Naïve Bayes Learning on Fraud Detection

Saiprakashreddy Balupalli
11546701
University of North Texas
Texas, USA
SaiprakashreddyBalupalli@my.unt.edu

Laxmi Sahithi Nagarapu
11549809
University of North Texas
Texas, USA
Laxmisahithinagarapu@my.unt.edu

Shiva sai Konda
11596850
University of North Texas
Texas, USA
ShivaSaiKonda@my.unt.edu

Sowmya Varanganti
11605359
University of North Texas
Texas, USA
sowmyavaranganti@my.unt.edu

Tirumala Siva Nagarjuna Mothukuri
11510377
University of North Texas
Texas, USA
Tirumalsivamothukuri@my.unt.edu

***Abstract*—Since the dawn of the human species, people have had the idea of taking loans, but it is now taking on many different forms. Except for the banking industry, which requires security for official loans, this includes personal exchanges of loans for payback based on personal track records and enjoying loans as the proceeds of daily contributions. The goal of providing loans to private persons and businesses is to stimulate the economy while earning money for the lenders through the interest on the loans. Credit history judging by humans is ineffective due to the volume and diversity of data; case-based, analogy-based reasoning and statistical methodologies have been used but in the twenty-first century, these traditional methods are incapable of detecting fraudulent efforts. Based on training and testing of a labeled dataset, this work employs supervised machine learning more specifically the Naive Bayes, to anticipate fraudulent activities in loan administration and take a decision on loan approval. Model is implemented in probabilistic approach using Bayes theorem. model is designed to analyze the income and other background details of applicant and classify the eligibility requirement. To improve the model's classification accuracy K-fold cross validation techniques is applied.Compared to the previous work done on the loan data set we have got good results in this model. The previous work was 71.11 percentage accuracy whereas in our paper we were able to get 82.9 percentage accuracy with Naive Bayes and 79 percentage by performing K fold cross validation to Naive Bayes model.**

I. Introduction

In order to boost the economy and make money for the lenders, private individuals and corporations are given loans. Due to the volume and diversity of data, human credit history judgment is an ineffective evaluation approach. Case-based, analogy-based reasoning, and statistical methodology have been used, but in the twenty-first century, these traditional techniques are unable to detect fraudulent actions. Using standard audit processes to identify fraud detection might be challenging due to a lack of understanding of their characteristics. Data mining techniques, which assert to have sophisticated categorization and prediction capabilities, may make it easier for auditors to find managerial wrongdoing [1]. Machine learning and related techniques are most utilized, which includes artificial neural networks, Naive Bayes, Decision Trees, logistic regression and support vector machine (SVM). The Naive Bayes classifier assumes that characteristics are independent of class, which considerably simplifies learning [2]. The Naive Bayes machine learning classifier attempts to predict a class known as the outcome class using probabilities and conditional probabilities of how many times it occurred from the training data. This type of learning, known as supervised learning, is extremely effective, quick, and accurate for real-world circumstances. [3].

There are some more supervised learning algorithms like KNN (K-Nearest Neighbor), DT (Decision Tree), MLP (Multi-Layer Perceptron) other than Naive Bayes. The 'K-Nearest Neighbors' algorithm (K-NN) is a supervised learning technique that is non-parametric. It is a method for categorizing data that calculates the likelihood that a data point belongs to one group or the other based on which group the data points closest to it belong. A decision tree is a flowchart-like structure in which each leaf node denotes a class name, each internal node denotes a "test" on an attribute, and each branch denotes the test's result. A multilayer perceptron (MLP) is a feedforward Artificial Neural Network that creates a collection of outputs from a set of inputs. A directed graph connecting the input and output layers of an MLP is made up of numerous layers of input nodes. Ensembling techniques also often used to enhance the models with increased performance.In [4] paper hybrid classification method, which is the mixture of KNN and Naive Bayes is used. They have compared the hybrid classification method with Voting based classification technique. The hybrid classification method will reduce the complexity of the model and increases various parameters like recall, precision and reduce execution time. As per analysis the hybrid classification model is more accurate and shows high

performance when compared to voting based classification technique

In [5] paper presents a survey of various techniques used in credit card fraud detection mechanisms. They studied applications of machine learning like Naïve Bayes, Logistic regression, Random Forest with boosting and shows that it proves accurate in deducting fraudulent transaction and minimizing the number of false alerts. If these algorithms are applied into bank credit card fraud detection system, the probability of fraud transactions can be predicted soon after credit card transactions. And a series of anti-fraud strategies can be adopted to prevent banks from great losses and reduce risks. The objective of the study was taken differently than the typical classification problems in that they had a variable misclassification cost. Precision, recall. f1-score, support and accuracy are used to evaluate the performance for the proposed system. By comparing all the three methods, they found that random forest classifier and Naive Bayes with boosting technique is better than the logistic regression.

The Bayesian classifier or Naive Bayes was put up against KNN, DT, and MLP to demonstrate the effectiveness of the strategy. When different feature extraction techniques are combined with various classifiers, the results can vary up to 90 percent. Artificial Neural Networks (ANN), SVM, Decision Trees, KNN algorithms, Bayesian classifiers, etc. are examples of intelligent classification techniques. A complex structure and a sluggish training convergence are ANN's drawbacks. Despite being widely utilized, SVM also has complex parameter optimization problems. Due to its superior generalization ability and great efficiency, the naïve Bayes classifier (NBC) might be a more practical choice [6]. Naive Bayes is highly scalable and efficient algorithm for text classification using word frequencies as features [7]. Multilabel Naïve bayes can handle high volumes of data for multilabel classification.In [3] paper presents the implementation of Naive Bayes and KNN algorithm on same credit card dataset to calculate the precision of algorithms to identify the fraudulent transactions in the dataset. Experimental results depict that both classifiers work differently for the same dataset. The purpose is to enhance the precision, accuracy and increase the flexibility of the algorithm. Credit Card Fraud Detection for given data set has been done using Naive Bayes and KNN individually with the precision of approximately 95% and 90% respectively. Hence, we will implement Naïve Bayes on the Loan data set to evaluate the profile of applicant using his/her various personal and income wise check and classify the applicant as genuine eligible application or fraudulent ineligible application. Model learns the criteria for already approved/rejected applications and henceforth predict the unseen applications. The model helps in reducing time and effort to great extent in initial background check of applicants. The data generated from source point being raw and dirty needs to be cleaned. Noise removal is removing the features responsible for generating repetitive constant information in dataset [7], feature extraction is a key preprocessing step where features are combined and analyzed for building precise model. Feature selections techniques performed as part of Data preprocessing to input the model with fine data. [7] Models can be trained with different trails of training data portions to obtain the optimal training set, but latest techniques K-fold cross validation can be implemented on model to obtain accurate results [8]. Traditional machine Learning approach involves 75percent of data for training and 25 for testing model but K fold cross validation is one of the efficient techniques to evaluate the model by dividing dataset into different folds training model on specific folds and evaluate on specific folds. This cross validation helps model to understand at most patterns in data thereby generating a precise model.

The project is designed as follows, post introduction we plan to describe the working and logic and methodology of model. In following section Our case study, work steps are illustrated. Finally experimental findings and results are shown with future scope of project in machine learning approach.

II. Methodology

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. Company wants to automate the loan pre approval process To achieve this an peer evaluation should be performed on application form filled by applicants about their certain back ground details. Company aims to target specific segment of audience for their business. Each record in dataset refers to specific loan application. Data set contains following features such as-
Unique features-
1) Loan ID – unique ID number for each Loan application
Categorical features-
2) Gender – The gender of applicant (Male, Female)
3) Marital status- Applicant is Married or not (Yes/No)
4) Dependents- No of dependents on applicant (0,1,2,3+)
5) Education – Applicants graduate educational status (Graduate/ Non Graduate)
6) Self Employed – if applicant is self employed or not (Yes/No) Numerical features-
7) ApplicantIncome – Monthly income of applicant in grands
8) CoapplicantIncome-Monthly income of coapplicant in grands
9) LoanAmount – Loan amount requested in grands
10) Loan_AmountTerm- Tenure of Loan payment in months
11) Credit_History- If applicant has previous credit history
12) Property_Area – location of property for loan (urban, semiurban,rural)
13) Loan_Status – Decision of pre approval for Loan
The key features involved in approval of loan are annual income of applicant, credit history because these help in determination of how sound the applicant is financially and help in identifying the probability of genuineness or faultiness of the profile. Once the profile is identified as genuine pre approval is granted and profile pushes for further manual evaluation of documents for approval on the other hand if the profile is identified as fraudulent then pre approval is

not granted. This saves a lot of time and effort for financial evaluation of profiles by company. So the target variable is LoanStatus and remaining all other are dependent variables.

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome
Male	No	0	Graduate	No	5849	0.0
Male	Yes	1	Graduate	No	4583	1508.0
Male	Yes	0	Graduate	Yes	3000	0.0
Male	Yes	0	Not Graduate	No	2583	2358.0
Male	No	0	Graduate	No	6000	0.0
...	...	...	...	...	...	...

LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
NaN	360.0	1.0	Urban	Y
128.0	360.0	1.0	Rural	N
66.0	360.0	1.0	Urban	Y
120.0	360.0	1.0	Urban	Y
141.0	360.0	1.0	Urban	Y
...	...	...	...	...

Fig. 1. sample dataset

It has already been proved that model performs better on cleaned data without bias in order to clean the data following Steps in Data preprocessing are performed. Identifying the Duplicate records and Null values are intial steps as duplicates and null values lead to baised model.

```
#count and display number of Duplicates in data
Duplicates=Data[Data.duplicated()]
Duplicates.count()
```

```
Loan_ID              0
Gender               0
Married              0
Dependents           0
Education            0
Self_Employed        0
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount           0
Loan_Amount_Term     0
Credit_History       0
Property_Area        0
Loan_Status          0
dtype: int64
```

Fig. 2. No of Duplicates

```
print('Number of duplicates in Loan Dataset:',duplicates[duplicates[0]==True].count())
print('Number of Null values in Loan Dataset:')
data.isnull().sum()
```

```
Number of duplicates in Loan Dataset: 0    0
dtype: int64
Number of Null values in Loan Dataset:
Loan_ID              0
Gender              13
Married              3
Dependents          15
Education            0
Self_Employed       32
ApplicantIncome      0
CoapplicantIncome    0
LoanAmount          22
Loan_Tenure         14
Credit_History      50
Property_Area        0
Loan_Status          0
```

Fig. 3. Nullvalues

Renaming columns – Loan_Amount_Term is not efficient way to represent the data so it has been renamed to Loan_Tenure. The next step is to identify if there are any duplicate records in data because duplicates lead to overfitting of model and infer incorrect results. So we calculated the total number of duplicate values and found there are no duplicates in data. The following step in data cleaning is to identify the Null values in data as they could lead to misfit of model. If there is a huge data available for training we can drop the records having null values but in our case there is limited data and cant afford to drop records so we have replaced the missing Null values with mean of corresponding column [7].While the categorical Null values are replaced with mode.

```
data_1[['ApplicantIncome','CoapplicantIncome','LoanAmount','Loan_Tenure']].corr()
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Tenure
ApplicantIncome	1.000000	-0.116605	0.570909	-0.045306
CoapplicantIncome	-0.116605	1.000000	0.188619	-0.059878
LoanAmount	0.570909	0.188619	1.000000	0.039447
Loan_Tenure	-0.045306	-0.059878	0.039447	1.000000

Fig. 4. Correlation

Once the data is cleaned we could analyze the data using different visualization techniques. The key numeric variables are used to identify the relationships among them by calculating the corelation cofficient. Now the distribution of varibles is seen inorder to get all round idea. The distribution of ApplicantIncome is shown below using histogram based on which we can infer most of the applicants have monthly income below 15000 USD.

```
data['ApplicantIncome'].hist(bins=5)
```

```
<AxesSubplot:>
```

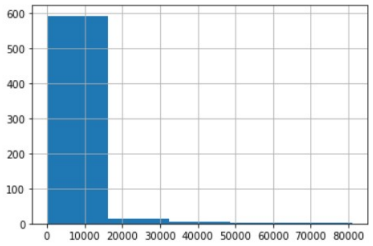


Fig. 5. Income distribution

Another histogram is developed for LoanAmount feature to understand what is over all pattern of loan requests capital from company. It can be seen that majority of people are rasing a loan request for capital of 80,000USD - 150000 USD.

Correlation matrix has been plotted to identify the strength of relation ships between features. Key performance indicators can be identified from the matrix. Now we can infer LoanAmount, ApplicantIncome are the main attributes in determining weather a person is eligible for loan disbursement.

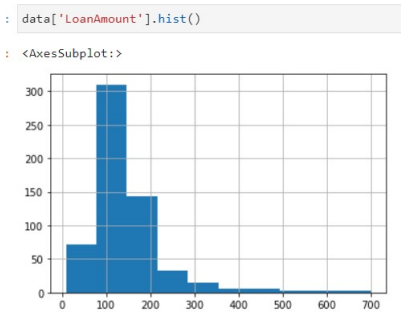


Fig. 6. LoanAmount distribution

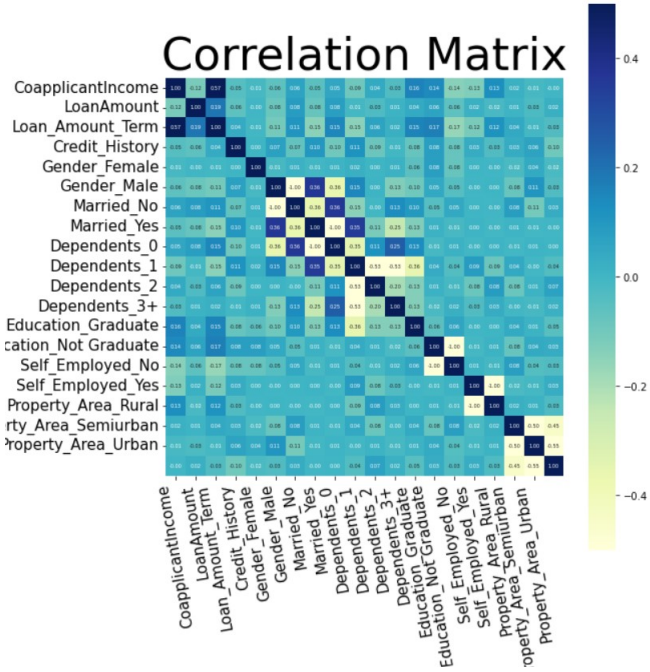


Fig. 7. Corelation matrix

A jointplot has been plotted between LoanAmount and HouseholdIncome which is the combined income of Applicant and Coapplicant.It depicts the scatter plot along with distribution of data. Also a scope for forecast of is determined from which loanamount can be estimated from House hold Income.

A biased data generates unbalanced dataset which is no more useful for building machine learning models. It leads to biased model with incorrect predictions. There are several types of bias comes in picture with data. Selection bias is the type of bias occurred when a specific group is excluded from study assuming the Loan data has been collected from all customers of financing company our Loan data set is free from sampling bias.

Recall bias occurs during data collection step, if a estimated or approximate value is taken instead of exact value. It could be because of database design issues and data collection environment issues. Assuming the data collected exactly and

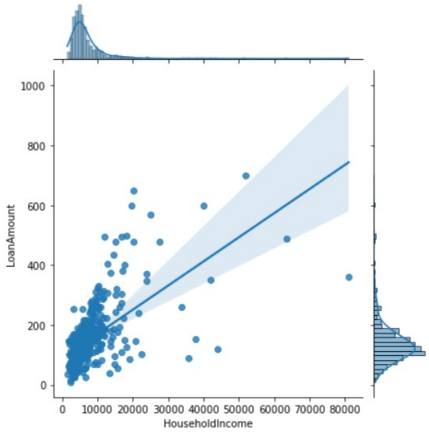


Fig. 8. Joint Plot

most of the key features involved are categorical in Loan Dataset Recall bias is minimized.

Measurement bias is another important type of bias often seen. It occurs mainly because of measuring the features in different standard units. we have We have eliminated the measurement bias by following the standardization technique. It is also refereed as Z-score normalization is a process of rescaling the features so that they will have properties of gaussian distribution. Subtracting the features from mean and dividing with standard deviation this brings all numerical units into a standard unit by eliminating the bias.The preprocessing is implemented using sklearn.preprocessing() technique. Final obtained data is now bias free and balanced dataset.

It is required that you have a fundamental understanding of the following concepts to understand Naive Bayes:
Conditional probability: a measure of the probability of event A occurring given that another event has occurred.
Joint Probability: a measure that determines the probability of two or more occurrences happening simultaneously.
Proportionality: Refers to the relationship between two quantities that are multiplicatively connected to a constant, or in simpler terms, whether their ratio yields a constant.
Bayes Theorem: Bayes' Theorem describes the probability of an event (posterior) based on the prior knowledge of conditions that might be related to the event. [1]–[15] The Naive Bayes machine learning classifier aims to predict an outcome class based on probabilities and conditional probabilities of its occurrence from the training data. For real-world circumstances, this sort of learning is extremely effective, quick, and accurate. It is also known as supervised learning. It is inspired by Bayes Theorem which states the following equation:

$$= P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

To make this equation easier to grasp, it may be rewritten using X (input variables) and y (output variable).

$$= P(A|B) = \frac{P(X|Y) * P(Y)}{P(X)}$$

We may rephrase P(X—Y) as follows because of the naïve assumption that variables are independent given the class:

$$P(X|y) = P(X1|Y) * P(X2|Y) * P(X3|Y) \ldots . P(Xn|Y)$$

P(X) is a constant because we are solving for y, so we may eliminate it from the equation and substitute a proportionality in its place.

$$P(Y|X) \alpha P(X|Y) * P(Y)$$

Argmax is simply an operation that finds the argument that gives the maximum value from a target function.

$$Y = \text{argmax}\,[P(Y) * \pi^{n}_{i=1} P(X_i|Y)]$$

Types of Naïve Bayes Model:

There are three types of Naive Bayes Model, which are given below:

Gaussian: The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution.

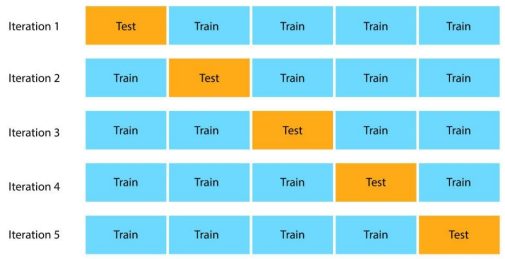
Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed. It is primarily used for document classification problems, it means a particular document belongs to which category such as Sports, Politics, education, etc. The classifier uses the frequency of words for the predictors.

Bernoulli: The Bernoulli classifier works like the Multinomial classifier, but the predictor variables are the independent Booleans variables. Such as if a particular word is present or not in a document. This model is also famous for document classification tasks.

Here we use gaussian naive bayes the data set is divided into 75% and 25% including dependent and target features. 75% of data is used for training the model and 25% is for validating the model. Model is trained using training data here model learns the probability of features LoanID, Gender, MarritalStatus, Dependents, Education, SelfEmployed, ApplicantIncome, CoapplicantIncome, LoanAmount, LoanTenure, CreditHistory, propertyArea given the condition LoanStatus is already known. Now the trained model performs on unseen test data where model predicts the probability of LoanStatus. Hence thenaive bayes model performs on Loan Dataset.

K-Fold Cross-Validation:

[2]In applied machine learning, cross-validation is a statistical technique that helps in model comparison and selection. In this model the data is split into 'k' number of subsets. Out of those k subsets one subset is kept has validation set to test the model remaining k-1 subsets are used to train the model. In the second experiment, completely different k subset is utilized for testing while the remaining k-1 are used for training. The experiment is repeated by using completely different k-1 subsets for training and the remaining k subset for testing where 10 experiments will be conducted subsequently. The output is the average of the results from all experiments.



The naive bayes model is implemented with k-fold cross validation technique to evaluate how naive bayes learns and performs on fold technique.The classification of model is determined using confusion matrix by measuring False positivity and False negativity.

III. Results and Conclusion

For detecting loan fraud, many top machine learning algorithms are applied. The experimental evaluation in this proposed system makes use of the proposed models Naive Bayes and K-fold cross validation. It is well known that models work better with bias-free, cleansed data. Finding duplicate entries in the data is the next action taken, since they cause the model to overfit and produce false results. When we tried finding out the duplicate entries there were none in our data set. By proceeding further, we know that null values have a negative impact on any machine learning algorithm's performance and accuracy. The null values are detected replaced them with the mean of the column. Then the categorical values are changed to numerical values. The data is studied using several visualization approaches once the data has been cleansed. By computing the correlation coefficient, the important numerical variables are employed to determine the connections between them. Based on the histogram used to display the distribution of applicant income, it can be concluded that the majority of applicants earn less

than $15,000 USD per month. Another histogram that was created for the Loan Amount feature helps to identify the general trend of loan requests for money from businesses. It was obvious that the majority of people were asking for loans for between 80,000 and 150000 USD. Unbalanced datasets produced by biased data are useless for the construction of machine learning models. A biased data generates unbalanced dataset which is no more useful for building machine learning models. It results in a model that is skewed and makes bad predictions. When a particular group is omitted from a study, bias of this kind occurs. Another significant and frequently seen type of bias is measurement bias. By using the standardization procedure, we have removed the measurement bias. Assuming that the information was correctly gathered, and the Loan Dataset contains classifications for the bulk of the significant characteristics. Recall bias is reduced Z-score normalization is the process of rescaling the features to give them gaussian distribution characteristics. By eliminating the bias, all numerical units are converted into a standard unit by subtracting the features from the mean and dividing by the standard deviation. The sklearn.preprocessing() approach is used to implement the preprocessing. The final data set is now a balanced, unbiased dataset. With the help of training data, the Naive Bayes machine learning classifier attempts to predict an outcome class using probabilities and conditional probabilities of its occurrence. This kind of learning is very quick, accurate, and effective for real-world situations. In this study, we use Gaussian Naive Bayes on a data set that is split into 25% dependent features and 75% target features. 25% of the data is used to validate the model, and 75% of the data is used to train the model. Now, the Naïve Bayes model is built, and the model is evaluated, and the accuracy is calculated and displayed which was 82 percentage. Then visualization was developed to plot confusion matrix and displayed the summary stats of the classifier. It shows 87 records has true and 15 records has false which were correctly classified, but I showed 3 true records has false and 18 false records has true. In addition, Cross-Validation was also performed, which is a statistical method used in applied machine learning that aids in model comparison and model selection. The data in this model is divided into k subgroups. The remaining k-1 subsets are utilized to train the model after one of those k subsets is kept as validation data to test the model. In the second experiment, the remaining k-1 are used for training while an entirely separate k subset is used for testing. 10 experiments will then be carried out after the experiment is repeated using completely different k-1 selections for training and the remaining k subset for testing. The output represents the average of all experiment outcomes. To assess how the naïve bayes model learns and performs on fold approach, the k-fold cross validation method is used. After using Naive Bayes alone, we were able to get an accuracy of up to 82.9 percent, but when using K-fold, the accuracy dropped to 79 percent. This leads us to the conclusion that Naive Bayes alone provides good accuracy rather than using additional techniques like K fold to Nave Bayes. By assessing True positivity and False negativity, the confusion matrix is used to classify the model. ROC curve is generated between true positive rate and false positive rate to identify the performance of the model. In real time situations banks accept the 80-percentage accuracy model and consider it has a good fit to grant the loans as the data is very huge and variant. In the previous work on the same dataset which was provided in Kaggle they have used several machine learning models like Logistic Regression, K - Nearest Neighbor, Support Vector Machine, Naïve Bayes, Decision Tree, Random Forest, and Gradient Boost but the maximum accuracy that they could get was 71.11 percentage. But, In the model which is presented in this paper we could achieve 82.9 percentage with Naïve Bayes and 79 percentage by using K fold cross validation.

REFERENCES

[1] G. D. Coderre, "Detection of fraudulent financial statements based on naïve bayes classifier', using data analysis techniques to detect fraud, global audit publications," 1999.

[2] J. H. I.Rish and T. Jayram, "An empirical study of the naive bayes classifier' an analysis of data char," *International Journal of Advanced Science and Technology*.

[3] R. K. N. K. D. K. M. S. Sai Kiran, Jyoti Guru, "Credit card fraud detection using naïve bayes model based and knn classifier," *Interntaional Journal of Advance Research Ideas And Innovations In Technology*, vol. 4.

[4] S. K. Darshan Kaur, "Machine learning approach for credit card fraud detection (knn and naïve bayes)," *The Online Journal of Science and Technology*.

[5] A. Bhanusri, "Credit card fraud detection using machine learning algorithms," *Quest Journals Journal of Research in Humanities and Social Science*, vol. 8, pp. 04–11.

[6] Y. Z. Z. T. B. X. R. C. Zhang X, Cong Y, "An investigation on early fault diagnosis based on naive bayes model," *Early Fault Detection Method of Rolling Bearing Based on MCNN and GRU Network with an Attention Mechanism, Shock and Vibration*, pp. 1–13, 2021.

[7] K. Q. Priyanga Chandrasekar, "The impact of data preprocessing on the performance of naïve bayes classifier," *2016 IEEE 40th Annual Computer Software and Applications Conference*.

[8] Z. R. Nafizatus Salmi, "2019 iop conf. ser.: Mater. sci. eng. 546 052068."

[9] D. H. W. Dr. S. Smys, "Naive bayes and entropy based analysis and classification of humans and chat bots," *Journal of ISMAC*, 2021.

[10] D. M. C. Sushma, "A credit card fraud detection using naïve bayes and adaboost," *International Journal of Scientific and Engineering Research*, vol. 10.

[11] A. Husejinović1, "Credit card fraud detection using naive bayesian and c4.5 decision tree classifiers," *Periodicals of Engineering and Natural Sciences*, vol. 8, no. 1, pp. 1–5.

[12] S. V. Toon Calders, "Three naive bayes approaches for discrimination-free classification," 2010.

[13] D. Beckerman, "Budgeted learning of naive-bayes classifiers'. a tutorial on learning with bayesian networks," *Microsoft Re¬search*, 1995.

[14] M. J. Pazzani, "learning on optimal navie bayes classifier'. searching for atribute dependencies in bayesian classifiers," *In preliminary Papers of the Intelligence and Statistics*, pp. 424–429, 1996.

[15] A. P. A. L. Sudiksha Wandre, Shefali Desai, "Credit card fraud detection using knn and navies bayes algorithm," *Journal of emerging technologies and innovative research(JETIR)*, vol. 9, 2022.