcy using five machine learning algorithms such as Adaboost, Decision tree, J48, Bagging, and Random Forest.

Table-3 we mention the accuracy of five years and find the best accuracy rate of 97%. So by analyzing this table we see that the best classifier is Bagging which gives the maximum of the best accuracy than others. In 1st year it gives 97% accuracy, in 2nd year it 96%, in 3rd year 95%, in 4th year 96% and in 5th year it gives 96% accuracy. We also see that the 1st-year accuracy of other classifiers is 97% but the average bagging accuracy is higher than other classifiers. So bagging is the best model in this paper.

Table:4 indicates that there have been a lot of current methods used to estimate Bankruptcy by Machine Learning classifiers and also other methods. So here are a few province approaches and their accuracy. They are compared to our proposed model, and it is noted that The proposed analysis obtained an accuracy of 97%.

Evaluation Outcome Machine Learning Algorithm to Measure Bankruptcy

Table - 1.1: Classification result analysis for 1st year

Model name	Label	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Adaboost	No bankrupt	0.99	0.88	0.97	0.99	0.98	0.85	0.99
	bankrupt	0.13	0.01	0.50	0.13	0.20	0.85	0.26
Decision Tree	No bankrupt	0.99	0.63	0.98	0.99	0.99	0.88	0.99
	Bankrupt	0.36	0.00	0.75	0.38	0.49	0.86	0.47
J48	No bankrupt	0.99	0.50	0.98	0.99	0.99	0.80	0.98
	Bankrupt	0.49	0.01	0.77	0.49	0.60	0.80	0.43
Bagging	No bankrupt	0.99	0.52	0.96	0.98	0.98	0.92	0.99
	Bankrupt	0.48	0.00	0.82	0.94	0.63	0.92	0.66
Random Forest	No bankrupt	0.99	0.72	0.97	0.99	0.99	0.91	0.99
	Bankrupt	0.28	0.00	0.89	0.28	0.43	0.91	0.63

Table-1.2: Classification result analysis for 2nd year

Model name	Label	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Adaboost	No	0.99	0.92	0.96	0.99	0.98	0.79	0.98
	bankrupt bankrupt	0.08	0.08	0.57	0.08	0.14	0.79	0.23
Decision tree	No	1.00	1.00	0.96	1.00	0.98	0.50	0.96
	bankrupt bankrupt	0.00	0.00	0.00	0.00	0.00	0.50	0.04
J48	No	0.99	0.55	0.98	0.99	0.99	0.79	0.98
	bankrupt bankrupt	0.44	0.01	0.73	0.45	0.55	0.79	0.38
Bagging	No	0.99	0.64	0.97	0.99	0.98	0.88	0.99
	bankrupt bankrupt	0.35	0.00	0.96	0.36	0.52	0.88	0.62
Random Forest	No	0.99	0.82	0.97	0.99	0.98	0.88	0.99
	bankrupt bankrupt	0.19	0.00	0.85	0.19	0.30	0.88	0.50

Table-1.3: Classification result analysis for 3rd year:

Model name	Label	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Adaboost	No bankrupt	0.99	0.99	0.96	0.99	0.98	0.82	0.99
	bankrupt	0.05	0.00	0.58	0.05	0.09	0.82	0.27
Decision tree	No bankrupt	1.00	0.97	0.95	1.00	0.98	0.59	0.96
	bankrupt	0.03	0.00	0.75	0.03	0.06	0.59	0.10
J48	No bankrupt	0.99	0.59	0.97	0.99	0.98	0.80	0.98
	bankrupt	0.41	0.01	0.64	0.41	0.50	0.80	0.39
Bagging	No bankrupt	0.99	0.69	0.98	0.99	0.98	0.90	0.99
	bankrupt	0.31	0.00	0.89	0.31	0.46	0.90	0.57
Random Forest	No bankrupt	0.99	0.85	0.96	0.99	0.98	0.89	0.99
	bankrupt	0.15	0.00	0.78	0.15	0.26	0.89	0.47

Table-1.4: Classification result analysis for 4th year

Model name	Label	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Adaboost	No	0.99	0.92	0.95	0.99	0.97	0.80	0.98
	bankrupt bankrupt	0.07	0.00	0.63	0.07	0.13	0.80	0.27
Decision tree	No	0.99	0.73	0.96	0.99	0.98	0.81	0.98
	bankrupt bankrupt	0.27	0.01	0.75	0.27	0.39	0.81	0.38
J48	No	0.99	0.59	0.97	0.99	0.98	0.79	0.98
	bankrupt bankrupt	0.41	0.01	0.65	0.41	0.51	0.79	0.40
Bagging	No	0.99	0.95	0.95	0.99	0.97	0.82	0.99
	bankrupt bankrupt	0.05	0.00	0.81	0.05	0.09	0.82	0.29
Random Forest	No	0.99	0.87	0.95	0.99	0.98	0.89	0.99
	bankrupt bankrupt	0.13	0.00	0.73	0.13	0.23	0.89	0.45

Table-1.5: Classification result analysis for 5th year

Model name	Label	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	PRC Area
Adaboost	No bankrupt	0.99	0.79	0.95	0.99	0.97	0.86	0.98
	bankrupt	0.22	0.00	0.88	0.22	0.35	0.86	0.52
Decision tree	No bankrupt	0.99	0.65	0.95	0.99	0.97	0.89	0.99
	bankrupt	0.35	0.01	0.81	0.35	0.49	0.89	0.59
J48	No bankrupt	0.98	0.39	0.97	0.98	0.98	0.82	0.97
	bankrupt	0.61	0.02	0.72	0.61	0.65	0.82	0.57
Bagging	No bankrupt	0.99	0.44	0.97	0.99	0.98	0.94	0.99
	bankrupt	0.56	0.01	0.86	0.57	0.68	0.94	0.77
Random Forest	No bankrupt	0.99	0.64	0.95	0.99	0.97	0.93	0.99
	bankrupt	0.36	0.01	0.80	0.36	0.50	0.93	0.64

Table-2: Confusion metrics for MLT bankruptcy prediction

Model name	Label	1st year PP	1st year PN	2nd year PP	2nd year PN	3rd year PP	3rd year PN	4th year PP	4th year PN	5th year PP	5th year PN
Adaboost	No Bankrupt	6723	33	9749	24	9991	17	9255	22	5488	12
	bankrupt	237	34	368	32	471	24	477	38	322	88
Decision tree	No Bankrupt	6722	34	9773	0	10004	4	9232	45	5467	33
	bankrupt	171	100	400	0	480	15	377	138	267	143
J48	No Bankrupt	6715	41	9707	66	9892	116	9164	113	5403	97
	bankrupt	136	135	221	179	291	204	303	212	161	249
Bagging	No Bankrupt	6747	9	9767	6	9988	20	9271	6	5463	37
	bankrupt	140	131	257	143	340	155	356	159	182	228
Random Forest	No Bankrupt	6747	9	9760	13	9986	22	9252	25	5463	37
	bankrupt	294	77	326	74	419	76	446	69	261	149

Table-3: Accuracy from different kinds of algorithms in five years

Model name	Accuracy (1st year)	Accuracy (2nd year)	Accuracy (3rd year)	Accuracy (4th year)	Accuracy (5th year)
Adaboost	96%	96%	95%	94%	94%
Decision tree	97%	96%	95%	95%	94%
J48	97%	96%	95%	95%	95%
Bagging	97%	97%	95%	96%	96%
RandomForest	97%	96%	95%	95%	94%

Table-4: Proposed model and Accuracy

Ref.	Method	Accuracy
Zelenkov et al.[7]	TSCM	0.934%
Devi et al. [10]	SVM	95%
Alaminos et al. [15]	Logistic Regression	Not assigned
Martin et al. [17]	Ant-miner algorithm	Not assigned
Chen et al. [18]	Bosting	Not assigned
Shumway [19]	Hazard Model	Not assigned
Kim et al [14]	GA	Not assigned
Proposed method	Bagging	97%

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion:

The execution of the several techniques used for Predicting bankruptcy. By rating, the technique was calculated under several acting metrics like accuracy, specificity precision, and sensitivity that compared with machine learning and other traditional statistical analysis. To raise the accurate bankruptcy prediction machine learning techniques have been used effectively. Accuracy is one of the extreme perfection ways due to its great economic impact. In this paper, we find out accuracy among 5 years. We find out the highest accuracy in bagging that obtains 97%. Bagging is a popular ensemble method that has been applied with massive victory to several machine learning problems. In this paper, Bagging is finding out the best classifier.

5.2 Future work:

This research shows that the bankruptcy report is useful in recognizing the consequences of bankruptcy. So, we decided that we will work on it on a large scale. For example, we will work with the garments sector. We will use different types of methods for better and perfect results. We will work on deep learning so that it can process data and make decisions by following the workings of the human brain. Develop an application that uses artificial intelligence to indicate when a model should be used to predict bankruptcy in future research. It will also work on how to improve the eCommerce marketplace.

CONTRIBUTION FROM THIS R&D PROJECT

International publication: Scopus indexed

Comparison of Different Machine Learning Algorithms for Detecting Bankruptcy: Maria Sultana Keya, Himu Akter, Md. Atiqur Rahman, Md. Mahbobur Rahman, Amit Chakroborty Chhoton, Md. Sabab Zulfiker : Accepted in 6th International Conference on Inventive Computation Technologies ICICT 2021, India from 20th- 22nd January 2021.

REFERENCES

[1] Wikipedia contributors. (2020, November 18). Bankruptcy. In Wikipedia, The Free Encyclopedia. Retrieved 20:10, November 21, 2020, from <https://en.wikipedia.org/w/index.php?title=Bankruptcy&oldid=989417913>

[2] Wikipedia contributors. (2020, November 18). Bankruptcy. In Wikipedia, The Free Encyclopedia. Retrieved 20:10, November 21, 2020, from <https://en.wikipedia.org/w/index.php?title=Bankruptcy&oldid=989417913>

[3] Wikipedia contributors. (2020, November 5). White-collar crime. In Wikipedia, The Free Encyclopedia. Retrieved 20:12, November 21, 2020, from https://en.wikipedia.org/w/index.php?title=White-collar_crime&oldid=987266832

[4] Wikipedia contributors. (2020, August 16). Corporations Act 2001. In Wikipedia, The Free Encyclopedia. Retrieved 20:15, November 21, 2020, from https://en.wikipedia.org/w/index.php?title=Corporations_Act_2001&oldid=973387889

[5] Office of the Superintendent of Bankruptcy, available at <http://www.ic.gc.ca/eic/site/bsf-osb.nsf/eng/br02805.html>, last accessed 8.20, January 2020.

[6] Wikipedia contributors. (2020, November 7). Insolvency and Bankruptcy Code, 2016. In Wikipedia, The Free Encyclopedia. Retrieved 20:22, November 21, 2020, from https://en.wikipedia.org/w/index.php?title=Insolvency_and_Bankruptcy_Code,_2016&oldid=987483654.

- [7] Zelenkov, Y., Fedorova, E., & Chekrizov, D. (2017). Two-step classification method based on genetic algorithm for bankruptcy forecasting. *Expert Systems with Applications*, 88, 393-401.
- [8] Loukeris, N., & Matsatsinis, N. (2006). Corporate financial evaluation and bankruptcy prediction implementing artificial intelligence methods. *WSEAS Transactions on Business and Economics*, 3(4), 343.
- [9] Loukeris, N., & Matsatsinis, N. (2006). Corporate financial evaluation and bankruptcy prediction implementing artificial intelligence methods. *WSEAS Transactions on Business and Economics*, 3(4), 343.
- [10] Tomczak, S. (2014). Comparative analysis of liquidity ratios of bankrupt manufacturing companies. *Business and Economic Horizons*, 10(3), 151-164.
- [11] Devi, S. S., & Radhika, Y. (2018). A survey on machine learning and statistical techniques in bankruptcy prediction. *International Journal of Machine Learning and Computing*, 8(2), 133-139.
- [12] Fernández, M. Á., Laguillo, G., del Castillo, A., & Becerra, R. (2019). Focused vs unfocused models for bankruptcy prediction: Empirical evidence for Spain. *Contaduría y Administración*, 64(2), e96
- [13] Chen, M. Y. (2011). Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. *Computers & Mathematics with Applications*, 62(12), 4514-4524.
- [14] Erdogan, B. E. (2013). Prediction of bankruptcy using support vector machines: an application to bank bankruptcy. *Journal of Statistical Computation and Simulation*, 83(8), 1543-1555.
- [15] Kim, M. J., & Han, I. (2003). The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems with Applications*, 25(4), 637-646.
- [16] Alaminos, D., del Castillo, A., & Fernández, M. Á. (2016). A global model for bankruptcy prediction. *PloS one*, 11(11), e0166693.
- [17] Laitinen, E. K., & Laitinen, T. (2000). Bankruptcy prediction: Application of the Taylor's expansion in logistic regression. *International review of financial analysis*, 9(4), 327-349.
- [18] Martin, A., Lakshmi, T. M., & Venkatesan, V. P. (2014). An analysis on qualitative bankruptcy prediction rules using ant-miner. *IJ Intelligent System and Applications*, 1, 36-44.

- [19] Chen, T., He, T., Benesty, M., Khotilovich, V., & Tang, Y. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1-4.
- [20] Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *The journal of business*, 74(1), 101-124.
- [21] Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert systems with applications*, 58, 93-101.
- [22] Martin, A., Aswathy, V., Balaji, S., Lakshmi, T. M., & Venkatesan, V. P. (2012, March). An analysis on qualitative bankruptcy prediction using fuzzy ID3 and ant colony optimization algorithm. In *International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME-2012)* (pp. 416-421). IEEE.

APPENDICES

Abbreviation

AdaBoost = Adaptive Boosting

KNN = k-nearest neighbors

SVM = Support-Vector Machines

RFE = Recursive Feature Elimination

SMOTE = Synthetic Minority Over-sampling Technique

RFC = Request For Comments

DTC = Decision Tree Classifier

Appendix: Research Reflections

At the beginning of this research work, we had very little experience with machine learning algorithms about detection and prediction. Our supervisor was very kind and helpful to us. He gave us valuable guidance and helped us a lot. In this working time of research, we learned much new information, new techniques, use of new algorithms, and worked with different methods. At the time when we start working with this, we face many problems but gradually we become more and more familiar with these algorithms.

Finally, by doing the research we have gained courage and been inspired to do more research in the future.

PLAGIARISM REPORT

Bankruptcy report

ORIGINALITY REPORT

12%

SIMILARITY INDEX

11%

INTERNET SOURCES

3%

PUBLICATIONS

9%

STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	5%
2	Submitted to St. Petersburg High School Student Paper	1%
3	hdl.handle.net Internet Source	1%
4	www.slideshare.net Internet Source	1%
5	en.wikipedia.org Internet Source	<1%
6	www.coursehero.com Internet Source	<1%
7	export.arxiv.org Internet Source	<1%
8	dspace.library.uvic.ca Internet Source	<1%
9	dro.deakin.edu.au Internet Source	<1%

10	Submitted to ESCP-EAP Student Paper	<1 %
11	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1 %
12	A. Martin, T. Miranda Lakshmi, V. Prasanna Venkatesan. "An Analysis on Qualitative Bankruptcy Prediction Rules using Ant-Miner", International Journal of Intelligent Systems and Applications, 2014 Publication	<1 %
13	utpedia.utp.edu.my Internet Source	<1 %
14	archive.ics.uci.edu Internet Source	<1 %
15	academic.oup.com Internet Source	<1 %
16	www.rug.nl Internet Source	<1 %
17	www.ijser.org Internet Source	<1 %
18	ethesisarchive.library.tu.ac.th Internet Source	<1 %
19	bspaace.buid.ac.ae Internet Source	<1 %
20	Błażej Prusak. "Review of Research into Enterprise Bankruptcy Prediction in Selected Central and Eastern European Countries", International Journal of Financial Studies, 2018 Publication	<1 %
21	skemman.is Internet Source	<1 %
22	www.criminaldefensene.com Internet Source	<1 %

Exclude quotes On Exclude matches < 10 words
Exclude bibliography On