iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** V **Month of publication:** May 2022

DOI: https://doi.org/10.22214/ijraset.2022.43665

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com

Bank Loan Approval Prediction Using Data Science Technique (ML)

Subhiksha R[1], Vaishnavi L[2], Shalini B[3], Mr. N. Manikandan[4]
[4]*Assistant Professor, Department of Computer Science and Engineering, Agni College of Technology, Chennai, Tamil Nadu, India*

Abstract: ***Banks are making major part of profits through loans. Loan approval is a very important process for banking organizations. It is very difficult to predict the possibility of payment of loan by the customers because there is an increasing rate of loan defaults and the banking authorities are finding it more difficult to correctly access loan requests and tackle the risks of people defaulting on loans. In the recent years, many researchers have worked on prediction of loan approval systems. Machine learning technique is very useful in predicting outcomes for large amount of data. In this paper, four algorithms are used such as Random Forest algorithm, Decision Tree algorithm, Naive Bayes algorithm, Logistic Regression algorithm to predict the loan approval of customers. All the four algorithms are going to be used on the same dataset and going to find the algorithm with maximum accuracy to deploy the model. Henceforth, we develop bank loan prediction system using machine learning techniques, so that the system automatically selects the eligible candidates to approve the loan.***
Keywords: Loan approval, Loan Default, Random Forest algorithm, Decision Tree algorithm, Naive Bayes algorithm, Logistic Regression algorithm, Loan prediction, Machine learning.

I. INTRODUCTION

A loan is the major source of income for the banking sector of financial risk for banks. Large portions of a bank's assets directly come from the interest earned on loans given. The activity of lending loans carry great risks including the inability of borrower to pay back the loan by the stipulated time. It is referred as "credit risk". A candidate's worthiness for loan approval or rejection was based on a numerical score called "credit score". Therefore, the goal of this paper is to discuss the application of different Machine Learning approach which accurately identifies whom to lend loan to and help banks identify the loan defaulters for much-reduced credit risk.

II. LITERATURE SURVEY

TITLE 1: Improving Information Quality in Loan Approval Processes for Fair Lending and Fair Pricing
AUTHOR: M. Cary Collins
YEAR: 2013
DESCRIPTION: Bank data management on loan approval processes has great room for improvements of information quality and data problems prevention especially with regards to fair lending and fair pricing practices. They first reviewed briefly typical data collection protocols deployed at many financial institutions for loan approval and loan pricing. Federal regulations mandate portions of these data protocols. While discussing the data capture and analysis for fair lending, they illustrated some initial key steps currently needed for improving information quality to all parties involved.

TITLE 2: Loan Credibility Prediction System Based on Decision Tree Algorithm
AUTHOR: Sivasree M S, Rekha Sunny T
YEAR: 2015
DESCRIPTION: Data mining techniques are becoming very popular nowadays because of the wide availability of huge quantity of data and the need for transforming such data into knowledge. Data mining techniques are implemented in various domains such as retail industry, biological data analysis, intrusion detection, telecommunication industry and other scientific applications. Techniques of data mining are also be used in the banking industry which help them compete in the market well equipped. In this paper, they introduced a prediction model for the bankers that will help them predict the credible customers who have applied for a loan. Decision Tree Algorithm is being applied to predict the attributes relevant for credibility. A prototype of the model has been described in this paper which can be used by the organizations for making the right decisions to approve or reject the loan request from the customers.

©IJRASET: All Rights are Reserved

TITLE 3: Loan Approval Prediction based on Machine Learning Approach
AUTHOR: Kumar Arun, Garg Ishan, Kaur Sanmeet
YEAR: 2016
DESCRIPTION: With the enhancement in the banking sector, lots of people apply for bank loans but the bank has its limited assets which it grants to only limited people , so finding out to whom the loan can be granted is a typical process for the banks. So, in this paper , they tried to reduce this risk by selecting the safe person so as to save lots of bank efforts and assets. It was done by mining the previous records of the people to whom the loan was granted before and on the basis of these records the machine was trained using the machine learning model which gave the most accurate result. The main goal of this paper is to predict if loan assignment to a specific person will be safe or not. This paper has into four sections (i) Collection of data (ii) Comparing the machine learning models on collected data (iii) Training the system on most promising model (iv) Testing the system.

III. EXISTING SYSTEM

Anomaly detection relies on individuals' behaviour profiling and works by detecting any deviation from the norm. When it is used for online banking fraud detection, it suffers from three disadvantages. First, for an individual, the historical behaviour data are often too limited for profiling his/her behaviour pattern. Second, because of the heterogeneous nature of transaction data, there is no uniform treatment to various attribute values, which will become a potential barrier for development of the model and for further usage. Third, the transaction data are highly skewed, and it becomes a challenge for utilizing the label information effectively. Anomaly detection often suffers from poor generalization ability and a very high false alarm rate. We argue that individuals' limited historical data for behaviour profiling and fraud data's highly skewed nature could account for this defect. Since it is straightforward to use information from other similar individuals, similarity measurement itself becomes a great challenge due to heterogeneous nature of attribute values.

A. Disadvantages

1) They had proposed a mathematical model and machine learning algorithms were not used.
2) Class Imbalance problem was not addressed and the proper measure were not taken.

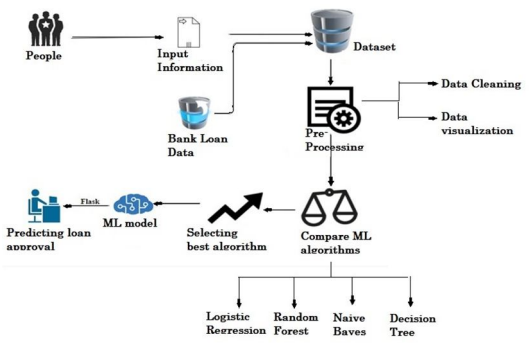
B. Proposed System

In our proposed system, we combine datasets from different sources to form a generalized dataset and use four machine learning algorithms such as Random forest, Logistic regression, Decision tree and Naive bayes algorithm on the same dataset .The dataset we collected for predicting given data is split into training set and test set in the ratio of 7:3. The data model which was created using Machine learning algorithms are applied on training set and based on maximum test result from the four algorithms, the test set prediction is done using the algorithm that has maximum performance. After that, we deploy the model using Flask Framework.

C. Advantages

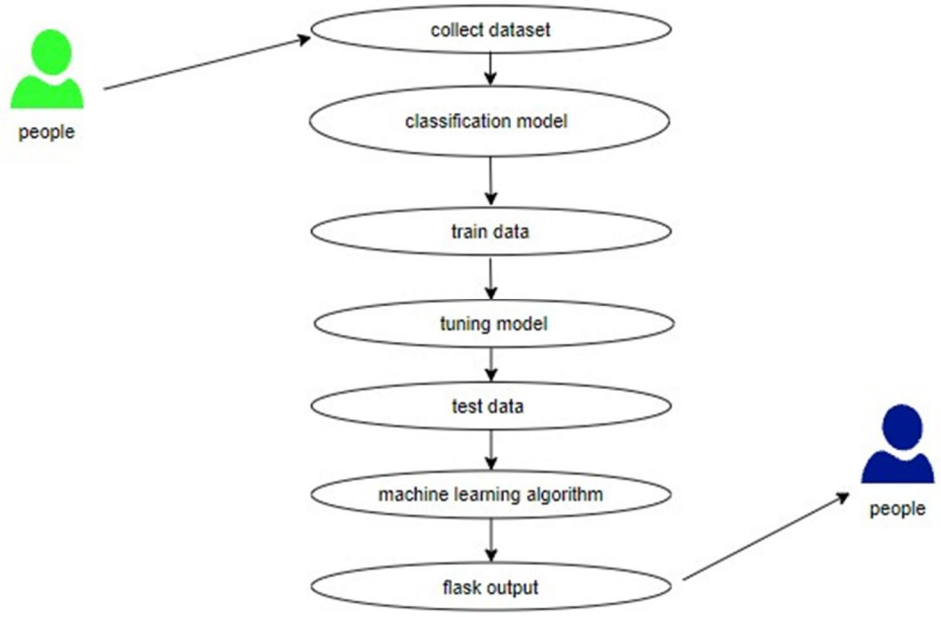
1) Performance and accuracy of the algorithms can be calculated and compared.
2) Class imbalance can be dealt with machine learning approaches.

IV. SYSTEM ARCHITECTURE



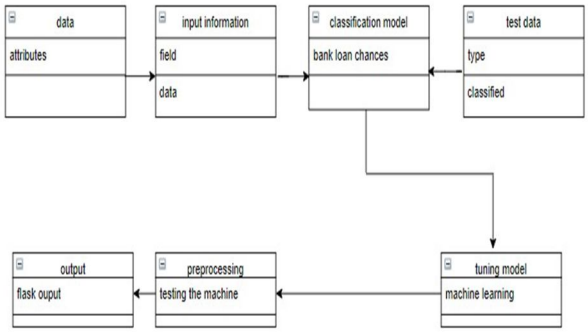
The dataset is obtained by gathering lot of required datasets and combining them to produce a generalised dataset. The dataset thus produced is pre-processed i.e., the dataset is cleaned before doing data visualization. Then the four algorithms are applied on the same pre-processed dataset and calculated for the best performed algorithms among them. Then the best algorithm is used to train the model and test it to check how accurate the algorithm can predict the output. Then we deploy that model to predict if bank loan can be approved or not for a specific candidate.

V. USE CASE DIAGRAM



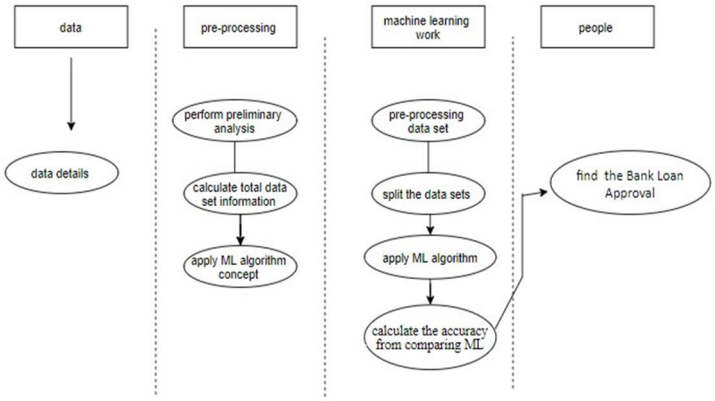
Use case diagrams are used for high level requirement analysis of a system. So, when analysing the requirements of a system, the functionalities are captured in use cases. So, uses cases are nothing, but the functionalities of the system written in an organized manner.

A. Class Diagram



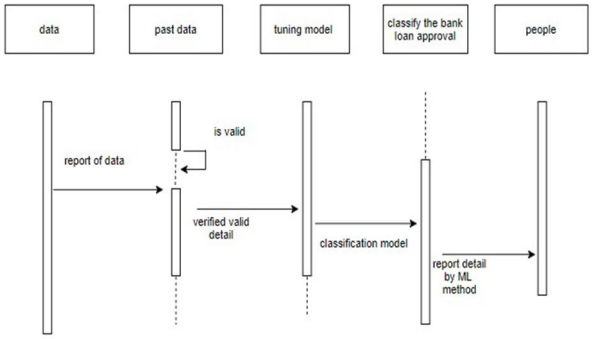
Class diagram is generally a graphical representation of the static view of the system and represents different aspects of the application. A collection of class diagrams will represent the whole system.

B. Activity Diadram



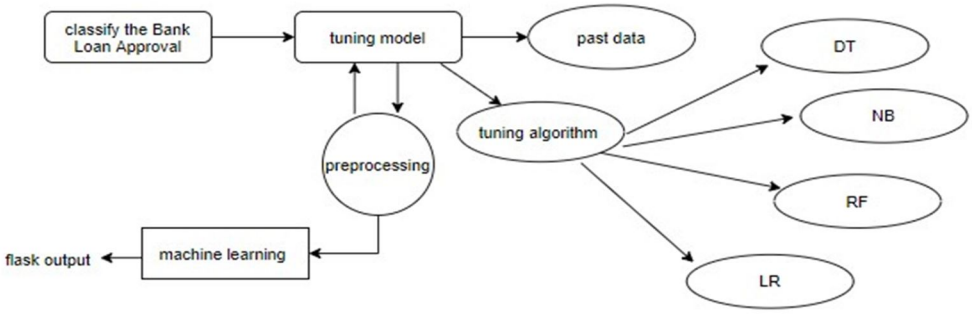
Activity diagrams not only visualize the dynamic nature of a system, but they are also used for constructing executable system by using forward and reverse engineering techniques. Activity diagram is some time considered as the flow chart, but it is not.

C. Sequence Diagram



Sequence diagrams model the flow of logic within our system in a visual manner, enabling both to document and validate our logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modelling, which focuses on identifying the behaviour within the system.

D. Entity Relationship Diagram



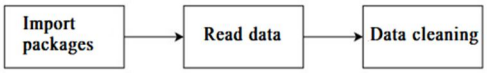
An entity relationship diagram (ERD) is a graphical representation of an information system that depicts the relationships among people, objects, places, concepts or events within that system. An Entity relationship model is a data modelling technique that helps define business processes and can be used as the foundation for a relational database.

VI. LIST OF MODULES

A. Data Pre-processing
B. Data Analysis of Visualization
C. Comparing Algorithms
D. Deployment Using Flask

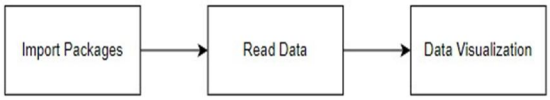
VII. DATA PRE-PROCESSING

Data Pre-processing is a technique used for converting the raw data into a clean data set. Whenever a data is gathered from different sources, it is collected in raw format which is not feasible for the analysis of the model. For achieving better results, the model in Machine Learning method of the data has to be in a proper manner. Some of the specific Machine Learning model needs information to be in a particular format, for example, Random Forest algorithm don't support null values. Therefore, to proceed further with random forest algorithm, null values need to be managed from the original raw dataset. In data pre-processing, we carry out data cleaning task with the help of Python's Pandas library. Data cleaning is a process of removing missing, incomplete or duplicate data. The steps and techniques used for data cleaning will vary from each dataset. When combining multiple datasets, there are so many possibilities for the data to be duplicated or mislabeled or incomplete. The main objective of data cleaning is to detect and remove errors to increase the value of data in analytics and decision making to get accurate outputs.



VIII. DATA ANALYSIS OF VISUALIZATION

Data visualization is a technique helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupted data, outliers, and much more. It can be used to express and demonstrate main relationships between plots and charts that are more visceral. Sometimes data does not make sense until we can look at in a visual form, such as with charts, graphs and plots. Being able to quickly visualize the data samples is an important skill both in applied statistics and in applied machine learning. It can also be used to remove outliers to get better and accurate results. It is implemented by using Python's Matplotlib library.



IX. COMPARING ALGORITHMS

Before comparing algorithms, we build a Machine Learning Model using Python's Scikit-Learn libraries. In this library package, we must do pre-processing, linear model with logistic regression method, cross validating by K-Fold method, ensemble with random forest method and tree with decision tree classifier. Additionally, we split the train set and test set in order to predict the result by comparing accuracy.
To find the algorithm with best performance, we use the following performance metrics:

A. Confusion Matrix

Confusion matrix is one of the performance metrics used to find the correctness and accuracy of the model. It has the following four parameters:

B. False Positives (FP)

A person who will pay is predicted as defaulter where actual class is no, and predicted class is yes. E.g., if actual class says a person is not married but predicted class tells that the person is married.

C. False Negatives (FN)

A person who will pay is predicted as defaulter where actual class is yes but predicted class is no. E.g., if actual class value indicates that the person is married, and predicted class tells that person is not married.

D. True Positives (TP)

A person who will not pay is predicted as defaulter where the value of actual class is yes, and the value of predicted class is also yes. E.g., if actual class value indicates that the person is married, and predicted class also tells the same thing.

E. True Negatives (TN)

A person who pays is predicted as defaulter where the value of actual class is no, and value of predicted class is also no. E.g., if actual class says the person is not married and predicted class also tells the same thing.

True Positive Rate (TPR) = True Positives / (True Positives + False Negatives)

False Positive Rate (FPR) = False Positives / (False Positives + True Negatives)

F. Accuracy

Accuracy is the most important performance metrics which is the ratio of observations that are correctly predicted to the total observations. Higher accuracy means that the model produces accurate results but only when we have symmetric datasets where values of false positive and false negatives are almost the same.

Accuracy = (TP+TN) / (FP+FN+ TN+TP)

G. Precision

Precision is the ratio of positive observations correctly predicted to the positive observations totally predicted. High precision rates relates to the low false positive rates of the dataset. We have got 0.876 precision which is really good.

Precision = TP / (FP+TP)

H. Recall

Recall is the ratio of correctly predicted positive observations to the all observations in actual class – yes i.e., the proportion of positively observed values correctly predicted which is nothing but the proportion of actual defaulters that the model will correctly predict.

Recall = TP / (FN +TP)

I. F1 Score

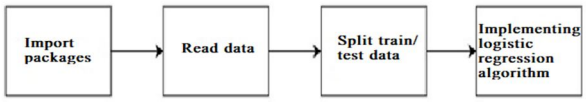
F1 Score is basically the average weight of Precision and Recall. Therefore, the F1 score takes the values of both false positives and false negatives into consideration. F1 score can be found out if there is an uneven class distribution. If the values of false positives and false negatives are too different, it's better to have a look at both Precision and Recall.

F1 Score = {(Precision * Recall) * 2} / (Precision + Recall)

X. LOGISTIC REGRESSION

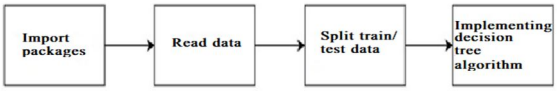
Logistic regression is a machine learning classification supervised algorithm that is used for predicting the probability of a categorical dependent variable. It is a statistical method that is used for analysing a dataset where there are one or more independent variables which determines an outcome. The outcome is measured with a dichotomous variable (which means there are only two possible outcomes). The primary objective of logistic regression is to find the best fitting model for describing the relationship between dependent variables and a set of independent variables. In logistic regression, the dependent variable is binary that contains data which is coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue V May 2022- Available at www.ijraset.com



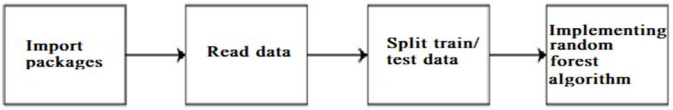
XI. DECISION TREE CLASSIFIER

Decision tree is used for building classification models in the form of a tree structure. It basically breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches which ends in another decision node or a leaf node, and a leaf node represents the final decision. The topmost decision node in a tree called root node corresponds to the best predictor. Decision trees are used for handling both categorical and numerical data. It uses an if-then rule set which is mutually exclusive and exhaustive for classification of datasets. The rules are being learnt sequentially using the training dataset one at a time. Every time a rule is learned, the tuples which are covered by the rules are being removed. This process is continued on the training set until it reaches a terminating condition. It is constructed in a top-down recursive divide-and-conquer manner and all attributes in the dataset should be categorical. Or else, they should be discretized in advance. Attributes in the top of the tree have more impact towards classification which are identified using the information gain concept.



XII. RANDOM FOREST ALGORITHM

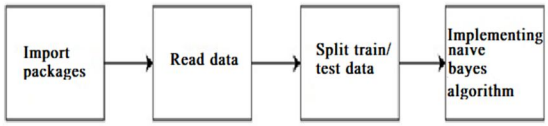
Random forest is one of the supervised machine learning algorithm which is based on ensemble learning method. Ensemble learning is a type of learning method where we can join different types of algorithms or join same algorithm multiple times to form a very powerful prediction model. It operates by constructing a multitude of decision trees at training time and outputting the class i.e., the mode of the classes of the individual trees. Random decision forests correct the decision trees' habit of over fitting their training set. In classification problems, each tree in the forest is used for predicting the category to which the new record belongs. Finally, the new record is being assigned to the category that wins the majority of the votes.



XIII. NAÏVE BAYES CLASSIFIER

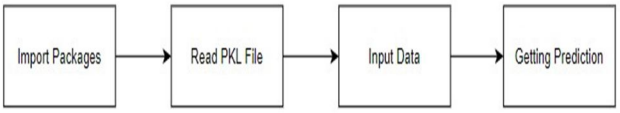
Naive Bayes is a supervised machine learning statistical classification technique based on Bayes Theorem. It is one of the most simplest supervised machine learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm among other algorithms. Naive Bayes classifiers have very high accuracy and speed on very large datasets. Naive Bayes classifier assumes that the effect of a specific feature in a class of a dataset is always independent of other features. For example, a loan applicant gets loan or not depending on his/her income, previous loan and transaction history, age, and location. Even if the above features are interdependent, these features are still considered independent. This assumption simplifies computation which is the reason why it is considered as naive. The above assumption is called class conditional independence. The probability of a class value when the value of an attribute is given is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value of a dataset, we have a probability of a data instance belonging to that specific class. To make a prediction, we need to calculate probabilities of the instance belonging to each class and select the class value that has the highest probability.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



XIV. DEPLOYMENT USING FLASK

After finding the algorithm that has the maximum performance among the four algorithms using performance metrics, we convert that model into a PKL file and then we deploy that model using FLASK framework in Python to build a user interface. So, when a customer enters his/her details in the user interface, and enters the submit button, he/she will get the result if a loan can be approved or not for that person.



XV. CONCLUSION

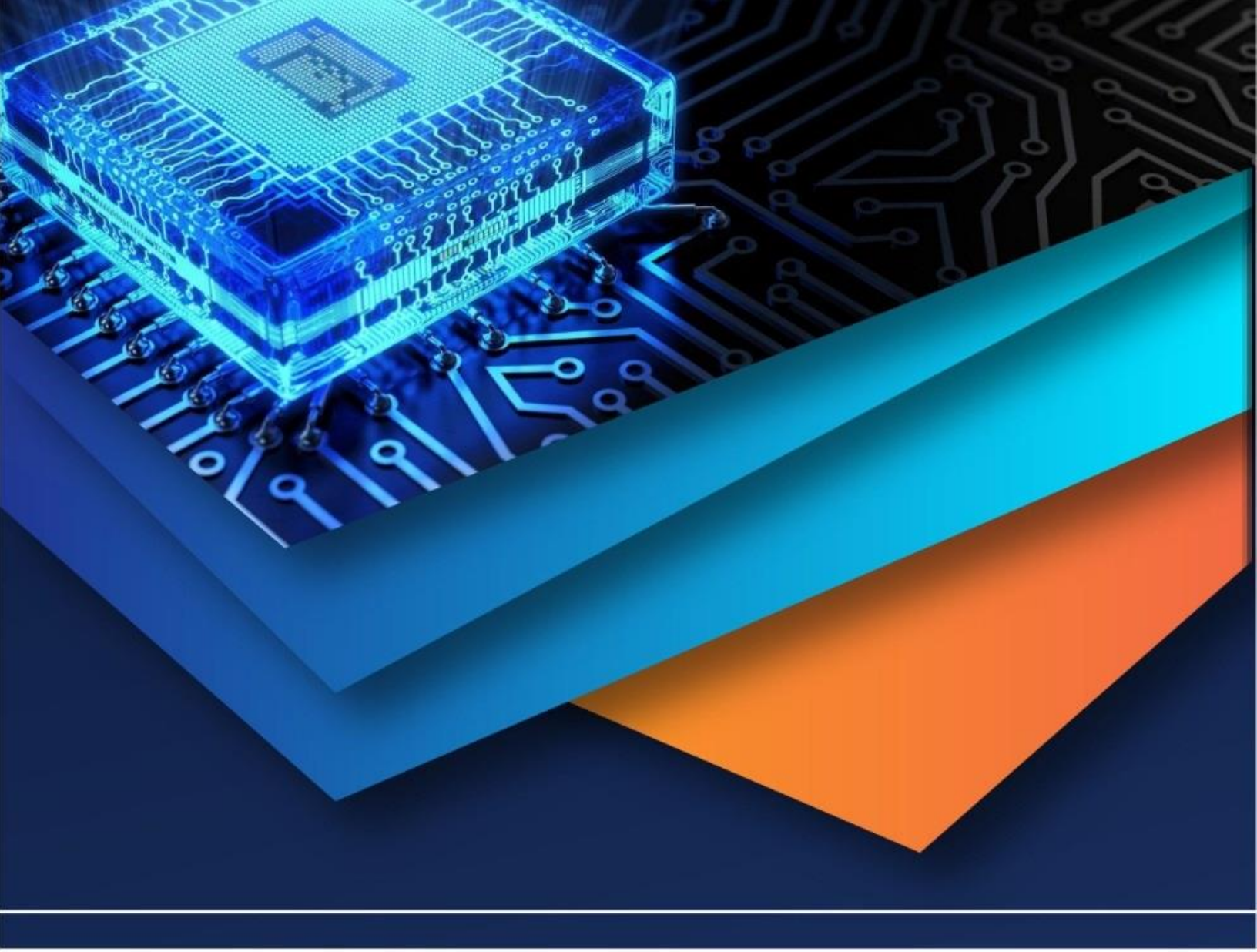
The analysis starts from data cleaning and processing missing value, exploratory analysis and finally model building and evaluation of the model. The best accuracy on public test set is when we get higher accuracy score and other performance metrics which will be found out. This paper can help to predict the approval of bank loan or not for a candidate.

XVI. FUTURE WORKS

We can make the Bank Loan Approval prediction to connect with Cloud for future use to optimize the work to implement in Artificial Intelligence environment.

REFERENCES

[1] Arun Kumar, Ishan Garg, and Sanmeer Kaur, "Loan Approval Prediction Using Machine Learning Approach," 2018.
[2] K. Hanumantha Rao, G. Srinivas, A. Damodhar, and M. Vikas Krishna at International Journal of Computer Science and Telecommunications published an article titled "Implementation of Anomaly Detection Technique Using Machine Learning Algorithms" (Volume2, Issue3, June 2011).
[3] G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial banks using machine learning classifier," International Conference on Intelligent Sustainable Systems (ICISS), 2017.
[4] "AzureML based analysis and prediction of loan applicants creditworthy," by Alshouiliy K, Alghamdi A, and Agrawal D P I n 2020, Third International conference on information and computer technologies.
[5] "Developing prediction model of loan risk in banks using data mining Machine Learning and Applications," Hamid A J and Ahmed T M, 2016.
[6] M. Li, A. Mickel, and S. Taylor "Should this loan be approved or denied?" published a paper in the Journal of Statistics Education in 2018.
[7] A. Vinayagamoorthy, M. Somasundaram, and C. Sankar, "Impact of Personal Loans Offered by Banks and Non-Banking Financial Companies in Coimbatore City," 2012.
[8] M. Cary Collins, Ph.D., and Frank M. Guess, Ph.D., MIT's Information Quality Conference, 2000, "Improving information quality in loan approval processes for fair lending and fair pricing."
[9] Arun Kumar, Ishan Garg, and Sanmeet Kaur, "Loan approval prediction based on machine learning approach," National Conference on Recent Trends in Computer Science and Information Technology, 2016.
[10] Sivasree M S and Rekha Sunny T, "Loan Credibility Prediction System Using Decision Tree Algorithm," International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 09, September-2015.
[11] Jiří Doležal, Jiří Šnajdr, Jaroslav Belás, Zuzana Vincúrová, "Model of the loan process in the context of unrealized income and loss prevention", Journal of International Studies, Vol. 8, No 1, 2015, pp. 91-106. DOI: 10.14254/2071-8330.2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)