**FORMAN CHRISTIAN COLLEGE**

**(A CHARTERED UNIVERSITY)**

****

**COMP 360 Section A**

**Final Project**

**Group Members**

Zohaib Javed 21-10445

**Date of Submission**

Sunday, July 3, 2022

**Submitted to**

Miss Maria Taimoor

**Abstract**

With growing population and urbanization in Pakistan, more people are wanting to apply for house loans. According to the traditional system, clients apply for a house and wait for the bank staff to approve their loans. This is done after investigating their financial status and other factors. Which is rather hectic and long process, the program that we have made will automate the entire process, the customer will just need enter their credentials in the eligibility form, these details include, their gender, marital status, education, number of dependents, income loan amount, credit history and others. To automate the process, we have given a dataset to our software, this will help it identify the client segment, so it can filter out clients automatically.

**Introduction**

Pakistan is urbanizing at very high rate currently. Every big city is expanding with multiple ongoing projects to build big towns. Most people here are shifting to urban areas from rural areas, most of these people cannot afford their own house and are living on rented houses/apartments. With this growing problem more and more people are approaching different banks to get their home loans approved. The procedure for house loan approval is very hard and long process, which contains a very lengthy investigation and multiple employees are indulged in this task with hefty documentation. We are aiming to produce an automated system where on a simple portal a customer enters their credentials, and this system would automatically filter out the clients who are most likely to get their loan approved. The clients would find out if they are eligible and the bank would only invest their time and money with who they can do business with.

**Literature Review**

In this segment the work done in the domain of improving the banking sector using AI (machine learning and deep learning) will be discussed. With increase demand in Pakistan for house loans and world of technology has evolved the banking sector has needed to adapt with this change and with everyday passing the advancements in technology has grown exponentially. Some work done in this sector already is as follows:

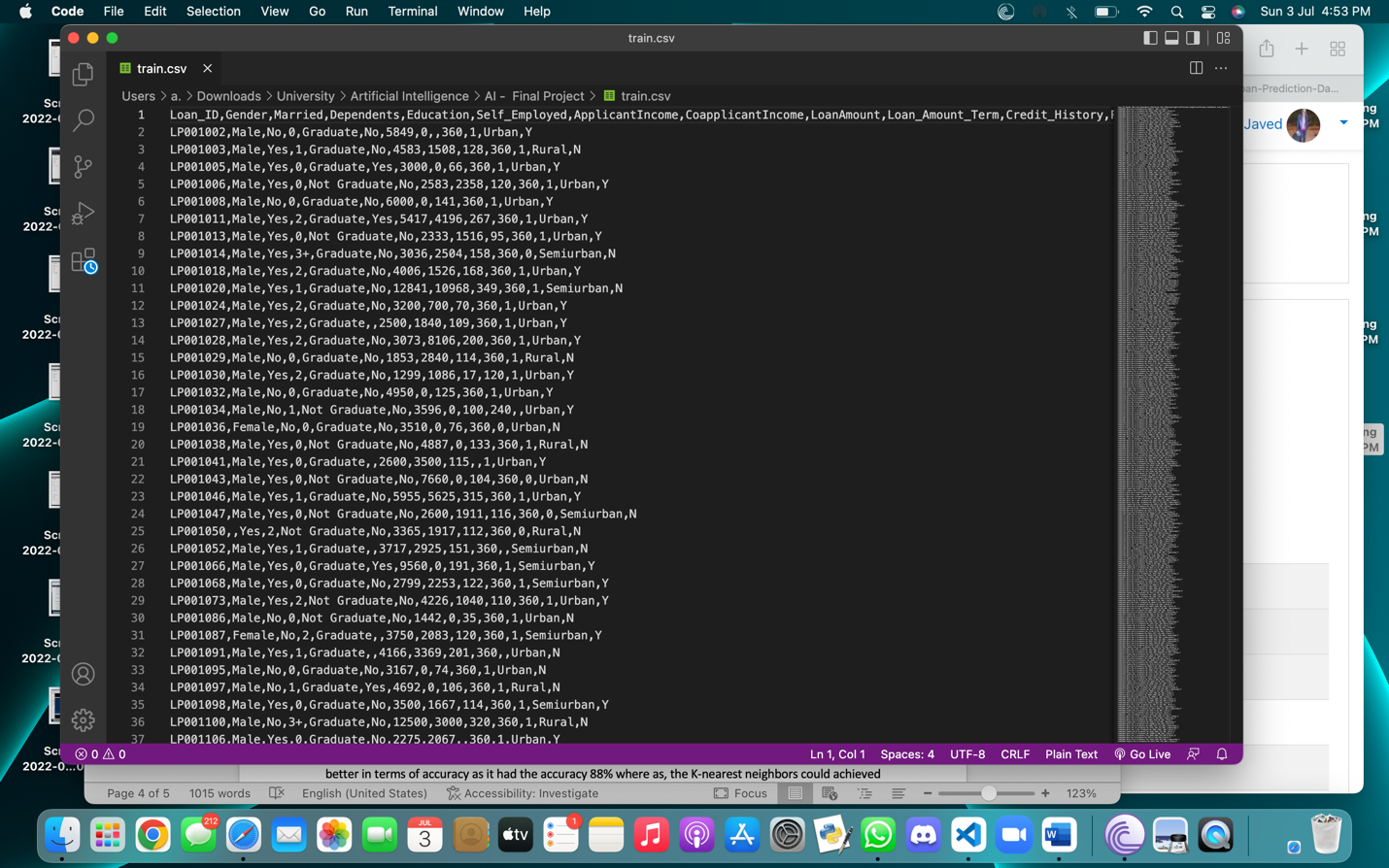
The Random Forest Algorithm was adopted by Lin Zhu et al. in paper [1] and Nazeeh Ghatasheh in paper [2] to construct a model for loan default prediction. Paper [1] concluded that random forest has much better accuracy (98%) than other algorithms like logistic regression (73%), decision trees (95%), and support vector machines (75%). The results of the paper [2] concluded that the random forest algorithm is one of the best options for credit risk prediction. Paper [2] also talked about the advantages of the algorithms, which are the competitive classification accuracy and simplicity. Paper [3] reviewed many methods available like logistic regression, k- nearest neighbors, random forest, neural networks, support vector machines, stochastic gradient boosting, Naive Bayes, etc. and concluded that it is nearly impossible to declare one best method of all. Nikhil Madane and Siddharth Nanda in paper [4] reviewed credit scoring of mortgage loans and made the following conclusions: Credit applications that do not pass certain requirements are often not accepted because the probability of them not paying back is high. Low-income applicants are more likely to get approval, and they are more likely to pay back their loans in time. **Methodology**

The classifiers we have used in this project are K-Nearest Neighbors and Naïve bayes classifier to predict if the bank loan would be approved or not.

K-Nearest Neighbors uses data and classify new data points based on similarity measures. Classification is done by a majority vote to its neighbors. The data is assigned to the class which has the nearest neighbors. As you increase the number of nearest neighbors, the value of k, accuracy might increase.

On the other hand, Naïve bayes classifier is based on bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high- dimensional training dataset. It helps in building the fast machine learning models that can make quick predictions. It predicts based on the probability of an object.

**Experiments and results**

We used a dataset available on the internet (referenced below), which is given below 

This dataset has 13 columns and 615 rows. We used this dataset to train or algorithm so it can predict if a loan will be approved or not. We split this data set into two parts, one for training purposes and one for testing. The testing data is 1/8th the size of training data. The maximum accuracy that we could achieve was 88%. As mentioned before, we used two different classifiers the naïve bayes method was better in terms of accuracy as it had the accuracy 88% where as, the K-nearest neighbors could achieved only 85% accuracy. We tried to visualize the data and were able to display the graph of k-nearest neighbors method.

Chart, scatter chart

Description automatically generated

**Conclusion**

Artificial intelligence is no doubt the future of tech world, and programs like this can reduce human effort to a high extent. As so, this program would help banking officer’s effort in approval for house loans. The long process of approvals through documentation would no longer be the necessity. Not only the employees would be as ease, but the client’s time would also be saved with fairly high accurate results.

**Resources**

[1] Zhu L, Qiu D, Ergu D, Ying C and Liu K 2019 A study on predicting loan default based on the random forest algorithm The 7th Int. Conf. on Information Technol. and Quantitative Management (ITQM) 162 pp 503–13

[2] Ghatasheh N 2014 Business analytics using random forest trees for credit risk prediction: a comparison study Int. Journal of Advanced Science and Technol. 72 pp 19–30

[3] Breeden J L 2020 A survey of machine learning in credit risk

[4] Madane N and Nanda S 2019 Loan prediction analysis using decision tree Journal of The Gujarat Research Society 21 p p 214–21

<https://github.com/shrikant-temburwar/Loan-Prediction-Dataset>