An Exploratory Investigation into Developing an Optimal Model for Predicting Credit Approval

Kategari Jyothirmai

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *This paper explores credit application data sourced from the Credit Approval dataset within the AI repository of the University of California (UCI). Its objective is to identify key factors influencing credit approval decisions and determine which data analysis techniques yield the most effective model for predicting application outcomes using the same algorithm. The project aims to develop a model capable of determining whether a financial institution should approve credit card applications from its customers. Such decisions are typically based on various factors such as creditworthiness, loan history, repayment track record, and income criteria. To achieve this, multiple classification techniques were employed, and various classification methods were compared to identify the most suitable one for the dataset. These findings can serve as valuable insights for consumers, while for financial industry professionals, it offers a model that could potentially automate the credit application approval process.*

I. INTRODUCTION

The decision to approve a credit card or loan heavily relies on the individual and financial profile of the applicant. Factors such as age, gender, income, employment status, credit history, and other attributes contribute to the approval decision. Credit analysis involves both statistical-quantitative and qualitative measures to assess the likelihood of a third party repaying the loan to the bank on time and predict its default risk. The analysis focuses on identifying, evaluating, and mitigating financial/other risks that may lead to losses for the institution while lending. Mismanagement of credit risk can have adverse effects on credit management; hence, evaluating credit approval is crucial before making any lending decision.

Recognizing the significance of credit assessment, numerous studies and research efforts have utilized data mining tools to enhance existing models and improve prediction accuracy. Data mining plays a vital role in uncovering insights, offering a wide array of tools, techniques, and methodologies for pattern recognition and data analysis. Various applications include fraud detection, market basket analysis, trend/sentiment analysis, market segmentation, and credit scoring. One such application explored in this study is Credit Approval, which involves classifying applications based on data mining classification methods. Banks consider financial and demographic information before establishing decision rules to mitigate risks. This process identifies individuals who may pose risks to credit approval, and the results help refine the credit approval strategy. Decision trees, support vector machines, and logistic regression are commonly used techniques for classification in credit risk assessment, yielding promising results.

The primary objective of this study is to apply data mining techniques to the credit approval dataset and design models for predicting approval decisions using supervised learning. [1].

II. TECHNIQUE

In this evaluation work, Supervised ML Algorithms like Decision Tree and KNN are discussed.

2.1 Decision Tree

A decision tree is essentially a tree, but the meaning of its parts is novel according to that of a standard tree. Each non-leaf centre point tends to an attribute, each branch tends to a yield, and each leaf centre point tends to one class [4]. Decision Tree is an unquenchable estimation, which assembles a decision tree in a recursive manner totally. Decision tree is a kind of overseen learning. According to the development of the decision tree, the decision tree can be segregated into matched decision tree and multi-branch tree [2][3]. For example, some decision tree computations simply produce twofold trees, while other decision tree estimations May convey non-matched trees. The going with will quickly introduce the decision tree age measure, pruning development and typical decision tree estimations.

The middle issue of the decision tree estimation is the method for picking attributes. After the decision model is set up, a particular estimation is used to prune the tree through the test set. Normally, attribute assurance depends upon information secure, information obtain extent, Gini coefficient and chi-square test.

The advancement of the decision tree is generally portrayed as :

1. Work the data getting ready set by client, starting with an empty tree, and thereafter dividing it into fitting classes subject to trademark testing.
2. Acquire data through planning dataset, building decision model through recursion from top.
3. Through assessment improvement computation then discover the possible division of every model set and use a specific estimation to prune the tree through the testing dataset.
4. After later pruning cycle to clear out peculiarities that might exist, finally outlining an absolute decision tree.

2.2 K-Nearest Neighbor (KNN) computation

KNN computation is one of the least demanding portrayal estimations and it is maybe the most used learning computations. KNN is a non-parametric, languid learning computation. Its inspiration is to use an informational index in which the data centre's are confined into a couple of classes to predict the request for another model point [5][7].

The K-NN working can be explained in light of the underneath computation:

- Step-1: Select the number K of the neighbors
- Step-2: Calculate the Euclidean distance of K number of neighbors
- Step-3: Take the K nearest neighbors as indicated by the decided Euclidean distance.
- Step-4: Among these k neighbors, count

III. EXPERIMENTAL RESULTS

This section presents findings and discussion on data-driven analysis of the Credit Approval dataset obtained from the UCI repository. The study utilized Weka, a tool developed by researchers at the University of Waikato in New Zealand. Written in Java, Weka offers a graphical user interface for interacting with data files and provides various facilities for developing machine learning methods and applying them to real-world data mining problems. Weka processes data in ARFF format, which includes special tags to indicate data file structure. It implements algorithms for data pre-processing and classification. The dataset comprises 690 instances with 16 attributes, divided into two classes: approved credit (307 instances) and rejected credit (383 instances). Analyses were conducted using 70% of the dataset for training and 30% for testing. The statistical summary of the dataset is illustrated in Figure-1

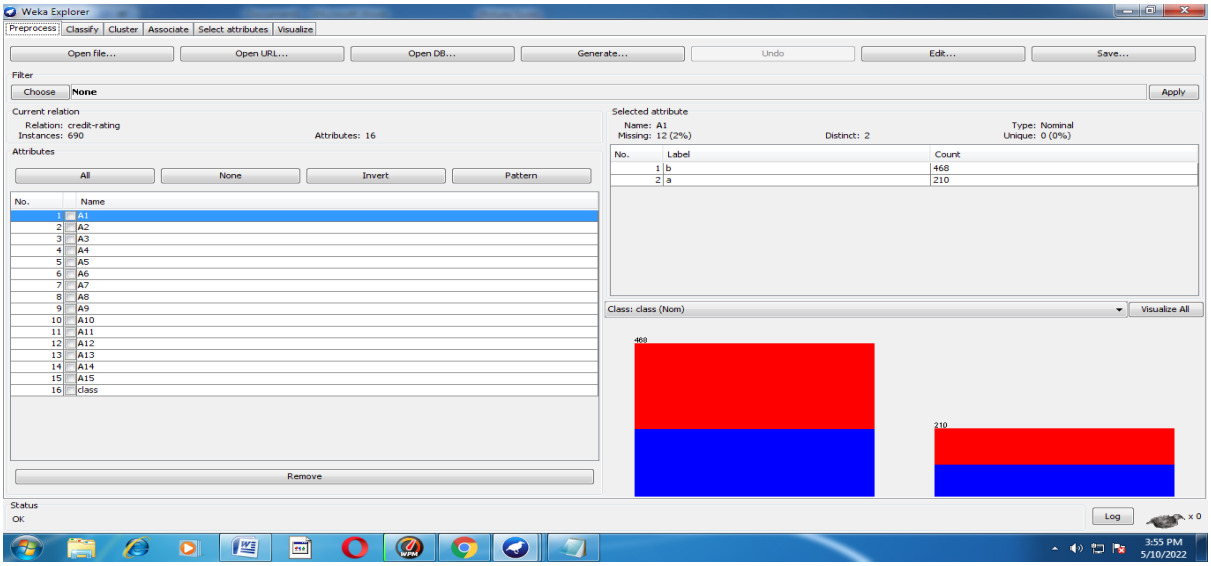

Figure-1: Statistical Information of Dataset

The results of decision tree and KNN classifiers are compared the on basis of correctly classified instances is shown in the table-1 and same shown in the figure-2.

Table-1
Experimental Results

Algorithm	Accuracy	Precision	Recall
Decision Tree	86	86	86
KNN	81	80.93	81

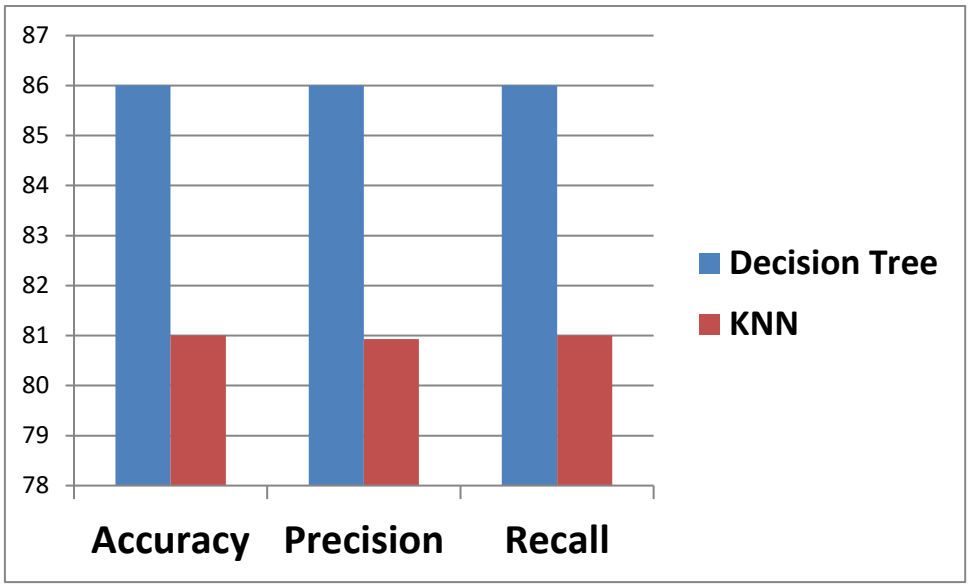

Figure-2: Classifier performance

From the figure-2, we notice the exhibition of decision tree classification has got 86% of Accuracy and the KNN has achieved the accuracy of 81%. So, the decision tree classification has got highest accuracy when compared to KNN. The screen shots of experimental results are shown in the figure-3 and figure-4.

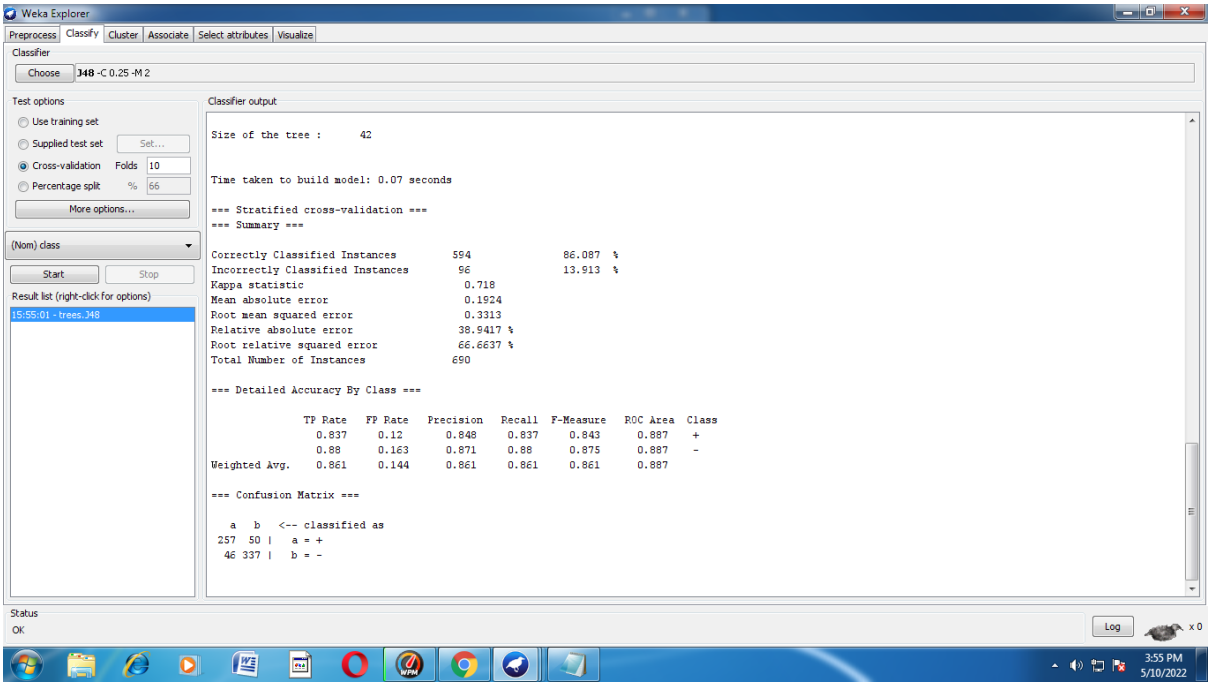

Figure-3: Screen shot of decision tree classifier

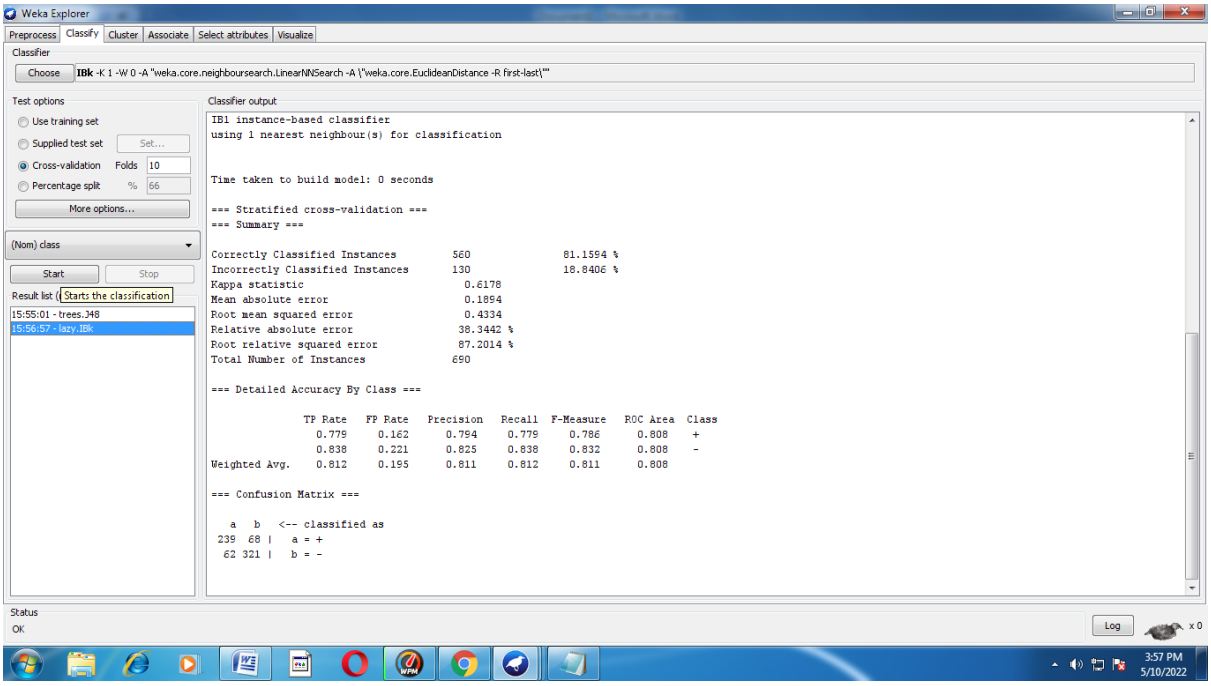

Figure-4: Screen shot of KNN classifier

IV. CONCLUSION

The model we developed serves as a predictor and analyzer, determining whether a financial institution will approve a card for a consumer. Through our observations, we identified key features crucial for credit approval decisions, including prior defaults, years of employment, credit score, and debt. Currently, factors considered encompass general details such as gender, age, credit reports, income, and employment duration. To enhance our model, additional factors like criminal history and asset ownership (both physical and liquid) could be considered. We utilized two classification algorithms, decision tree and KNN, to build and compare models, aiming to improve accuracy for practical application.

REFERENCES

[1] G.Ravi Kumar, K. Tirupathaiah and B.Krishna Reddy , "Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques", International Journal of Computer Sciences and Engineering, PP:842-846, e-ISSN: 2347 — 2693, Vol-7, Issue-6, June 2019

[2] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005.

[3] J.Han and M.Kamber, "Data Mining concepts and Techniques", the Morgan Kaufmann series in Data Management Systems, 3nd edition, Morgan Kaufmann, ISSN:9780123814791, 2011

[4] S. Rahamat Basha, J.Keziya Rani, JJC Prasad Yadav and G.Ravi Kumar, "Impact of feature selection techniques in Text Classification: An Experimental study", Advances in Engineering, Management and Sciences, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-3, PP: 39-51, ISSN (Print) 0973-8975, September 2019

[5] Surya Bhupal Rao, S.Rahamat Basha and G Ravi Kumar, "A comparative approach of text mining: classification, clustering and extraction techniques", Journal of Mechanics of Continua and Mathematical Sciences, PP:120-131, ISSN (Print) 0973-8975, Special Issue, No.5, January 2020

[6] UCI machine learning repository. http://archive.ics.uci.edu/ml/.

[7] Written H.I, Frank E, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann Publishers, 2005.