JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING ALGORITHMS FOR FINDING CREDIT SCORE PREDICTION FOR OPTIMAL OUTCOME

[1]Shaik Zareena, [2]P Jaya Surya, [3]T Divya Rani, [4]B. Sanjeev, [5]Dr. Subhani Shaik
[1,2,3]IV Year Students, [4]Assistant Professor, [5]Associate Professor
Dept. of Information Technology, Sreenidhi Institute of Science and Technology(Autonomous), Hyderabad, India

Abstract: In the rapidly evolving landscape of financial services, the accurate assessment of creditworthiness is paramount. This research paper delves into the realm of machine learning algorithms to ascertain their efficacy in predicting credit scores. The research employs a comprehensive dataset comprising diverse financial indicators and demographic information. By systematically evaluating the performance of each algorithm against this dataset, we aim to discern the strengths and limitations of different approaches in capturing the intricacies of credit risk. Our findings shed light on the most optimal machine learning algorithm for credit score prediction, considering both predictive accuracy and interpretability. The study compares various types of machines learning algorithms.Among all machine learning algorithms Logistic Regression is good in accuracy, F score, sensitivity and time also low. For time taken point of view KNN consume less time to finish the work but accuracy is low comparatively Logistic Regression. AdaBoost algorithm perform less compared to all machine learning algorithms and it take huge amount of time for execution of the data classification.

Keywords: Machine Learning Algorithms, Credit Score Prediction, Accuracy, Efficiency.

I. INTRODUCTION

The global financial landscape is undergoing a paradigm shift, propelled by advancements in technology and the increasing integration of machine learning algorithms in critical decision-making processes [1]. Among these, the assessment of creditworthiness stands as a cornerstone in financial institutions' ability to make loaningverdicts. Past credit scoring models, while reliable, often struggle to adapt to the dynamic nature of contemporary financial landscapes. In response, the application of machine learning algorithms has emerged as a promising avenue, offering the potential for enhanced predictive accuracy and a nuanced understanding of credit risk [2].

Against this backdrop, our research embarks on a comprehensive exploration of machine learning algorithms for credit score prediction, the overarching goal of identifying the most optimal approach. As the financial industry grapples with the need for sophisticated, data-driven solutions, understanding the comparative strengths and limitations of various algorithms becomes imperative [3]. This research seeks to address this need by conducting a thorough examination of established and cutting-edge machine learning models, ranging from traditional regression techniques to more complex neural networks [4].

Motivated by the challenges inherent in traditional credit scoring methods and the burgeoning capabilities of machine learning, our study not only endeavours to quantify the predictive power of these algorithms but also to unravel the interpretability and practical implications associated with their adoption in real-world financial settings [5]. Through a meticulous analysis of diverse datasets encompassing financial, demographic, and credit behaviour information, our research aims to provide a nuanced understanding of the performance nuances exhibited by each algorithm [6].

In doing so, this paper aspires to contribute to the broader discourse on the evolution of credit scoring methodologies, offering insights that resonate with financial institutions, regulatory bodies, and researchers alike [7]. By delineating the strengths and limitations of different machine learning algorithms, we hope to guide the development of more robust and adaptive credit scoring systems, fostering a new era of precision and reliability in the assessment of creditworthiness [8].

The following section 2 discuss with credit scoring system. Section 3 stated the results and analysis. Last section concludes the paper.

II. PROPOSED SYSTEM FOR CREDIT SCORING

This proposed system is used to improve the credit scoring systemprocedure and upsurge its efficiency and accuracy. The constituent of numerous classifier associationsnumerous classifiers to attainimprovedfalloutsat all of the specific classifier [9]. Most of the approaches for emergent classifiers turn around altering the working out dataset, developing classifiers on these n novel training sets, and integration the consequenceshooked on a soloconclusioninstruction. The following Figure oneportrays the proceduremovement of the recommendedensemble credit scoring system's model with multiple phases [10,13].

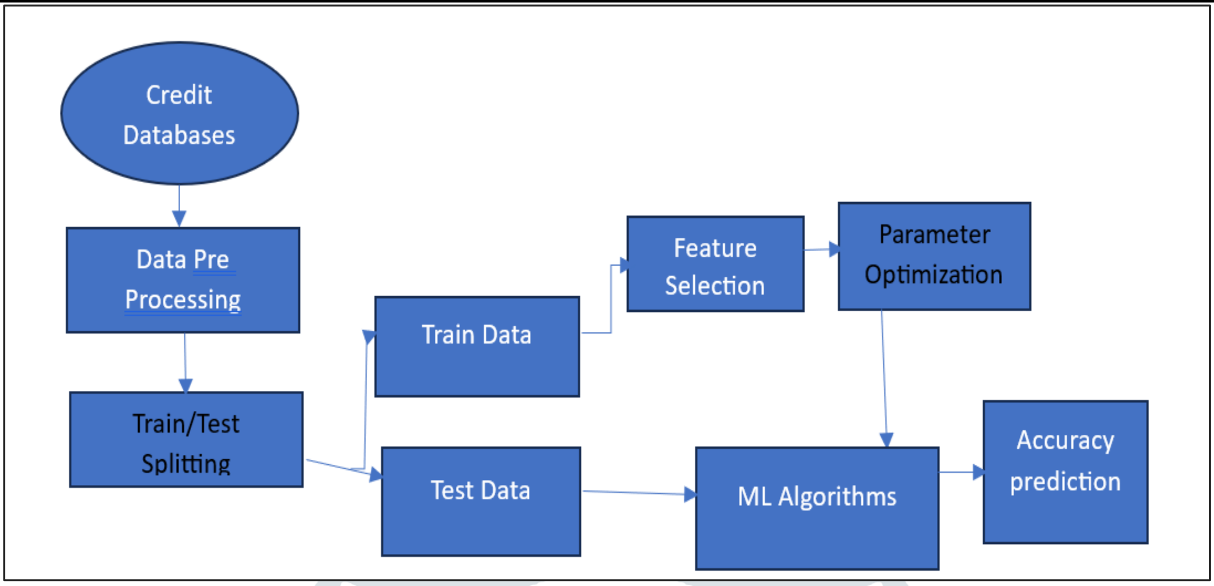


Figure 1. Block diagram of credit scoring system [11]

2.1 Credit datasets

This researchpracticesmulti credit datasets from conservative financial organizations. The UCI-ML provide all of the predictable credit datasets [12].They are most popular and frequently used by researchcontributors and reachable to the community. The recommendedtype was legalized using datasets from two countries names Australia and Germany.Athoroughexplanation is providing in table 1[14].

Table 1. Datasets description

Dataset	Attributes	Loans		
		Bad	**good**	**Total**
Germany	20	350	750	1000
Australia	14	393	679	1072

III. RESULTS AND ANALYSIS

The following results taken from real time data bases and use Google Colab for predict the results of multiple Machine Learning algorithms.

```
#importing all required modules

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


from google.colab import files
uploaded = files.upload()

     Choose files creditAnalysis.csv
     • creditAnalysis.csv(text/csv) - 9001 bytes, last modified: 21/12/2023 - 100% done
     Saving creditAnalysis.csv to creditAnalysis (4).csv


df = pd.read_csv('creditAnalysis.csv')
```

3.1 Data Pre-Processing

After loading the data, then preprocess the data for removing missing and abnormal values.

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 164 entries, 0 to 163
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 164 non-null    int64
 1   Gender              164 non-null    object
 2   Income              164 non-null    int64
 3   Education           164 non-null    object
 4   Marital Status      164 non-null    object
 5   Number of Children  164 non-null    int64
 6   Home Ownership      164 non-null    object
 7   Credit Score        164 non-null    object
dtypes: int64(3), object(5)
memory usage: 10.4+ KB
```

Table 2: Selection of data types with credit score labels

	Age	Gender	Income	Education	Marital Status	Number of Children	Home Ownership	Credit Score
Age	1.000000	0.235343	0.699464	0.170254	-0.517723	0.055390	-0.713803	0.205362
Gender	0.235343	1.000000	0.495738	0.248671	0.278362	-0.442139	-0.031519	-0.247729
Income	0.699464	0.495738	1.000000	0.369449	-0.471004	0.084547	-0.704928	0.083698
Education	0.170254	0.248671	0.369449	1.000000	-0.067797	0.047311	-0.397043	0.334424
Marital Status	-0.517723	0.278362	-0.471004	-0.067797	1.000000	-0.696984	0.708374	-0.205756
Number of Children	0.055390	-0.442139	0.084547	0.047311	-0.696984	1.000000	-0.497129	0.136517

Heat Maps are graphicdepictions of data that exploit color-coded systems. The main purpose of Heat Maps is to improvedimagine the volume of events within a dataset and support in directionalspectatorstoparts on data conceptions that substance mostly. The succeeding picture for age, gender and credit score heat map.

Figure 3: Heat Map

3.2 Class distribution of original and resampled data

```
Class distribution before resampling: 1    90
0    31
2    10
Name: Credit Score, dtype: int64
```

```
Class distribution after resampling: 1    90
0    90
```

```
2    90
Name: Credit Score, dtype: int64
```

The following picture shows the class distribution before and after original and resampled data respectively.

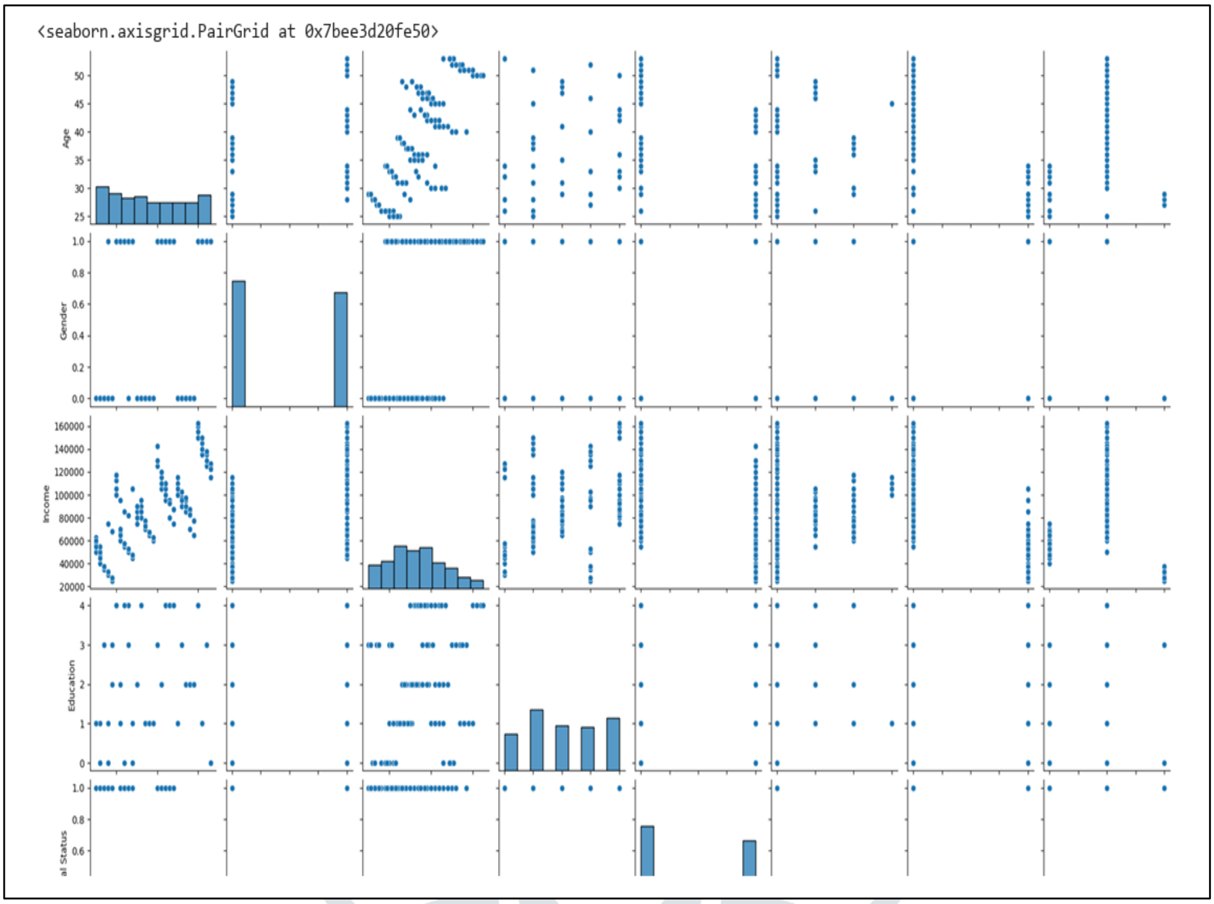


Figure 4: Class distribution of resampling data

The number of instances in each class

The following diagram for each class of three instances labelsLow, Average and High with 69%, 22% and 9% respectively.

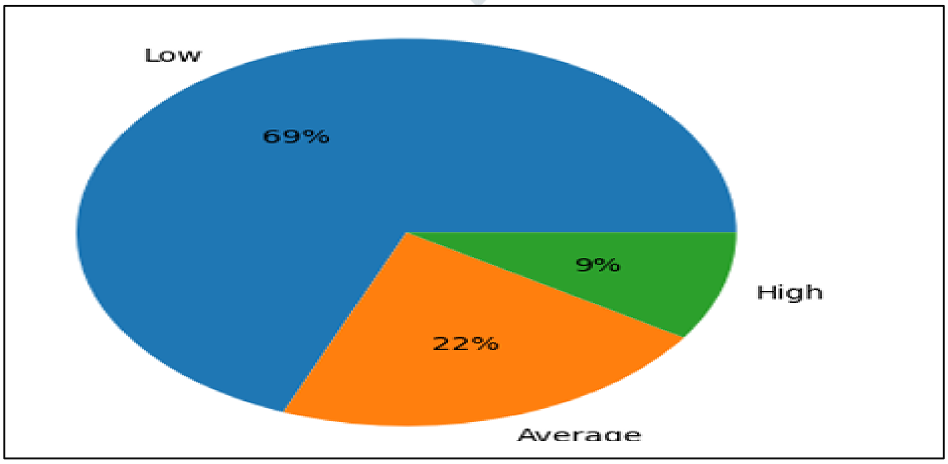
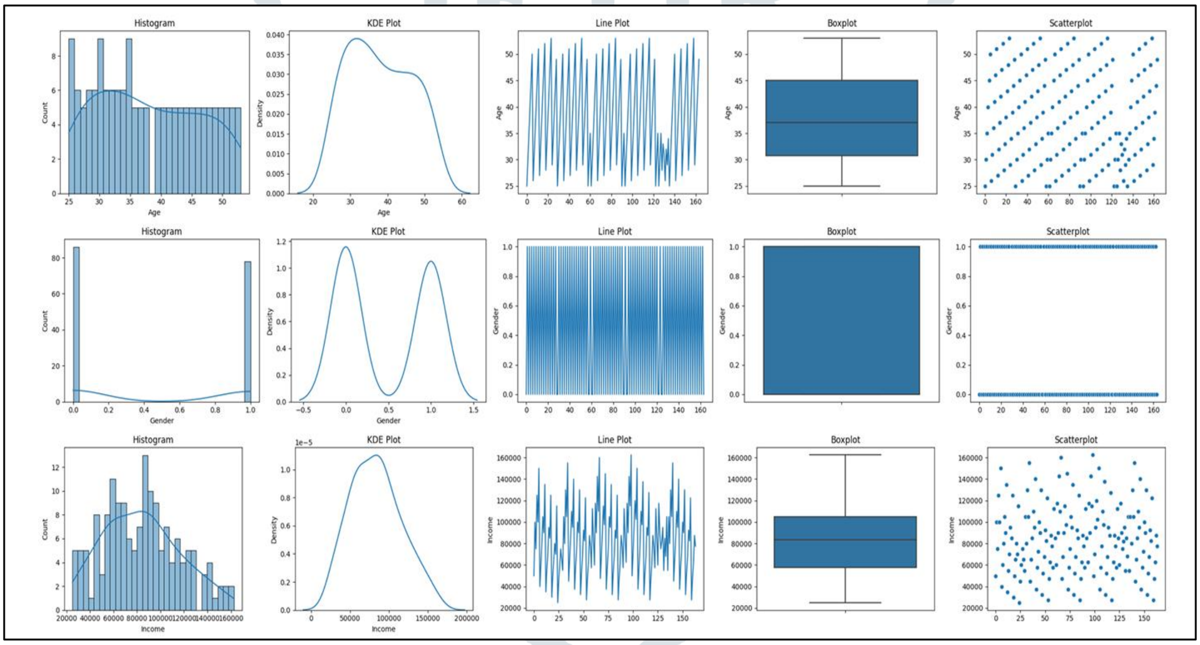
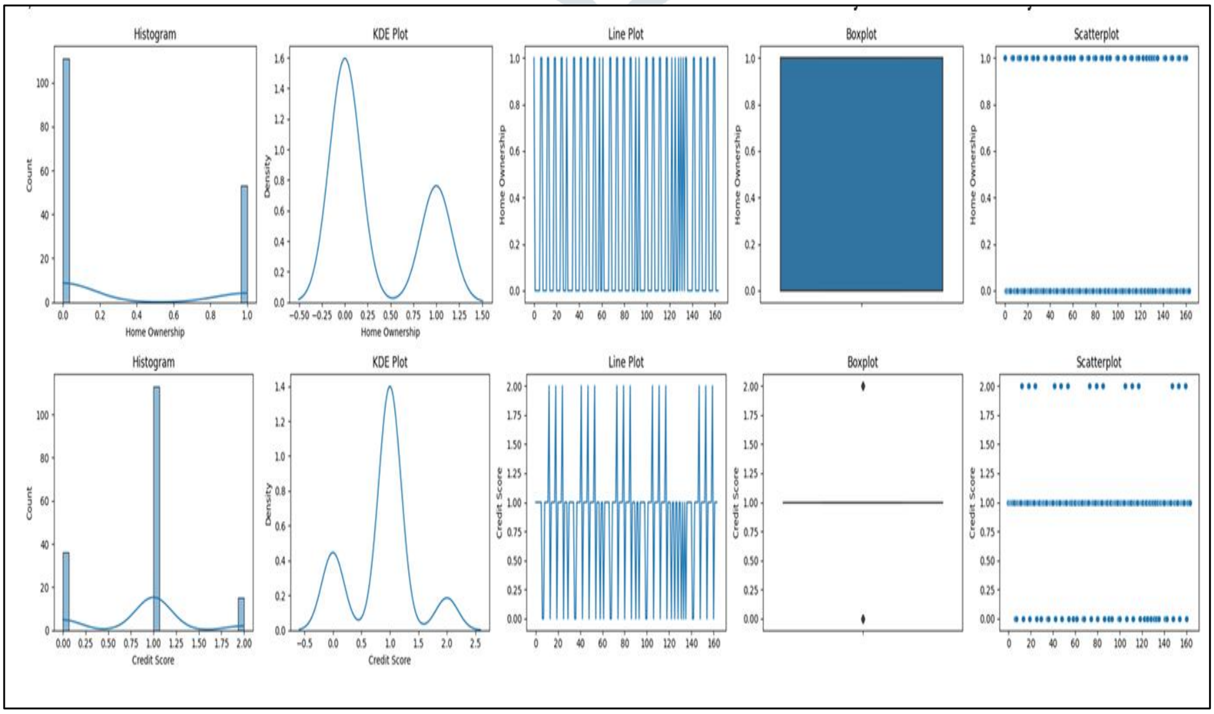


Figure 5: List of Class labels

```
def plots(df, variable):
  if df[variable].dtype != object:
    # define figure size
    fig, ax = plt.subplots(1, 5, figsize=(24, 4))

    # histogram
    sns.histplot(df[variable], bins=30, kde=True, ax=ax[0])
    ax[0].set_title('Histogram')

    # KDE plot
    sns.kdeplot(df[variable], ax=ax[1])
    ax[1].set_title('KDE Plot')

    # Line plot
    sns.lineplot(df[variable], ax=ax[2])
    ax[2].set_title('Line Plot')

    # boxplot
    sns.boxplot(y=df[variable], ax=ax[3])
    ax[3].set_title('Boxplot')

    # scatterplot
    sns.scatterplot(x=df.index, y=df[variable], ax=ax[4])
    ax[4].set_title('Scatterplot')

    plt.tight_layout()
    plt.show()

for i in df.columns:
    plots(df ,i)
```





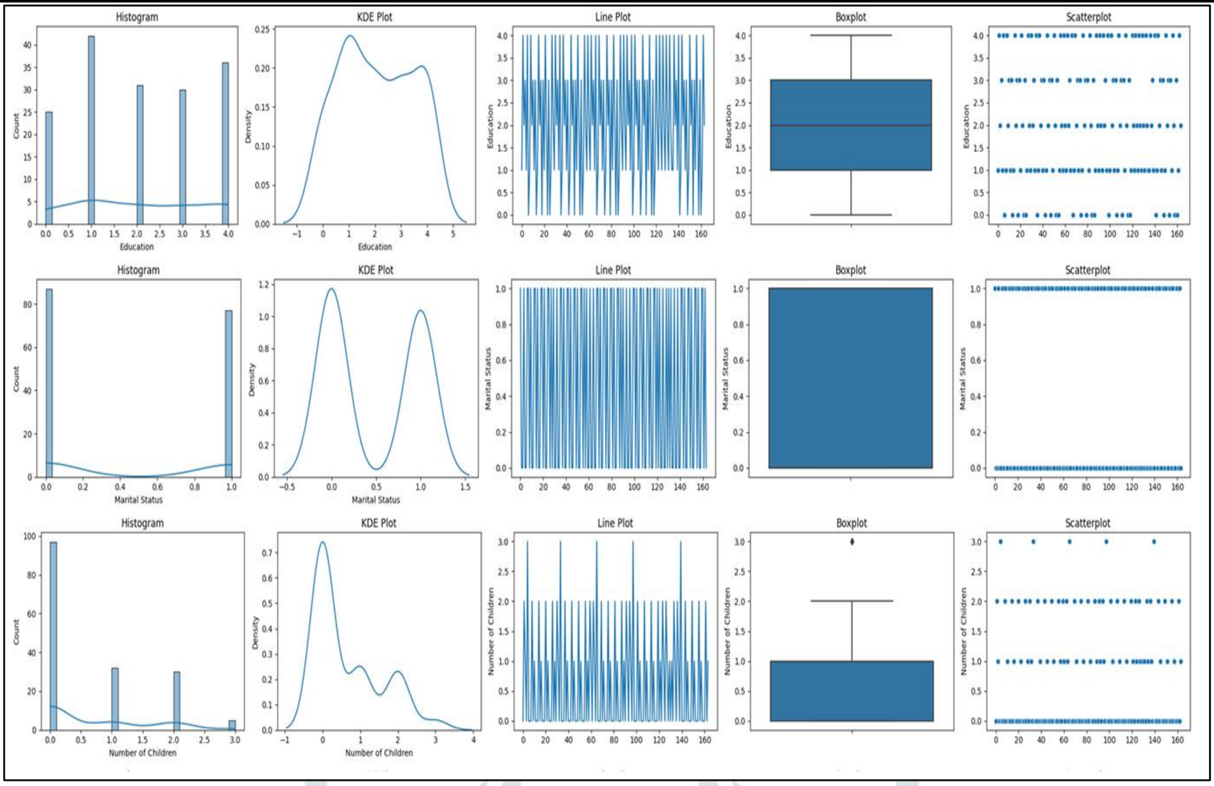


Figure 6: Multiple plots diagram for given credit data

IV. COMPARATIVE STUDY OF DIFFERENT MACHINE LEARNING ALGORITHMS

This chapter focus on comparison between different machine learning algorithms. The following diagram shows the comparison.

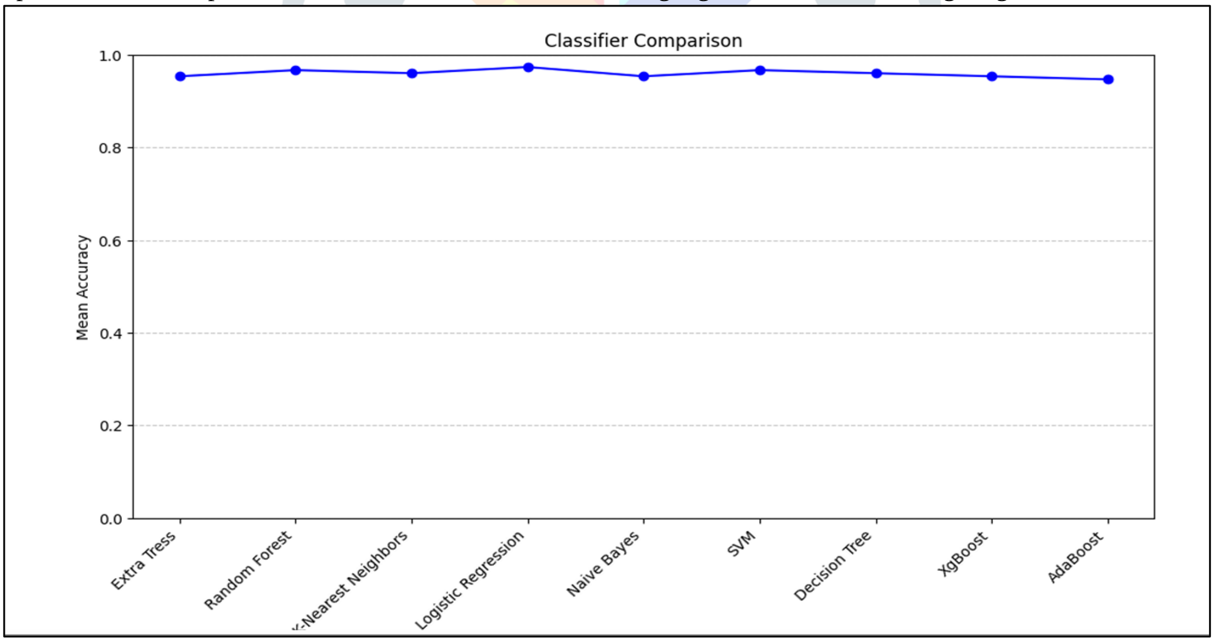


Figure 7: Comparative study of Multiple Machine Learning Algorithms

The following table shows the different attribute values like accuracy, F score, Sensitivityand time of multiple algorithms. Table shows the accuracy, F score, Sensitivity, Recall and time of Random Forest is 96.6%, 96.6%, 95.3% and 2.50 Sec. KNN shows the accuracy, F score, Sensitivity, Recall and time is 96%, 95.9%, 96% and 0.10 Sec. The Logistics Regression shows the accuracy, F score, Sensitivity, Recall and time is 97.3%, 97.3%, 96.6% and 0.28 sec. The naïve Bayes shows the accuracy, F score, Sensitivity, Recall and time is 95.3%, 95.3%, 95.3% and 0.03 sec. The Support Vector Machine shows the accuracy, F score, Sensitivity, Recall and time is 96.6%, 96.6%, 96.6% and 0.06 sec. The Decision tree shows the accuracy, F score, Sensitivity, Recall and time is 96.6%, 95.3%, 96.6% and 0.03 sec. The XGBoost algorithm shows the accuracy, F score, Sensitivity, Recall and time is 95.3%, 95.3%, 95.3% and 0.64 Sec. The AdaBoost algorithm shows the accuracy, F score, Sensitivity, Recall and time is 94.6%, 94.6%, 94.6% and 2.24 Sec.

Table 3: Table for Accuracy, F-Score, Sensitivity, Recall and Time

	Accuracy	F-score	Sensitivity (Recall)	Time (s)
Extra Tress	0.953333	0.952840	0.953333	1.911556
Random Forest	0.966667	0.966482	0.966667	2.502780
K-Nearest Neighbors	0.960000	0.959832	0.960000	0.105187
Logistic Regression	0.973333	0.973165	0.973333	0.289509
Naive Bayes	0.953333	0.953047	0.953333	0.031316
SVM	0.966667	0.966515	0.966667	0.062781
Decision Tree	0.966667	0.953250	0.966667	0.031113
XgBoost	0.953333	0.953149	0.953333	0.641274
AdaBoost	0.946667	0.946330	0.946667	2.241139

Among all machine learning algorithms Logistic Regression is good in accuracy, F score, sensitivity and time also low. For time taken point of view KNN consume less time to finish the work but accuracy is low comparatively Logistic Regression. AdaBoost algorithm perform less compare to all machine learning algorithms and it take huge amount of time for execution of the data classification.

V. CONCLUSION

The research employs a comprehensive dataset comprising diverse financial indicators, and demographic information. By systematically evaluating the performance of each algorithm against this dataset, we aim to discern the strengths and limitations of different approaches in capturing the intricacies of credit risk. Our findings shed light on the most optimal machine learning algorithm for credit score prediction, considering both predictive accuracy and interpretability. The study compares various types of machines learning algorithms.Among all machine learning algorithms Logistic Regression is good in accuracy, F score, sensitivity and time also low. For time taken point of view KNN consume less time to finish the work but accuracy is low comparatively Logistic Regression. AdaBoost algorithm perform less compare to all machine learning algorithms and it take huge amount of time for execution of the data classification.

The following algorithms shows the different attribute values like accuracy, F score, Sensitivityand time of multiple algorithms. Table shows the accuracy, F score, Sensitivity, Recall and time of Random Forest is 96.6%, 96.6%, 95.3% and 2.50 Sec. KNN shows the accuracy, F score, Sensitivity, Recall and time is 96%, 95.9%, 96% and 0.10 Sec. The Logistics Regression shows the accuracy, F score, Sensitivity, Recall and time is 97.3%, 97.3%, 96.6% and 0.28 sec. The naïve Bayes shows the accuracy, F score, Sensitivity, Recall and time is 95.3%, 95.3%, 95.3% and 0.03 sec. The Support Vector Machine shows the accuracy, F score, Sensitivity, Recall and time is 96.6%, 96.6%, 96.6% and 0.06 sec. The Decision tree shows the accuracy, F score, Sensitivity, Recall and time is 96.6%, 95.3%, 96.6% and 0.03 sec. The XGBoost algorithm shows the accuracy, F score, Sensitivity, Recall and time is 95.3%, 95.3%, 95.3% and 0.64 Sec. The AdaBoost algorithm shows the accuracy, F score, Sensitivity, Recall and time is 94.6%, 94.6%, 94.6% and 2.24 Sec.

REFERENCES

1. Rodgers W, Hudson R, Economou F. Modelling credit and investment decisions based on AI algorithmic behavioural pathways. Technological Forecasting and Social Change 2023; 191: 122471.
2. Alaei F, Alaei A, Pal U, Blumenstein M. A comparative study of different texture features for document image retrieval. Expert Systems with Applications 2019; 121: 97–114.
3. Zhang D, Zhou X, Leung SCH, Zheng J. Vertical bagging decision trees model for credit scoring. Expert Systems with Applications 2010; 37(12): 7838–7843.
4. Dumitrescu E, Hué S, Hurlin C, Tokpavi S. Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research 2022; 297(3): 1178–1192.
5. Xia Y, Zhao J, He L, et al. A novel tree-based dynamic heterogeneous ensemble method for credit scoring. Expert Systems with Applications 2020; 159: 113615.
6. Ch. Shravya, Pravallika and Subhani Shaik,"Heart disease prediction using Machine learning Techniques",International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue 6, 2019.
7. Shiva Keertan J and Subhani Shaik," Machine Learning Algorithms for Oil Price Prediction", International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.
8. KP Surya Teja, Vigneshwara Reddy and Subhani Shaik,"Flight Delay Prediction Using Machine Learning Algorithm XGBoost",Jour of Adv Research in Dynamical & Control Systems, Vol. 11, No. 5, 2019.
9. Dr. R. Vijaya Kumar Reddy, Dr. Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao," Breast Cancer Prediction using Classification Techniques", International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.
10. Mr. Sujan Reddy, Ms. Renu Sri andSubhani Shaik," Sentimental Analysis using Logistic Regression", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.
11. Ms. Mamatha, Srinivasa Datta and Subhani Shaik," Fake Profile Identification using Machine Learning Algorithms", International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.
12. R. Vijaya Kumar Reddy, Subhani Shaik,B. Srinivasa Rao, "Machine learning based outlier detection for medical data" Indonesian Journal of Electrical Engineering and Computer Science, Vol. 24, No. 1, October 2021.

13. Dr. Subhani shaik, IdaFann,"performance indicator using machine learning techniques",Dickensian journal, volume 22, issue 6, June, 2022.
14. C.S. Kiran Varma, Shaik Farzaan Ali, Sripada Jishnu, Dr. Shaik Subhani, "Traffic infringement system using machine learning",Dickensian journal, volume 22, issue 6, June, 2022.