

PROJECT ON:

LOAN APPROVAL PREDICTION

USING MACHINE LEARNING

(DETAILED PROJECT REPORT)

--BY MANGESH SHINDE

ABSTRACT

Loan prediction could be a quite common real-life problem that every retail bank faces at least once in its lifetime. Banks derive an out sized portion of their profits from loans. Despite the actual fact that several people are applying for loans, it is difficult to seek out a real applicant who will repay the loan. When the process is done manually, plenty of misunderstandings may occur so as to pick out the real applicant. To resolve this problem with banking sector, a loan prediction system supported machine learning is being developed, in order that the system can automatically select the qualified candidates. If done correctly, it can save plenty of man hours at the top of are tail bank. This can be beneficial to both the bank staff and therefore the applicant. The time it takes to sanction a loan is going to be drastically reduced. It is a classification problem where it predicts whether a loan would be approved or not to that particular customer. During a classification problem, discrete values are to be predicted based on a given set of independent variable(s). Banks are looking forward for automating the loan eligibility process (real time) based on customer detail provided while filling online application form to ease the manual work. These details are gender, marital status, education, number of dependents, income, loan amount, credit history and others. To automate this process, related customer segments are identified, those are eligible for loan amount so they will specifically target these customers.

INTRODUCTION

Lending loans is the core business part of almost every banks. The main portion of the bank’s profit is mostly dependent on the profit earned from the loans. Though bank approves loan after a regress process of verification and testimonial but still there is not any surety whether the chosen candidate is that the right person or not. Loan approval isa very important process for banking organizations. This process takes much time while doing it manually. Loan prediction is incredibly helpful for employee of banks like wise as for the applicant also. The aim of this project is to produce quick, immediate and simple way to choose the deserving applicants. It can provide special advantages to the bank. The loan prediction system can automatically calculate the weight of each feature taking part in loan processing and on new test data same features are processed with reference to their associated weight. A limit will be set for the applicant to check whether his/her loan are sanctioned or not. Loan status against particular Loan Id are sent to numerous department of banks so they will choose the real candidates who are applicable for the loan.

MODULE 1:

DATA HANDLING:

* The quantity & quality of your data dictate the

accuracy of our machine learning model.

* The outcome of this module is usually a representation of information which we are going to use for training.
* The test and train dataset contains information about Loan Applicants. There are 12 independent columns and 1 dependent column. This dataset includes attributes like Loan ID, gender, marital status, the level of education, applicant’s income etc.

MODULE 2:

PRE- PROCESSING PHASE:

* Cleaning can require to remove duplicates, correct errors, deal with missing values, Normalization, data type conversions and so on.
* We randomize data, which erases the consequences of the particular order in which we collected and/or otherwise prepared our data.
* Pre processing is used to remove all the lower case, symbols, names, spaces etc. for example any word goes through pre-processing stage and after this word will be processed and converted into English.
* All the NA and Null values are removed.

MODULE 3:

Selecting Model to train the data : Support Vector Machine : In this method, each data item is plotted in an n-dimensional space, where n represents the number of features, and each feature is represented by a set of coordinates. A hyperplane is chosen to distinguish the classes (possibly two) based on their characteristics . All the data that is required to predict loan status is trained applying SVM using the libraries imported.

Naive Bayes : Naive Bayes is a probabilistic algorithm, where the conditional probability is determined regarding the input features. On the other hand, during the dependent input features scenario , conditional probability is calculated twice resulting in improper results. Hence, for better prediction results with respect to Naive Bayes model, independent input features are selected and processed . This predicts the status of granting a loan based on the probability whether the customer can repay the loan or not.

Decision Tree: A decision tree is a type of supervised education algorithm (with a pre-defined target variable) that is commonly used in category problems. In this approach, we divide the population or sample into two or more homogeneous sets (or sub-populations) based on the most significant splitter/differentiator in input variables. Here loan status is set as the pre-defined target variable and it is predicted based on the trainedfeatures.

MODULE 4:

Evaluating the predictions using trained model:

For evaluating the trained model , we use the following features :

* Loan\_ID - The ID number generated by the bank which is giving loan.
* Gender - Whether the person taking loan is male or female.
* Married - Whether the person is married or unmarried.
* Dependents - Family members who stay with the person.
* Education - Educational qualification of the person taking loan.
* Self Employed - Whether the person is self-employed or not.
* ApplicantIncome - The basic salary or income of the applicant per month.
* CoapplicantIncome -The basic income or family members.
* LoanAmount - The amount of loan for which loan is applied.
* Loan\_Amount\_Term - How much time does the loan applicant take to pay the loan.11.
* Credit History - Whether the loan applicant has taken loan previously from same bank
* Property Area - This is about the area where the person stays ( Rural/Urban).
* Target Feature - Loan Status - Loan granted or not (Y, N)

Advantages

1. Improved accuracy: Machine learning algorithms can analyze large amounts of data and identify patterns that humans may miss, leading to more accurate predictions of loan approval.
2. Increased efficiency: Machine learning can automate the loan approval process, reducing the need for manual review and speeding up the approval process.
3. Customized solutions: Machine learning algorithms can be tailored to specific lending criteria, allowing lenders to create customized solutions for borrowers.
4. Reduced risk of bias: By using objective data analysis, machine learning can help reduce the risk of bias in loan approvals.
5. Lack of transparency: The use of machine learning algorithms can make it difficult to understand the decision-making process behind loan approvals.

Disadvantages

1. Limited explain ability: Lenders may find it difficult to explain their loan approval decisions to borrowers if they are based solely on machine learning algorithms.
2. Data accuracy issues: The accuracy of machine learning predictions is only as good as the data it is trained on. If the data is incomplete or inaccurate, the predictions may be similarly flawed.
3. Regulatory compliance: Lenders must ensure that their use of machine learning complies with all relevant regulations, including fair lending laws.

Drawback of Existing System

* Checking details of all applicants consumes lot of time and efforts.
* There is chances of human error may occur due checking all details manually.
* There is possibility of assigning loan to ineligible applicant.

Operating Environment – Hardware and Software

Hardware:

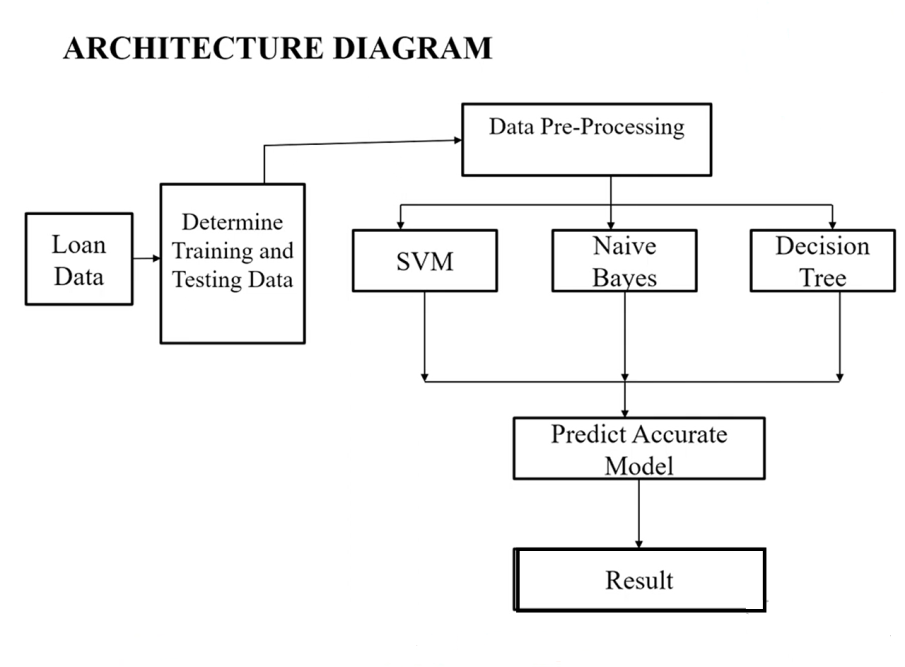
* Computer with minimum 8 GB RAM and a multi-core processor

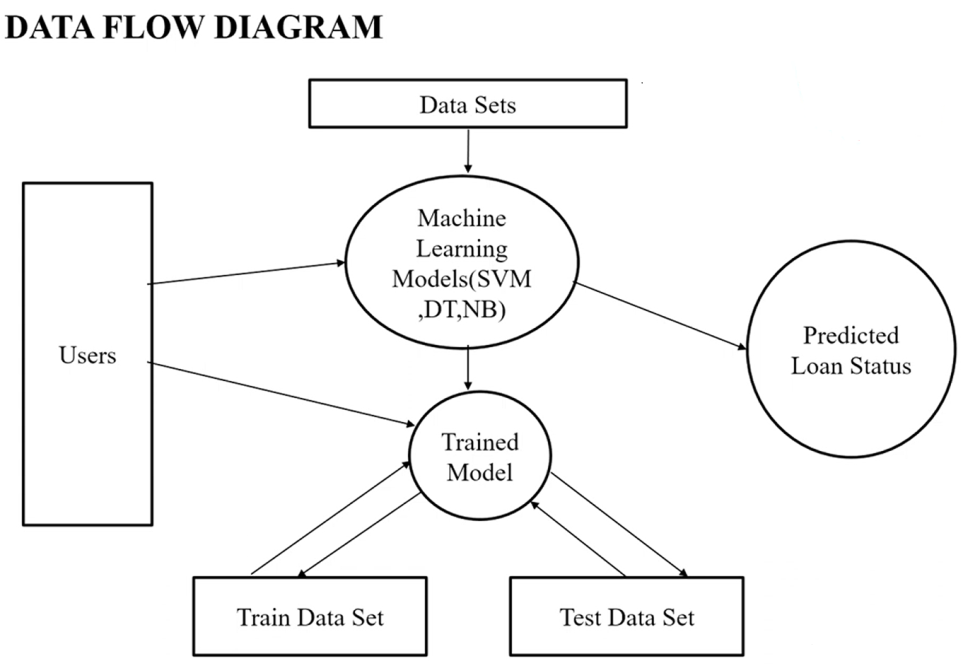
Software:

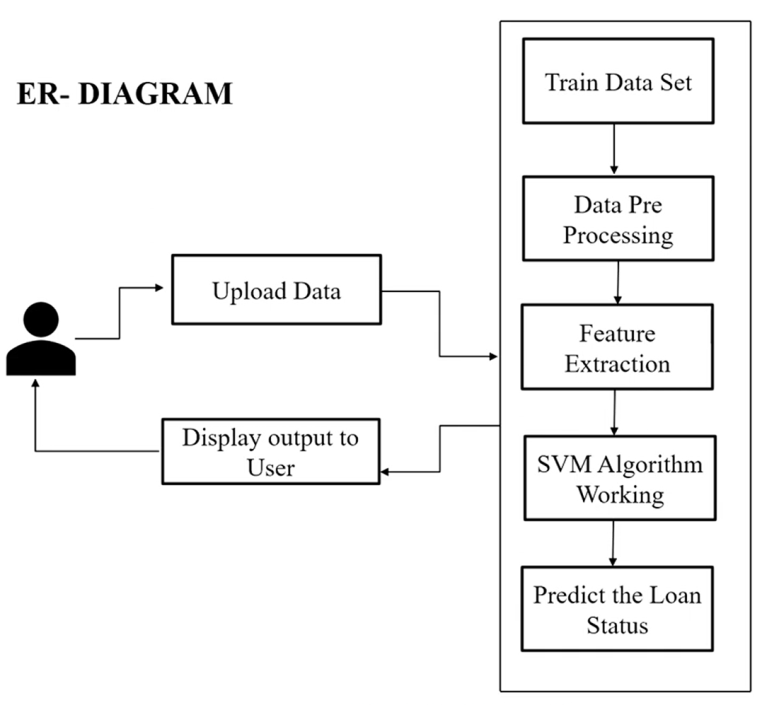
* Operating system: Windows, Linux, or macOS
* Python programming language (preferably v3.0 or later)
* Integrated Development Environment (IDE): Jupyter Notebook ( Anaconda Distribution ).

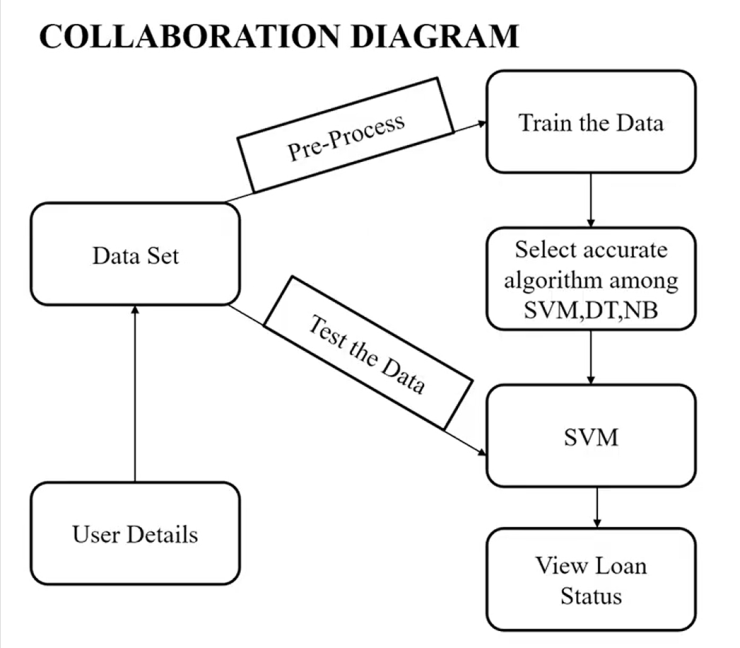
IMPLEMENTATION

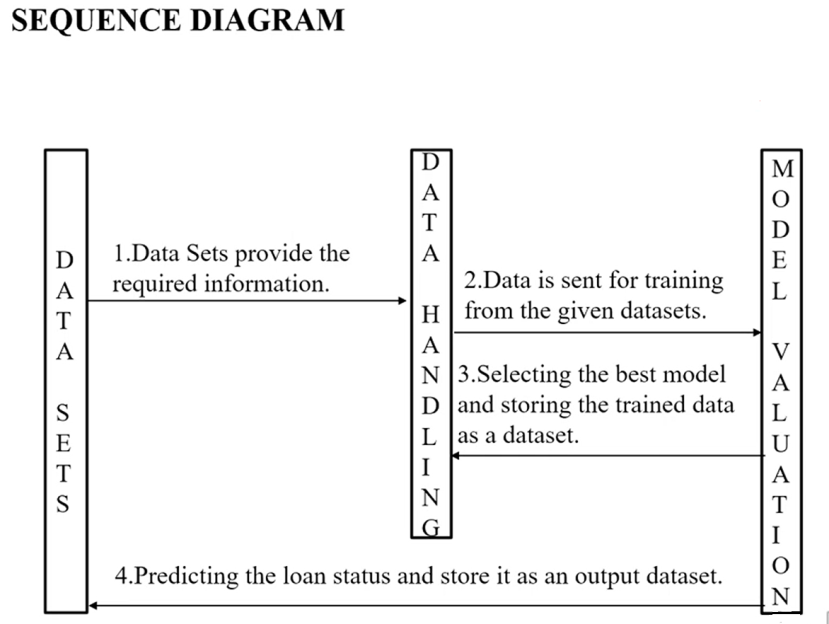
* ARCHITECTURE DIAGRAM
* DATA FLOW DIAGRAM
* ER DIAGRAM
* SEQUENCE DIAGRAM
* COLLABORATION DIAGRAM











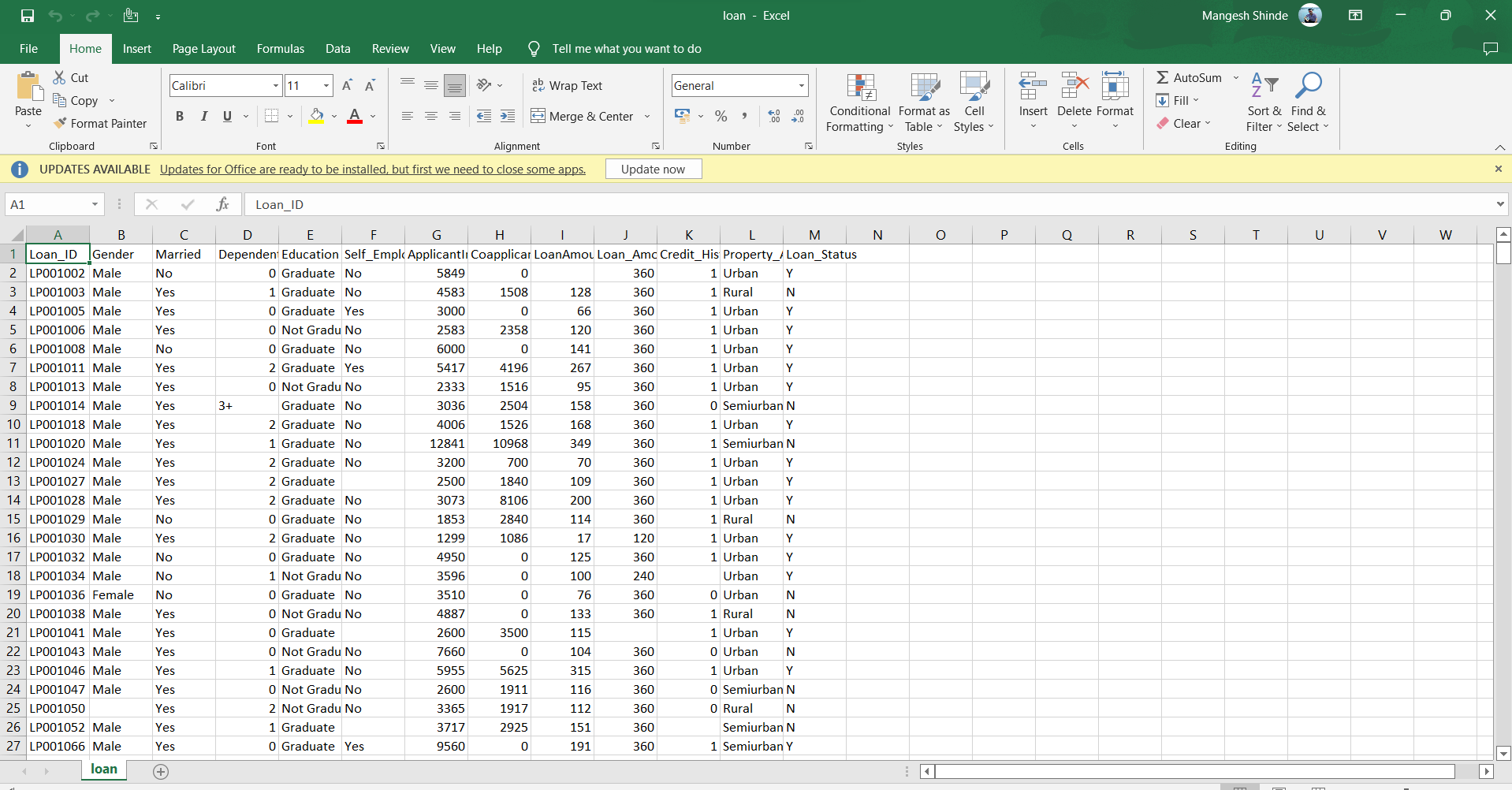
UNIT TESTING:

* Unit testing is usually an automatic test that verifies each unit or, isolated piece of coding functions.
* Unit tests are conducted individually, consisting of segments of code that work together to perform a selected function. Each unit test evaluates the coding written to confirm that it aligns with each said function.
* The quality of unit tests depends on the ability to foresee and properly implement cases that should be within the test suite.
* In this project, each machine learning model performs various functions like transform(), isnull() and other functions. All these functions can be tested individually and test it against the given data.

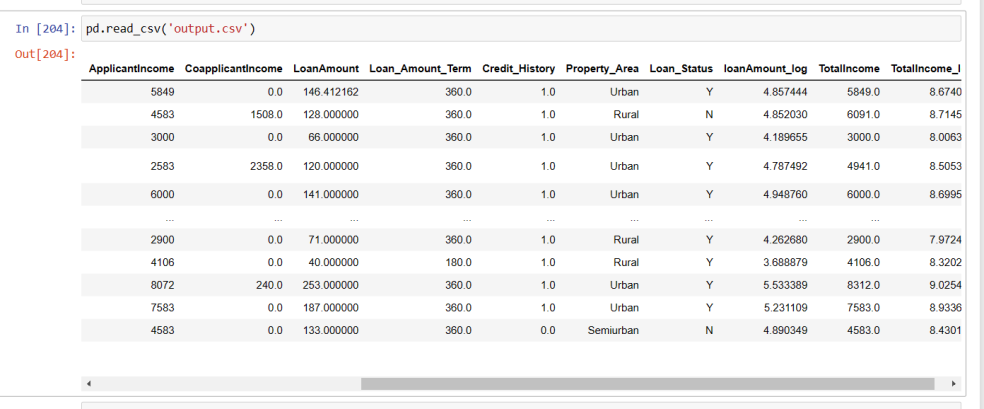
INTEGRATION TESTING:

* Integration Testing is defined as a kind of testing where software modules are integrated logically and tested as a group.
* Under Integration Testing all components that are developed are integrated and validated.
* This project consists of multiple software modules like SVM ,Decision tree and Naïve Bayes.
* The main purpose of integration testing is to combine modules within the application and test as a group to work out that they are working fine.

INPUT DATA SET



OUTPUT DATA SET



CONCLUSION

* From a proper analysis of machine learning algorithms, it can be concluded that this system can be implemented in the real world banking systems to avoid conflicts while approving loans.
* It can be concluded that Support Vector Machine , a supervised machine learning algorithm, has shown more precised accuracy among other algorithms at 73.17 % and hence can be used efficiently to solve banking problems.
* Customer details can also be understood more clearly

with this machine learning models.

* Increase in the digital credit system requires more such techniques to make the banking sector more secure and trustworthy , so that a customer can avail a bank loan with less effort.

Future Enhancements

* This project explores all the machine learning models and chose SVM among them with an accuracy of 73.17 %.
* This can be used in evaluating risk based credit systems. This can further be improved by using different methods of feature selection which can produce much better and accurate results .
* So further we can improve this model by combining different machine learning algorithms using different methods of ensembling like stacking , bagging and more.

