iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** VIII **Month of publication:** Aug 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.55132

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 11 Issue VIII Aug 2023- Available at www.ijraset.com*

# Loan Eligibility Prediction Using Machine Learning

Mr. V. Sravan Kiran[1], B. Teja Reddy[2], D. Uday Kumar[3], K. Sai Avinash Varma[4], T. Sheshi Kiran[5]
*Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*

***Abstract:*** ***Lending research has become a highly essential research area since it may assist prevent loan defaults and grant loans to those who would pay on time. Therefore, for it though, we devised a technique for machine learning known as the random forest method, and also the data was used in this. Whatever is necessary is gathered from internet sites, and the data gathered is normalized before being employed for researching and predicting output, and it is then delivered to the random forest method, which is employed in our research. Following that, we may use the program to determine if a person is eligible for a loan or not, and a bank might not exclusively target the wealthy. Clients are accessed for loan purposes, but it also accesses other aspects of a client, that play a significant role in credit giving choices and lending prediction tax evaders.***
***Keywords: Loan approval, Random Forest algorithm, Pandas, Matplotlib.***

## I. INTRODUCTION

Loan Prediction is extremely beneficial to both bank workers and applicants. The goal of this Project is to give a quick, straightforward approach to choose qualified candidates. Housing Financing Corporation handles all types of loans. They are present in all urban, moderately, and rural locations. When that corporation or bank checks the consumer's qualifying for the loan, the user applies for it. A corporation or bank wishes to automating the loan qualification procedure based on information given by the consumer while completing out a registration form. Gender, Family Status, Occupation, Dependents by Number, Income, Loan Term, Credit Score and other facts are included. This project used data from past bank clients to create Loans were granted based on a set of criteria. Therefore, the machine learning model, which itself is based on the random forest method, is taught on that record to produce correct results. The primary goal of this study is to forecast lending safety, in which case the data is processed first and foremost to eliminate incorrect values in the information set so that it may be utilized to train the algorithm. Decisions may be made employing statistical and prospective approaches generated by different algorithms that use machine learning. The random forest and its theoretical formulation are explained in this essay. This study uses random forest as a technology study to develop predicting and probability techniques to a particular complaint of mortgage loan forecasting aid. Using random forest as a method, this study specifically decides whether a loan for a certain set of papers from an application would be accepted. Additionally, this machine learning game's novel elements have real-world applications.

## II. LITERATURE SURVEY

A benchmark factor is required in all commercial banking businesses to assess whether to grant a loan to an individual applicant. The judgement call criteria do not have to be confined to a single property, they might comprise any number of qualities that must be taken into account. Money lenders may supply datasets including the pertinent information for their consumers. This dataset's properties will be used to construct an algorithm that will assess if a loan should really be approved for a certain customer. There are two possibilities conceivable: adoption or refusal. The built model must reach conclusions quicker than desired. Computer science may help with prediction, judgement, and learning with data. It has its own flavor. Data is the most important thing in the world, that have triggered a renaissance in the discipline of computer science. Machine learning techniques have produced a wide range of data product based. To acquire data for this model, I studied several articles. The writers of the article aimed to reduce the efforts put forth by banks by constructing a model employing a range of algorithms to learn and outlining which of the techniques can be right. The four factors of the paper were data collection, assessment of various machine learning methods on the data, providing complete and testing. They used a mapper to forecast the entries. Writers were looking for reviews in the research. Credit score of new mortgages and application criteria are created utilizing the inductive decision tree technique. The credit score has an impact on loan approval. Researchers developed a model to check if loan licensing is safe and it was discovered that limited clients seem to be more likely to be approved for loans because they're more probable to repay them.

This sample was gathered using Kaggle. The authors sought to develop a model by analyzing new mortgage creditworthiness and application criteria using the injection of different prediction technique. The credit history has an impact on getting approved. They developed a model to predict whether It has been demonstrated that reduced clients are now more likely to be accepted for loans because they're more prone to repaying them. The data set was gathered using Kaggle. The article's authors sought to measure creditworthiness and calculate the Loan payback is likely. They employed the randomized forest method in the paper. They employed the decision-tree approach within the paper. A test set is used to certify the form. The scientists used data collection to develop a model in the study and the apparatus is made up of three components.

## III. OVERVIEW OF THE SYSTEM

### *A. Existing System*

The original attempt to the problem was quite rudimentary, utilizing the decision tree method. • The decision-tree approach provides lower protection and is incapable of checking and evaluating massive volumes of data.

When we use the randomized forest technique, we do not supply enormous quantities of information, which reduces the odds of having a machine offer additional data for more including among if it fails, reducing the reliability of the output. This may result in errors in the result, that cannot be tolerated.

Disadvantages of Existing System

The technique employed produces imprecise results. b. The precision is lowered.

c. Because the data provided is limited, there is a possibility of errors.

.

### *B. Proposed System*

The difficulty with the current system is the fact that the information provided is restricted, and it is unable to analyze vast amounts of data, thus the findings may be inaccurate. • To address this issue, we employ more computer algorithms, including the random forest method, that educates the system with greater efficacy than the decision tree approach since it is a set of choice trees.

Advantages of Proposed System

- ✓ Improved precision. b. The outcome is more exact than the present structure.
- ✓ c. Because they are quicker at collecting the data, it can be readily separated, and customers who make payments on time obtain the loan, lowering the lender's loss.

### *C. Proposed System Design*

In this project work, I used five modules and each module has own functions, such as:

1) *Dataset:* Submissions are used to collect large datasets. Following the collection of datasets, the types of data are further subdivided into two types of data sets, with the training data used to prepare the artificial intelligence algorithm and the experiment number of observations was using to evaluate the model's accuracy.
2) *Preprocessing:* The combined findings of the macroeconomic study for all statistical characteristics are as follows: 16 There might be a few outliers or exceptions that must be addressed before integrating the data into the model. In addition, the data collection has some null values. These spaces must be filled with data in order to ensure the random forest technique may be consistently applied.

These are some of the situations that might happen when training the database.

- Dealing with incomplete data
- Dealing with excessive values
- Data integrating into model: d. Forecasting and discovering the likelihood of certain scenarios

3) *Data Collection:* We may gather the user's information using a web browser constructed with front end languages like HTML and PHP, in order that the information can be physically obtained from the customer as needed.
4) *Data Analysis:* This allows us to determine if a person is eligible for a loan or not, reducing the workload for both bank employees and loan applicants. With remarkable precision.

## IV. ARCHITECTURE

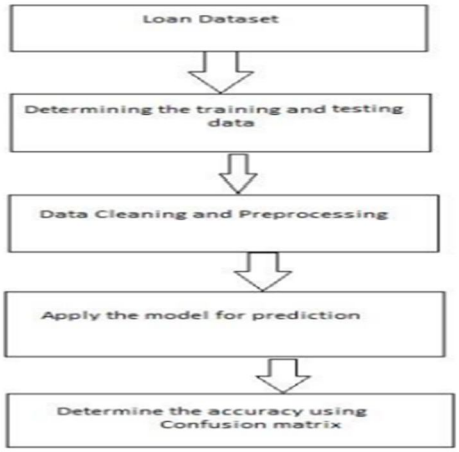



Fig 1: Frame work of Loan Prediction

## V. RESULTS SCREEN SHOTS

### A. *Dataset*

| Loan_ID | Gender | Married | Dependen | Education | Self_Empl | Applicantl | Coapplica | LoanAmou | Loan_Amo | Credit_His | Property_ | Loan_Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LP001002 | Male | No | 0 | Graduate | No | 5849 | 0 | | 360 | 1 | Urban | Y |
| LP001003 | Male | Yes | 1 | Graduate | No | 4583 | 1508 | 128 | 360 | 1 | Rural | N |
| LP001005 | Male | Yes | 0 | Graduate | Yes | 3000 | 0 | 66 | 360 | 1 | Urban | Y |
| LP001006 | Male | Yes | 0 | Not Gradu | No | 2583 | 2358 | 120 | 360 | 1 | Urban | Y |
| LP001008 | Male | No | 0 | Graduate | No | 6000 | 0 | 141 | 360 | 1 | Urban | Y |
| LP001011 | Male | Yes | 2 | Graduate | Yes | 5417 | 4196 | 267 | 360 | 1 | Urban | Y |
| LP001013 | Male | Yes | 0 | Not Gradu | No | 2333 | 1516 | 95 | 360 | 1 | Urban | Y |
| LP001014 | Male | Yes | 3+ | Graduate | No | 3036 | 2504 | 158 | 360 | 0 | Semiurban | N |
| LP001018 | Male | Yes | 2 | Graduate | No | 4006 | 1526 | 168 | 360 | 1 | Urban | Y |
| LP001020 | Male | Yes | 1 | Graduate | No | 12841 | 10968 | 349 | 360 | 1 | Semiurban | N |
| LP001024 | Male | Yes | 2 | Graduate | No | 3200 | 700 | 70 | 360 | 1 | Urban | Y |
| LP001027 | Male | Yes | 2 | Graduate | | 2500 | 1840 | 109 | 360 | 1 | Urban | Y |
| LP001028 | Male | Yes | 2 | Graduate | No | 3073 | 8106 | 200 | 360 | 1 | Urban | Y |
| LP001029 | Male | No | 0 | Graduate | No | 1853 | 2840 | 114 | 360 | 1 | Rural | N |
| LP001030 | Male | Yes | 2 | Graduate | No | 1299 | 1086 | 17 | 120 | 1 | Urban | Y |
| LP001032 | Male | No | 0 | Graduate | No | 4950 | 0 | 125 | 360 | 1 | Urban | Y |

This is a training dataset to train the model and to measure the accuracy of the model once it is trained

### B. *Testing Dataset*

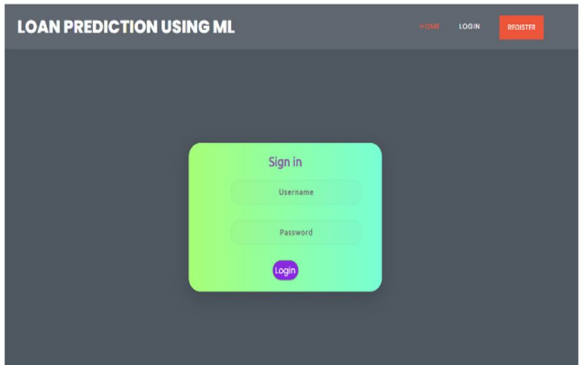
| Loan_ID | Gender | Married | Dependen | Education | Self_Empl | Applicantl | Coapplica | LoanAmou | Loan_Amo | Credit_His | Property_ | Loan_Status |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LP001002 | Male | No | 0 | Graduate | No | 5849 | 0 | | 360 | 1 | Urban | Y |
| LP001003 | Male | Yes | 1 | Graduate | No | 4583 | 1508 | 128 | 360 | 1 | Rural | N |
| LP001005 | Male | Yes | 0 | Graduate | Yes | 3000 | 0 | 66 | 360 | 1 | Urban | Y |
| LP001006 | Male | Yes | 0 | Not Gradu | No | 2583 | 2358 | 120 | 360 | 1 | Urban | Y |
| LP001008 | Male | No | 0 | Graduate | No | 6000 | 0 | 141 | 360 | 1 | Urban | Y |
| LP001011 | Male | Yes | 2 | Graduate | Yes | 5417 | 4196 | 267 | 360 | 1 | Urban | Y |
| LP001013 | Male | Yes | 0 | Not Gradu | No | 2333 | 1516 | 95 | 360 | 1 | Urban | Y |
| LP001014 | Male | Yes | 3+ | Graduate | No | 3036 | 2504 | 158 | 360 | 0 | Semiurban | N |
| LP001018 | Male | Yes | 2 | Graduate | No | 4006 | 1526 | 168 | 360 | 1 | Urban | Y |
| LP001020 | Male | Yes | 1 | Graduate | No | 12841 | 10968 | 349 | 360 | 1 | Semiurban | N |
| LP001024 | Male | Yes | 2 | Graduate | No | 3200 | 700 | 70 | 360 | 1 | Urban | Y |
| LP001027 | Male | Yes | 2 | Graduate | | 2500 | 1840 | 109 | 360 | 1 | Urban | Y |
| LP001028 | Male | Yes | 2 | Graduate | No | 3073 | 8106 | 200 | 360 | 1 | Urban | Y |
| LP001029 | Male | No | 0 | Graduate | No | 1853 | 2840 | 114 | 360 | 1 | Rural | N |
| LP001030 | Male | Yes | 2 | Graduate | No | 1299 | 1086 | 17 | 120 | 1 | Urban | Y |
| LP001032 | Male | No | 0 | Graduate | No | 4950 | 0 | 125 | 360 | 1 | Urban | Y |

This is a testing dataset to test the model without the results to check for the output of predictions

### C. *Registration Page Input Form*

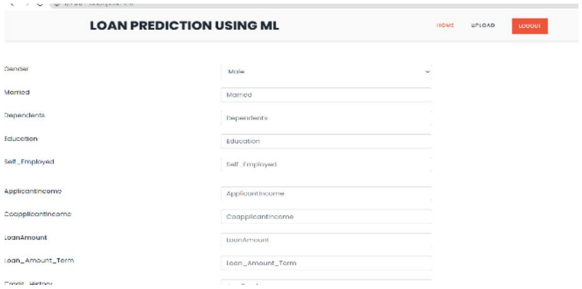

This is the Web page for registration of the end users in order to have an access to the website to check for their eligibility.

*D. Login Page Input Form*



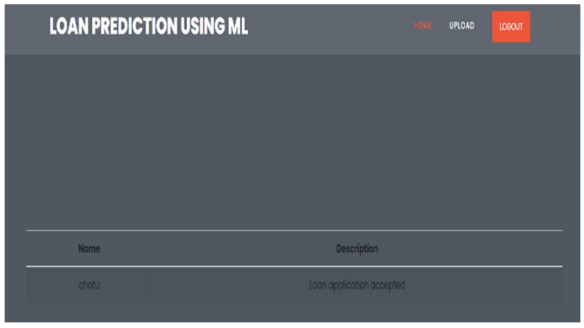

This is the login page to authenticate the end user to have an access to the facilities in the web page.

*E. Upload Details*




This is the input interactive web page wherein the user needs to manually enter the details of his personnel and post it to the backend to fetch the result

*F. Predicted Result*




This is the end result page where the user gets his eligibility result for the given input in the upload section.

## VI. MODEL COMPARISON

As a part of the project, the models under our consideration are:

*1)* Support Vector Machine
*2)* Logistic Regression
*3)* Random Forest Classifier

| Model | Accuracy |
|---|---|
| Logistic Regression | classification report of training data<br>precision recall f1-score support<br><br>0 0.95 0.45 0.61 154<br>1 0.80 0.99 0.88 337<br><br>accuracy 0.82 491<br>macro avg 0.87 0.72 0.75 491<br>weighted avg 0.84 0.82 0.80 491<br><br>classification report of testing data<br>precision recall f1-score support<br><br>0 0.83 0.39 0.54 38<br>1 0.78 0.96 0.86 85<br><br>accuracy 0.79 123<br>macro avg 0.81 0.68 0.70 123<br>weighted avg 0.80 0.79 0.76 123 |
| Support Vector Machine | cost of training model<br>precision recall f1-score support<br><br>0 0.96 0.47 0.63 154<br>1 0.80 0.99 0.89 337<br><br>accuracy 0.83 491<br>macro avg 0.88 0.73 0.76 491<br>weighted avg 0.85 0.83 0.81 491<br><br>cost of testing model<br>precision recall f1-score support<br><br>0 0.79 0.39 0.53 38<br>1 0.78 0.95 0.86 85<br><br>accuracy 0.78 123<br>macro avg 0.78 0.67 0.69 123<br>weighted avg 0.78 0.78 0.75 123 |
| Random Forest Classifier | cost of training model<br>precision recall f1-score support<br><br>0 0.99 0.90 0.94 154<br>1 0.95 1.00 0.98 337<br><br>accuracy 0.97 491<br>macro avg 0.97 0.95 0.96 491<br>weighted avg 0.97 0.97 0.96 491<br><br>cost of testing model<br>precision recall f1-score support<br><br>0 0.68 0.39 0.50 38<br>1 0.77 0.92 0.84 85<br><br>accuracy 0.76 123<br>macro avg 0.73 0.66 0.67 123<br>weighted avg 0.74 0.76 0.73 123 |

From the above designed models, we have considered Random Forest Classifier as a model with 97% train accuracy to predict the eligibility criteria of loan applicants.

## VII. CONCLUSION

Consequently, we built a system in which we can directly submit our information through to the homepage, and the input is then transferred to the data layer, wherein we employed our random forest approach to examine the data. This algorithm using random forests for client getting approved is a dependable and efficient way for determining the possibility of a consumer's loan acceptance. It has a great deal of precision in predicting loan favorable ratings and gives an easy approach to deciding on getting approved. It is additionally less susceptible to error because the system takes into account several parameters when calculating the risk of a home loan. Additionally, the Random Forest approach is very flexible and may be applied to large datasets.

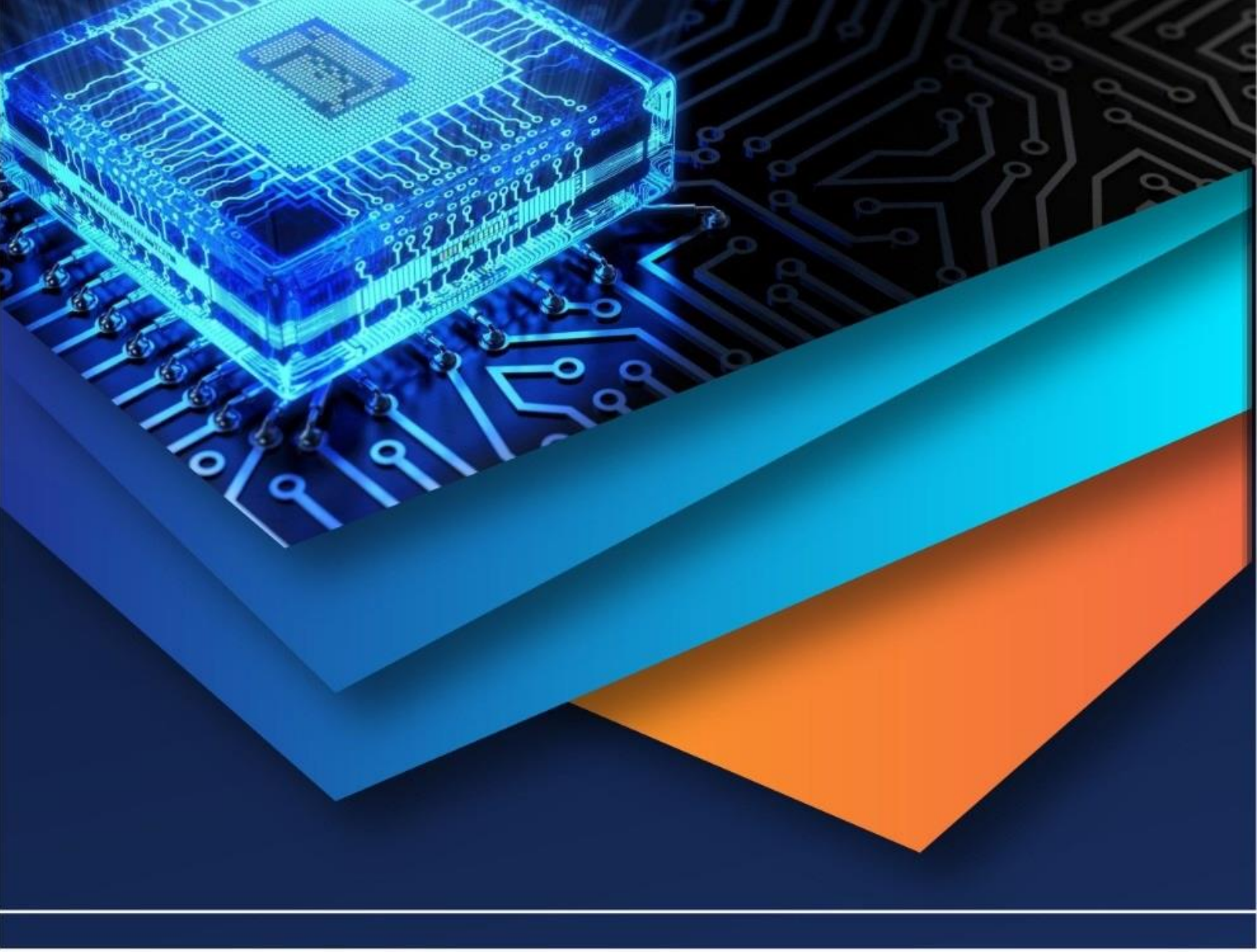
## VIII. FUTURE ENHANCEMENT

We may try to develop and improve the current techniques so that the correctness of the result is enhanced and the time required is decreased so that we can receive an outcome in a brief time, and we can attempt to integrate them for any active learning environment in order to ensure the banker's hard workload is lowered.

## REFERENCES

[1] Kumar Arun, Garg Ishan, Kaur Sanmeet, May-Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)
[2] Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei University of Technology, Hefei 2009, China

[3] Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C.Sealand on july 20, 2018.
[4] Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques, Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.
[5] K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).
[6] S.S. Keerthi and E.G. Gilbert. Convergence of a generalizeSMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002.
[7] Shiva Agarwal, “Describe the concepts of data mining”, Data Mining: Data Mining Concepts and Techniques, INSPEC Accession Number: 14651878, Electronic ISBN:978-0-7695-5013-8, 2013.
[8] Aboobyda, J. H., and M. A. Tarig. "Developing Prediction Model of Loan Risk in Banks Using Data Mining." Machine Learning and Applications: An International Journal (MLAIJ)3.1, 2016.
[9] A kindaini, Bolarinwa. “Machine learning applications in mortgage default prediction.” University of Tampere, 2017.
[10] Amir E. Khandani, Adlar J. Kim and Andrew Lo, “Consumer credit-risk modelsvia machinelearning algorithms and risk management in banking system”,J. Bank Financ., vol. 34, no. 11,pp. 27672787, Nov. 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY