ISSN: 0974-5823 Vol. 7 No. 3 March, 2022

International Journal of Mechanical Engineering

Developing a website for a bank's Machine Learning-Based Loan Prediction System

Dr. T C Thomas
Professor, Department of Management studies, Rajalakshmi School of Business, Kuthambakkam, Tamilnadu 600124, India

Dr. J P Sridhar
Associate Professor, Department of Electrical and Electronics Engineering, SJB Institute of Technology, Bangalore, Karnataka 560060, India

Dr. M J Chandrashekar
Associate Professor, Department of Electrical and Electronics Engineering, SJB Institute of Technology, Bangalore, Karnataka 560060, India

Dr. Makarand Upadhyaya
Associate Professor, Department of Management & Marketing, College of Business Administration, University of Bahrain, Bahrain

Dr. Sagaya Aurelia
Assistant Professor, Department of Computer Science, CHRIST University, Bangalore, Karnataka 560029, India

Abstract— Banking is a service that is used by almost everyone irrespective of their financial status. Most of the customers use banks for loan requirements. Bank workers tend to have a lot of work along with the validation of the eligibility of those who apply for loans. This research aims the development a website that can predict the eligibility of the borrower. The dataset which consists of previous records of credentials of the borrower and their eligibility is obtained from Kaggle. The data is then processed through some steps for maximum efficiency of the model. The machine learning models are then tested using the processed data. Three machine learning algorithms like Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and XGBoost were constructed and tested for accuracy. The XGBoost algorithm gives the highest accuracy of 91.6% and it is deployed on a website along with a user interface. This website is beneficial for both banks and borrowers. It reduces the tedious paperwork for the bank workers regarding the validation for eligibility for loans. If there is no such website, the borrowers submit the required papers to the banks and have to wait until /her papers are being processed. This process may even take months some time. It is fine if it turns out to be positive, but if the person is not eligible, then the time he/she wasted in waiting may cost a fortune. The loan procedure can be exhausting if the borrower is in some emergency. Such scenarios can be avoided if a website can predict the eligibility of the borrower bank-wise. The user can look for another bank if he/she doesn't fulfill the eligibility criteria of a particular bank.

Keywords— Banking system, machine learning, SVM, KNN, XGBoost, null value elimination, label encoding, correlation, website designing

I. INTRODUCTION

The banking sector is something that is used by people of almost every financial status. Banks are critical to a country's development because they provide funding for companies. Banks also play an important role in the transmission of monetary policy, which is one of the government's most important tools for supporting economic growth while preventing inflation. It acts as a middleman between people with spare income and others who require it for different business goals. In layman's terms, banks act as a savior in times of need for the people. There are two major services for which customers reach out to a bank. The first one is to save money and the second is to borrow money as a loan. The first service does not come with a lot of terms and conditions as it acts as a piggy bank. But the process of borrowing loans from banks includes tedious paperwork and a lot of credential verifications. However, the procedures for the application for loans have become tough for both bank workers and loan applicants. The workers may delay the process of credential verification of a particular application due to some other work resulting in the delay of the application process. On the other hand, the applicant may not be eligible for a loan from that particular bank but ends up wasting a lot of time waiting for the eligibility verification results. This situation can be avoided if a website can predict the eligibility of the applicant for a particular bank based on the details, he/she enters. This research aims the development such a website. It is done using machine learning and some preprocessing techniques. The processes were elaborated on in the upcoming chapters.

II. LITERATURE SURVEY

Researchers from the University of the West Indies, St. Augustine conducted research on the resilience of the economic growth due to the distress caused by the banking sectors. They stated that the Basel III implementations take shape in response to the Global Financial Crisis, banking sector stability has attracted a lot of attention during the previous decade. Despite the importance of these regulatory and supervisory improvements, they took into account the risk of a financial crisis and looked at the capacity of key banking stability metrics to support economic growth resilience. However, among high-income and middle-income nations, there exist disparities in banking sector liquidity and banking sector regulatory capital. Liquidity supports economic growth, although the benefit is primarily noticeable in high-income nations. They noted that overall financial stability is vital and that regulators and other policymakers must take into consideration the disparities [1]. Another group of researchers from the ISCTE Business school of Portugal stated that the usage of advanced technologies in the banking field is crucial for the security of data. They stated that stakeholders are putting a lot of pressure on traditional financial institutions to adapt to new technology. However, owing to the nature of this industry, data security cannot be compromised. These difficulties make it difficult to make judgments about how to address the challenges of incorporating artificial intelligence (AI), digital transformation, and cybersecurity into the banking sector. To address this problem, they used the DEMATEL technique to construct a realistic decision-support model. The success of a bank, as well as its capacity to recruit new clients and retain existing ones, is directly influenced by its reputation. They concluded that adding reality produces the greatest results [2]. Researchers from the Changchun University of China used the improvised KNN method to acquire higher efficiency. Following a detailed evaluation of the KNN classification algorithm's merits and shortcomings, two revised methods are given in this research to resolve the problem that KNN does not pre-process dataset, characterized by a long classification time and lower accuracy. All of the enhanced algorithms described in their study improve the overall classification effect, as well as classification accuracy and efficiency. The application of optimization algorithms will receive greater study attention in the future. KNN may be used to forecast and analyze data in particular [3]. The KNN algorithm is one of the most powerful algorithms. This statement is backed up by research conducted by a group of researchers at the School of Pedagogical and Technological Education of Greece. They've successfully implemented the KNN algorithm in such a delicate process of classifying environmental sounds. They looked at the issue of noise as a source of pollution and its harmful influence on human activities in their research. Based on these descriptors, a collection of eight discrete noise types was examined and a KNN-based model was trained using a set of test extracts. The categorization technique was employed, and a 70 to 85 percent success rate was reached. They want to examine the feature of Sound Event Detection in the future, which divides a continuous sound stream into distinct segments of sound events. This feature has been out manually in our instance, but attempts are underway to fully automate it [4].

The SVM approach works just as well as the KNN algorithm. A team of Spaniards used the SVM approach to predict geometrical accuracy. SPIF is said to be using SVM algorithms to construct models that can forecast the geometrical precision of molds produced from DX51 aluminized steel sheets. We created our models using this information. The SVM technique using a linear kernel (kappa = 0.80) accurately recognized 90% of the examples; this model was trained using data from the area between the real and theoretical profiles. This resulted in the best outcomes. In the bibliographies, these names are held in great regard. The molds were made using a contour parallel technique, according to the process maps [5]. The application of the method can also improve physical systems. This has been established by researchers from China's Guangxi University. In simulations, the approach has been proven to have a high degree of accuracy in detecting FDIAs. The two methods under consideration are exceptionally dependable, with a detection accuracy of more than 95%. The suggested SVM-based detection approach has an average detection time of 15.3 and 25.6 milliseconds, which corresponds to the CPPS real-time criterion for FDIA detection. One last step is to experiment to ensure that the suggested detection technique can differentiate between normal power system oscillations and FDIA [6]. To identify and forecast data, advanced machine learning algorithms such as XGBOOST are utilized. Researchers from a Chinese electric power provider utilized this system to forecast solar irradiation. Using the ensemble learning capabilities of XGBoost, this approach repeatedly derives the probability prediction results from the predicted values supplied by multiple trees. This approach necessitates less training time and allows for easier parameter changes in trials. This is quite useful in the field of engineering. The proposed model outperforms existing methods in terms of deterministic prediction accuracy on public real-world data sets. The suggested strategy outperforms the previous benchmark methods in terms of prediction accuracy [7].

III. MATERIALS AND METHODS

The data consisting of previous records of basic credentials like loan id, gender, marital status, dependents, education details, employment details, income details, loan amount, credit history, property details, and the status of the loan is obtained from Kaggle in the form of an excel sheet. The further processes done are pictorially represented in figure 1.

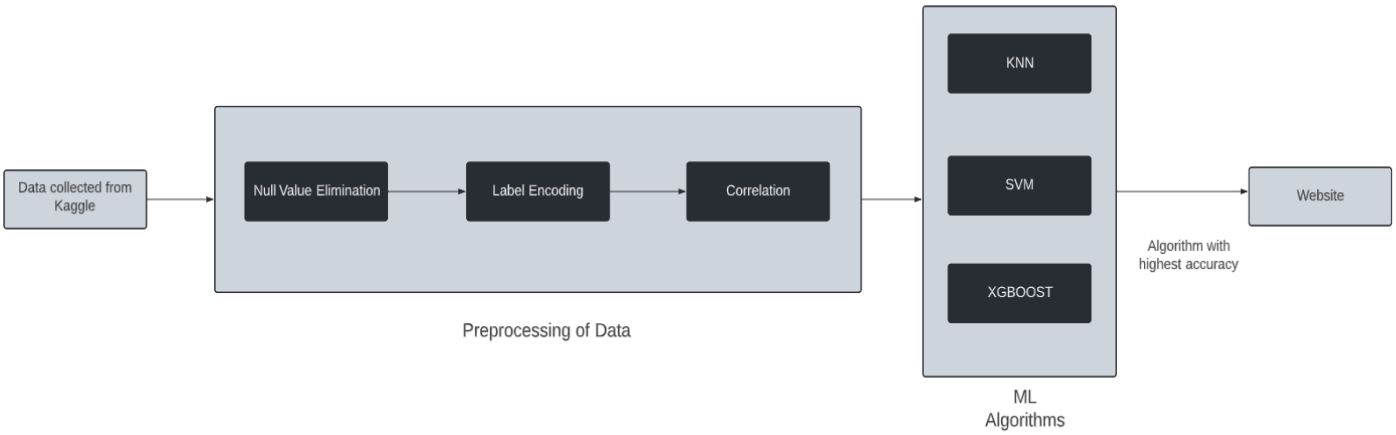

Fig 1. Workflow of the research

As shown in figure 1, the obtained data is then processed using the null value elimination method, the label encoding method, and the correlation method. The processed data is then used to test the machine learning models for higher accuracy. The model with the highest accuracy is deployed on a website in which the user will be entering his/her credentials to check the eligibility criteria of a particular bank.

IV. DATA COLLECTION AND PREPROCESSING

This chapter goes through how the data was gathered. This chapter will also go through the preparation procedures that were utilized on the data.

A. Data Collection:

The data consisting of previous records of basic credentials like loan id, gender, marital status, dependents, education details, employment details, income details, loan amount, credit history, property details, and the status of the loan is obtained from Kaggle in the form of an excel sheet. A glimpse of the obtained dataset is shown in figure 2.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Loan_ID	Gender	Married	Dependents	Education	Self Employed	Applicant Income	Coapplicant Income	Loan Amount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
2	LP001002	Male	No	0	Graduate	No	5849	0		360	1	Urban	Y
3	LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
4	LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
5	LP001006	Male	Yes	0	Not Gradu	No	2583	2358	120	360	1	Urban	Y
6	LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
7	LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
8	LP001013	Male	Yes	0	Not Gradu	No	2333	1516	95	360	1	Urban	Y
9	LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
10	LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
11	LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
12	LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
13	LP001027	Male	Yes	2	Graduate		2500	1840	109	360	1	Urban	Y
14	LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
15	LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
16	LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
17	LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
18	LP001034	Male	No	1	Not Gradu	No	3596	0	100	240		Urban	Y
19	LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
20	LP001038	Male	Yes	0	Not Gradu	No	4887	0	133	360	1	Rural	N

Fig. 2. Obtained dataset

B. Null value elimination:

It can be observed from the data that it contains some cells of the sheet contains null values which can result in improper results when tested. Thus, these null values are eliminated using the statistical elimination method [8]. After the elimination of null values, the data in the sheet will look as shown in figure 3.

Loan_ID	0	Loan_ID	0
Gender	13	Gender	0
Married	3	Married	0
Dependents	15	Dependents	0
Education	0	Education	0
Self_Employed	32	Self_Employed	0
ApplicantIncome	0	ApplicantIncome	0
CoapplicantIncome	0	CoapplicantIncome	0
LoanAmount	22	LoanAmount	0
Loan_Amount_Term	14	Loan_Amount_Term	0
Credit_History	50	Credit_History	0
Property_Area	0	Property_Area	0
Loan_Status	0	Loan_Status	0
dtype: int64		dtype: int64	

Fig. 3. Data after null value elimination

C. Label Encoding:

The data is then converted into categorical values to numerical values using the label encoding method. In supervised machine learning methods, it is mandatory to convert the categorical values to numerical values. Various methods are used to do so. In this research, the label encoding method is used. Label encoding is the process of translating labels into a numeric format so that they may be read by machines [9]. For instance, figure 4 explains the possible values of categories such as marital status, dependents, educational details, etc. By the end of the label encoding process, all the data will be converted into numerical forms for future purposes.

Gender	Unique Entries
Married	Female Male
Dependents	0 1 2 3+
Education	Not Graduate Graduate
Self Employed	Yes No
Property Area	Semiurban Urban Rural
Loan Status	Yes No

Fig. 4. Data Encoding

D. Correlation

Some categories in the obtained data are irrelevant for the validation of eligibility for loans. Processing the entire dataset including those values can cause unnecessary time consumption. It can be avoided by using the correlation method to remove all those needless categories [10]. Figure 5 is an example of the correlation process.

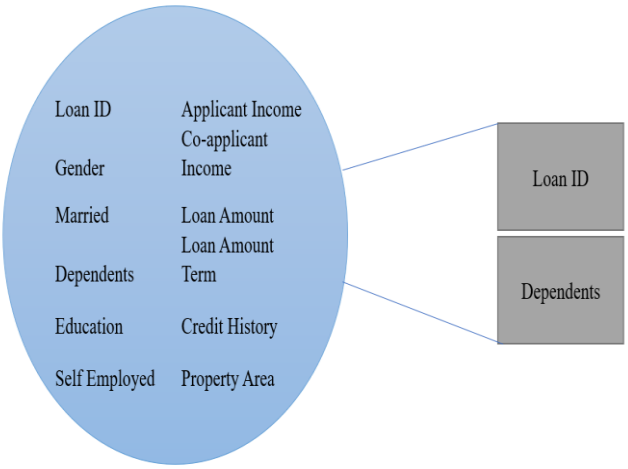

Fig. 5. Correlation

From figure 5, it can be inferred that the categories like the loan Id and the details of the dependents are unnecessary for the loan process. Thus, those values are eliminated from the dataset obtained.

V. CONSTRUCTION OF MACHINE LEARNING MODELS

The main part of the research is the machine learning algorithms. These algorithms act as the backbone of the website. Three machine learning algorithms were constructed and tested for this research. They are the KNN or the K-Nearest Neighbor algorithm, SVM or the Support Vector Machine algorithm, and the XGBOOST or the Extreme Gradient Boosting algorithm. These algorithms are explained further below.

A. KNN

The K-Nearest Neighbor technique is a straightforward classification algorithm used in data mining. It has been in use for quite some time. When dealing with vast volumes of data, the fundamental fault of this technique is that it is inefficient and inaccurate. For a long time, data pretreatment approaches like instance reduction and missing values imputation have been aimed at the KNN algorithm's inefficiency. As a result, the k-nearest neighbor approach is currently used to detect and rectify defective data, such as deleting noisy and redundant samples or imputing missing values [11]. This KNN technique turns huge data into smart data, which is data of adequate quality to produce good results from any data mining algorithm, such as machine learning. These K training samples reflect the unidentified collections' K closest neighbors. The similarity between the collections was measured using distance metrics such as Euclidean. The direction is determined using an actual selection of the k-nearest neighbor approach. Different distance functions, weighting methodologies, or combinations can result in k closest neighbors that aren't the same as each other.

B. SVM

The support vector machine algorithm is a classification technique that may be used to quickly address huge data classification problems. Support vector machines, in particular, may be used to handle multidomain applications in the big data context. To date, this classic machine learning method has shown to be a highly beneficial technology. The support vector machine is a supervised machine learning algorithm. SVM is superior at generalizing problems when it comes to statistical learning. The support vector machine technique makes predictions and judgments simpler using statistical learning theory. Based on a collection of training examples, each of which has been labeled as belonging to one of the many categories, an SVM training approach generates a model that predicts the category of a new example [12]. The support vector machine is one of the easiest algorithms to develop when it comes to picture classification. There are two types of picture categorization in the support vector machine approach. There are two types of patterns: linear and non-linear. Linear patterns are easily recognized or can be easily divided into low-resolution pictures. Non-linear picture patterns are difficult to recognize or can't be readily separated.

C. XGBoost

The term XGBoost is the abbreviation for eXtreme Gradient Boosting. In recent years, the XGBoost algorithm has emerged as a relatively new approach in the field of applied machine learning. It falls under the umbrella of integrated learning. Integrated learning entails creating and merging various learners to perform learning tasks. XGBoost is a Gradient Boosting Judgement Tree-based algorithm. In the XGBoost strategy for training susceptible learners, categorization trees are employed for the majority of the weak learners [13]. During the construction process, new learners are always added depending on the residue mistake from a prior weak learner loop. Following training, it uses weighted summation to build the resulting estimation technique. The gradient is used to build the new trainee in minimizing overall model inaccuracy. Ultimately, a more efficient model is constructed.

VI. RESULT AND DISCUSSION

Kaggle provides data in the form of an excel file containing records of basic credentials such as loan id, gender, marital status, dependents, education details, work details, income details, loan amount, credit history, property details, and loan status. The null values of the dataset are removed using the statistical null value removal method. The data is then converted into numerical form using the label encoding method. The converted data is then correlated to remove all the unnecessary categories. Then three machine learning models were developed using three machine learning algorithms. These algorithms include the KNN algorithm, the SVM algorithm, and the XGBOOST algorithm. The processed data is then used to test the efficiency of the models. The results were converted into confusion matrices for each algorithm for better understanding. A confusion matrix is a 2*2 matrix that has the predicted values on its x-axis and the actual values on the y-axis. When the data are tabulated in this form, it will be easier for the developers to interpret details from the matrix. Figure 6 contains the confusion matrix of the K-Nearest Neighbor algorithm.

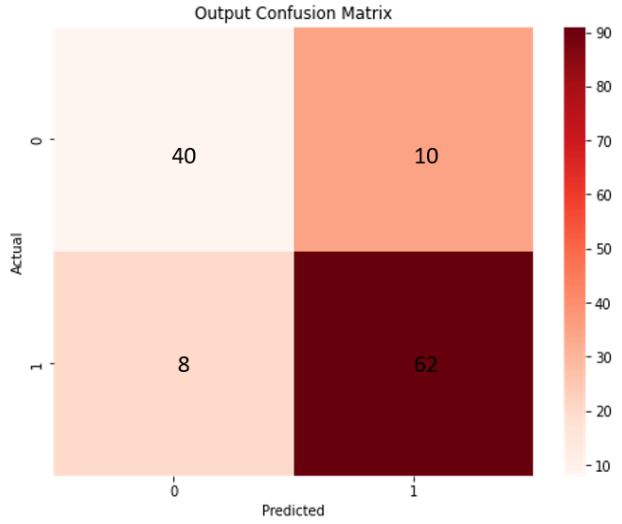

Fig. 6. Confusion matrix for KNN

From figure 6, it can be observed that the KNN algorithm predicted 102 entrees correctly including 40 no values and 62 yes values while it predicted 18 values in the wrong way. Figure 7 is the confusion matrix for the SVM algorithm.

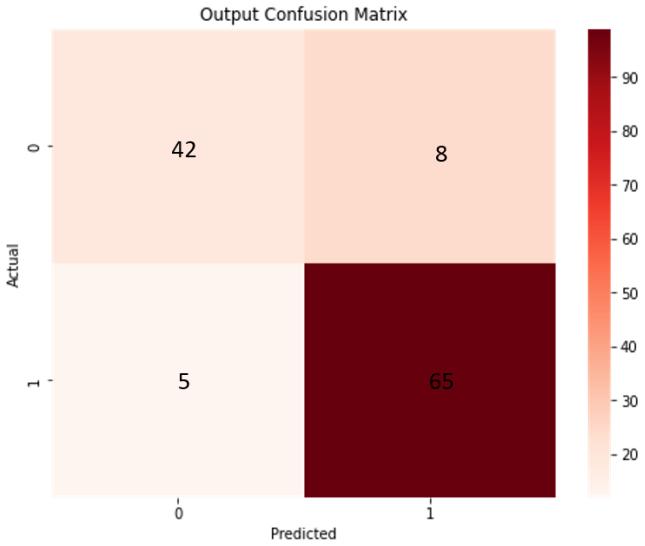

Fig. 7. Confusion matrix for SVM

From figure 7, it can be inferred that the SVM algorithm predicted 107 values correctly and 13 values wrongly. Just like SVM and KNN, the confusion matrix for the XGBOOST algorithm is shown in figure 8.

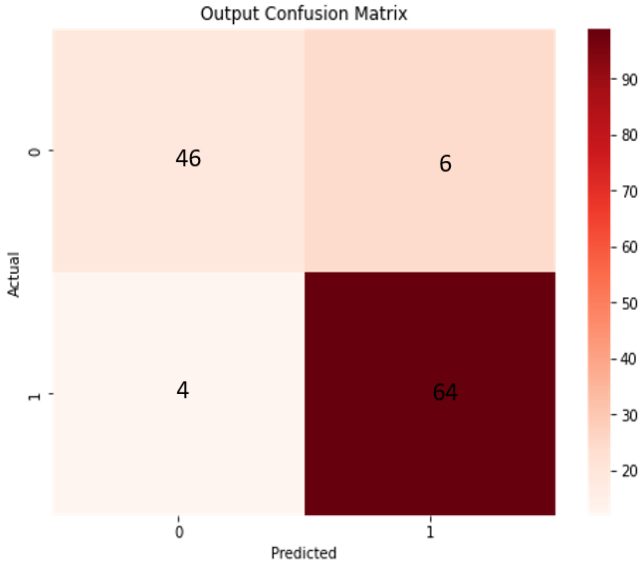

Fig. 8. Confusion matrix for XGBOOST

It can be seen that almost 110 values were predicted correctly by the XGBoost algorithm and it has only 10 wrong predictions. Though the fact that the XGBoost algorithm is more efficient than the other two is seen from the confusion matrix, the values are represented in form of a graph for clearer understanding. Figure 9 consists of the graphical representation of the accuracy value of all three algorithms.

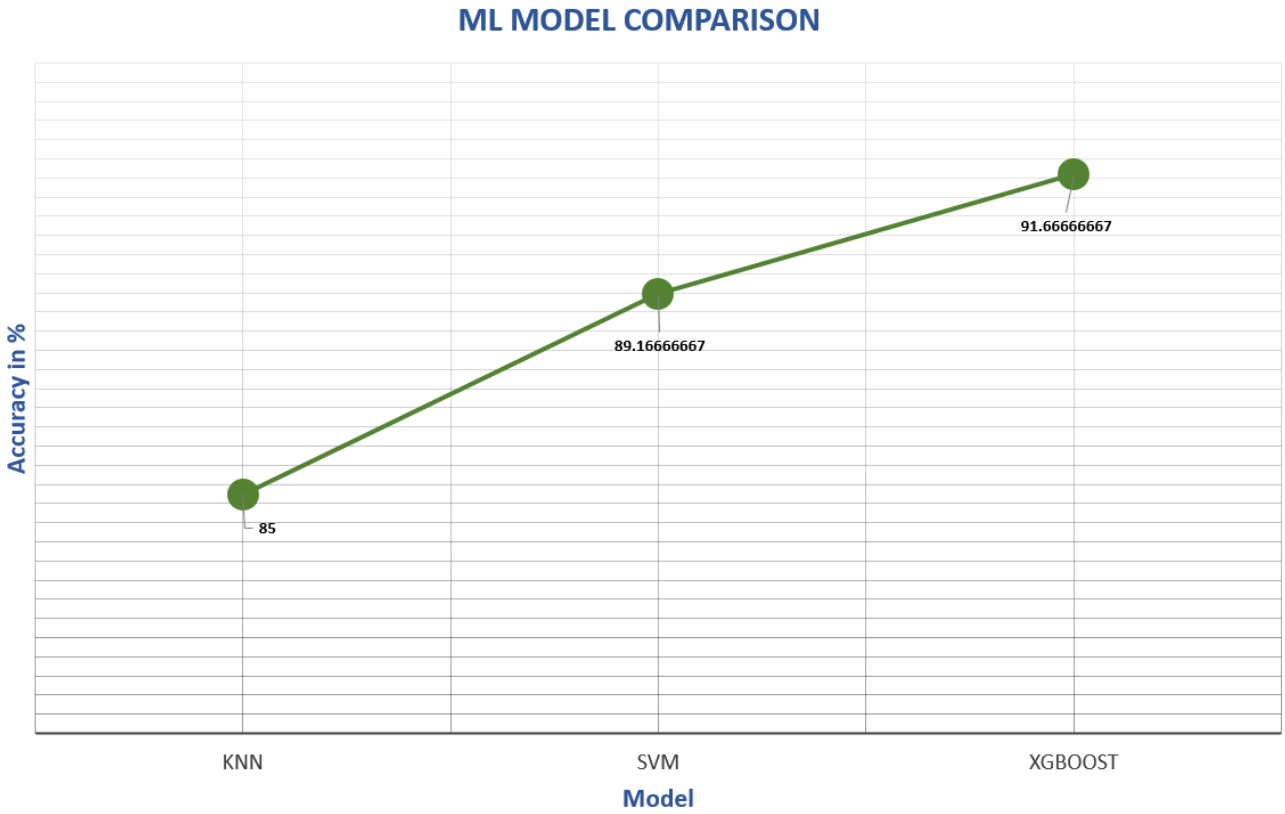

Fig. 9. Comparison of all algorithms

From figure 9, it can be seen that the accuracy percentage of the KNN algorithm is the lowest at 85%. The SVM has the second-largest accuracy with 89.16%. While the XGBOOST algorithm had the highest accuracy percentage of 91.66%. Thus, it can be concluded that the XGBOOST algorithm is the best algorithm to predict the eligibility of the borrower.

A website is designed using the hypertext markup language and it is powered by the XGBOOST algorithm. It allows the user to enter his/her credentials and displays the output of whether he/she fulfills the eligibility criteria of a particular bank. The website looks as shown in figure 10 when the user completes entering all the required values.

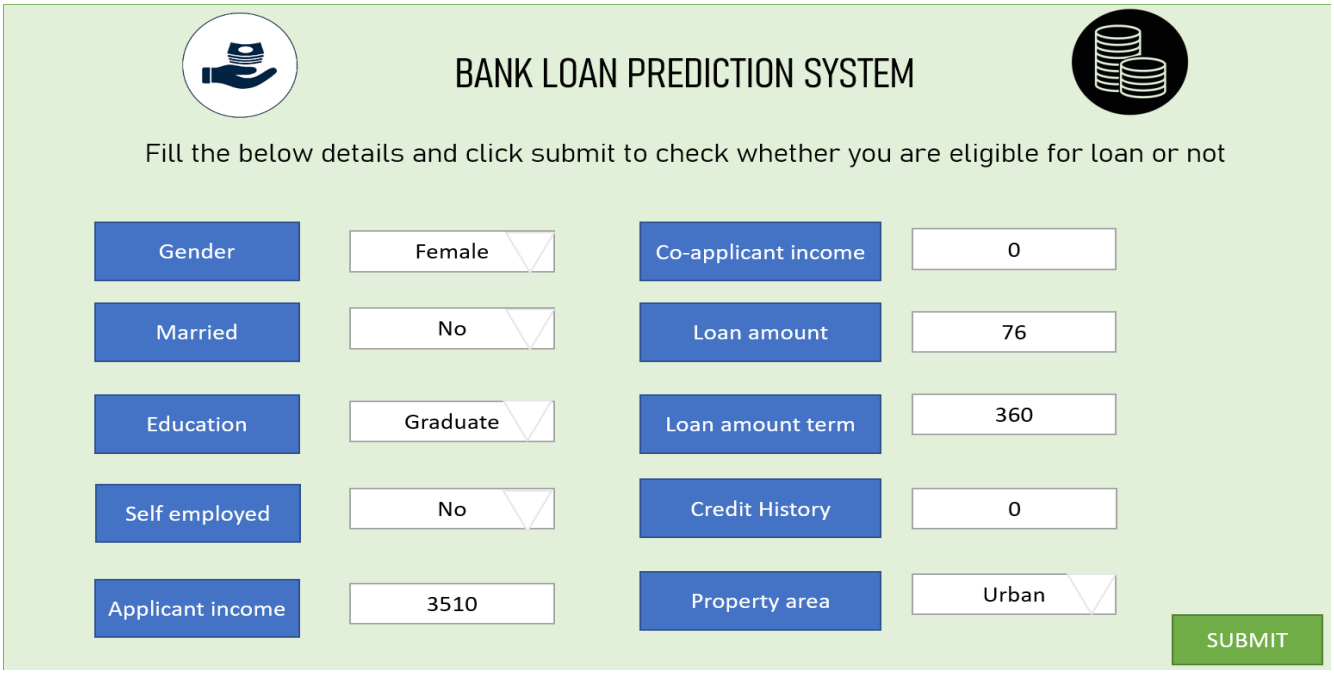

Fig. 10. Home page of the website after the credentials are filled

It can be seen from figure 10 that some of the categories have drop-down lists as inputs like gender, employment details, etc while some categories have textboxes. Credentials like the income of the applicant and loan amount are collected from the user using a textbox. Once the user is done entering the credentials, he clicks on the submit button on the bottom right of the website and it leads to the page where the output is displayed. The sample output page of the website is shown in figure 11.

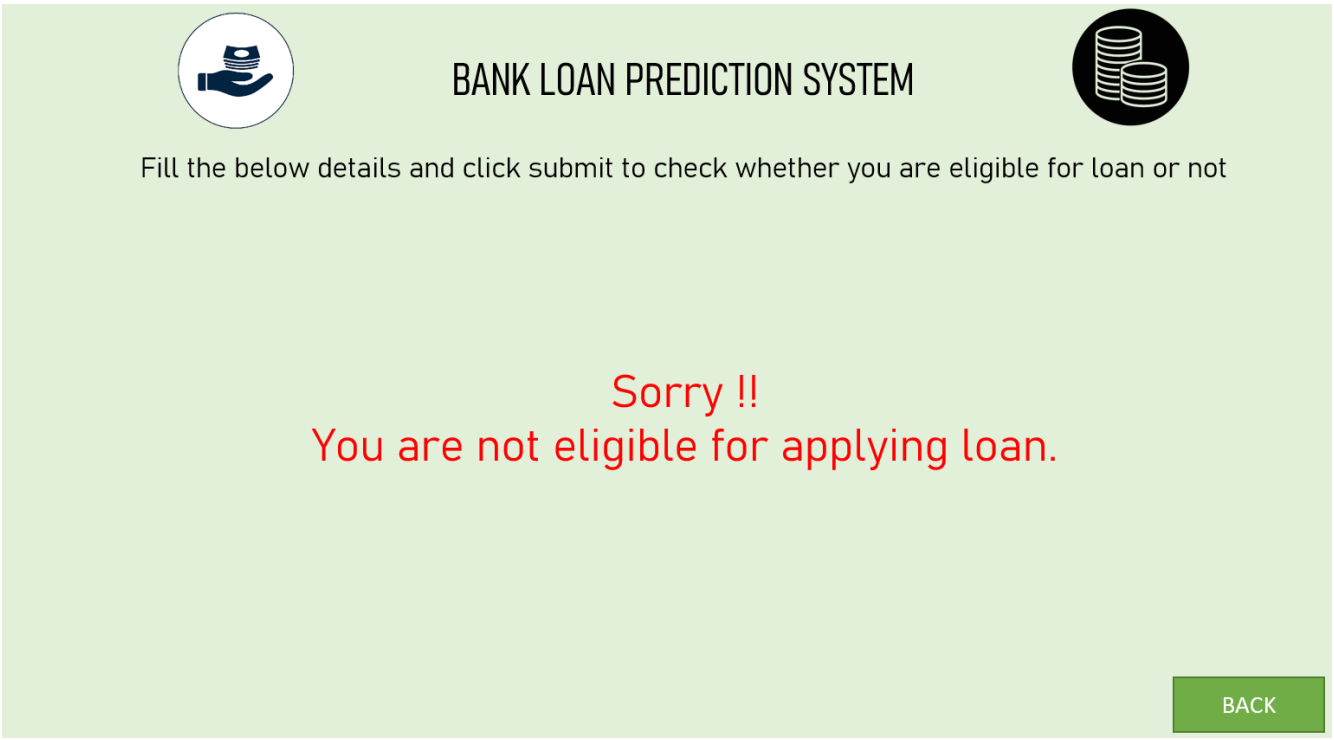

Fig. 11. Output page of the website

VII. CONCLUSION

A complete set of data consisting of all the user credentials along with the loan status of the previous loan applicants is collected in the form of an excel sheet from Kaggle. The data is then processed using certain methods like the statistical null value elimination method, label encoding, and correlation to ensure the highest accuracy. Meanwhile, three machine learning models were developed using the SVM algorithm, KNN algorithm, and the XGBOOST algorithm. The models are then tested using the processed data. The results are obtained in the form of a confusion matrix and an accuracy graph. In the end, it can be found that

the XGBOOST algorithm is the most efficient algorithm that can be used for eligibility prediction. Thus, a website is developed using HTML and the XGBOOST model is deployed on that website. When the website is deployed in real-time applications, it can be seen that it can reduce the wastage of time of the borrowers and it will also reduce the workload of the bank workers.

REFERENCE

[1] Robert Stewart, Murshed Chowdhury, Banking sector distress and economic growth resilience: Asymmetric effects, The Journal of Economic Asymmetries, Volume 24,2021,

[2] Ana Rita D. Rodrigues, Fernando A.F. Ferreira, Fernando J.C.S.N. Teixeira, Constantin Zopounidis, Artificial intelligence, digital transformation and cybersecurity in the banking sector: A multi-stakeholder cognition-driven framework, Research in International Business and Finance, Volume 60,2022,

[3] Haiyan Wang, Peidi Xu, Jinghua Zhao, Improved KNN algorithms of spherical regions based on clustering and region division, Alexandria Engineering Journal, Volume 61, Issue 5,2022,

[4] Eleni Tsalera, Andreas Papadakis, Maria Samarakou, Monitoring, profiling and classification of urban environmental noise using sound characteristics and the KNN algorithm, Energy Reports, Volume 6, Supplement 6, 2020,

[5] Pablo E. Romero, Oscar Rodriguez-Alabanda, Esther Molero, Guillermo Guerrero-Vaca, Use of the support vector machine (SVM) algorithm to predict geometrical accuracy in the manufacture of molds via single point incremental forming (SPIF) using aluminized steel sheets, Journal of Materials Research and Technology, Volume 15, 2021,

[6] Xiaoping Xiong, Siding Hu, Di Sun, Shaolei Hao, Hang Li, Guangyang Lin, Detection of false data injection attack in power information physical system based on SVM–GAB algorithm, Energy Reports, Volume 8, Supplement 5, 2022,

[7] Xianglong Li, Longfei Ma, Ping Chen, Hui Xu, Qijing Xing, Jiahui Yan, Siyue Lu, Haohao Fan, Lei Yang, Yongqiang Cheng, Probabilistic solar irradiance forecasting based on XGBoost, Energy Reports, Volume 8, Supplement 5, 2022,

[8] Jeffery von Ronne , Michael Franz , Niall Dalton , Wolfram Amme
"Compile Time Elimination of Null- and Bounds-Checks "; Third workshop on feedback directed and dynamic optimization

[9] Steven S.W. Lee, Kuang-Yi Li, Alice Chen, Linear random code-based label encoding scheme for label swapping free optical packet switching networks, Optical Switching and Networking, Volume 12,

[10] Houjian Zhao, Xiaowei Li, Yingjie Wu, Xinxin Wu, Friction factor and Nusselt number correlations for forced convection in helical tubes, International Journal of Heat and Mass Transfer, Volume 155, 2020,

[11] Zhou Zhou, Gangquan Si, Haodong Sun, Kai Qu, Weicheng Hou, A robust clustering algorithm based on the identification of core points and KNN kernel density estimation, Expert Systems with Applications, Volume 195, 2022,

[12] Yuhua Sha, Zhenzhi He, Jiawei Du, Zheyingzi Zhu, Xiangning Lu, Intelligent detection technology of flip chip based on H-SVM algorithm, Engineering Failure Analysis, Volume 134, 2022,

[13] Haipei Dong, Fuli Wang, Dakuo He, Yan Liu, The intelligent decision-making of copper flotation backbone process based on CK-XGBoost, Knowledge-Based Systems, Volume 243, 2022,