**Project Name:Loan Approval System using ML**

**SDLC (Software Development Life Cycle)**

The Software Development Life Cycle is a systematic process for building software that ensures the quality and correctness of the software built. SDLC process aims to produce high quality software which meets customer expectations. The software development should be completed within the pre-defined time frame and cost.

**SDLC Phases**

The entire SDLC process is divided into the following stages

● Phase 1: Requirement collection and analysis

● Phase 2: Feasibility study

● Phase 3: Design

● Phase 4: Coding

● Phase 5: Testing

● Phase 6: Installation/Deployment

● Phase 7: Maintenance

**Platform Knowledge**

**What is data science?**

Data science is the study of data to extract meaningful insights for business.It is a multi disciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence and computer engineering to analyze large amounts of data. This analysis helps data scientists to ask and answer questions like what happened, why it happened, what will happen, and what can be done with the results.

**Importance of data science**

Solves real problem using data.Modern organizations have lots of data.Online systems and payment portals capture more data in the fields of e-commerce, medicine, finance, and every other aspect of human life. We can process the information and predict the results

**OBJECTIVE:**

As the needs of people are increasing, the demand for loans in banks is also frequently getting higher every day. Banks typically process an applicant's loan after screening and verifying the applicant's eligibility, which is a difficult and time-consuming process. In some cases, some applicants default and banks lose capital. The machine learning approach is ideal for reducing human effort and effective decision making in the loan approval process by implementing machine learning tools that use classification algorithms to predict eligible loan applicants.

**FEATURE ENGINEERING:**

Transformation: All the categorical variables are stored in a new dataframe. One hot encoding:Dummy variables are created for all categorical variables using get\_dummies function from pandas and all encoded columns are created. Taking the best estimators, training algorithm is performed on Test data.

The min-max scaling method is used for numerical features.

Min-max scaling is similar to z-score normalization in that it will replace every value in a column with a new value using a formula. In this case, that formula is:

*m = (x -xmin) / (xmax -xmin)*

Where:

• *m* is our new value

• *x* is the original cell value

• *xmin* is the minimum value of the column

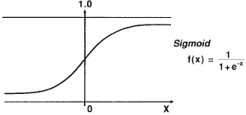
• *xmax* is the maximum value of the column

Using this formula, we will see that the values of each column will now be between zero and one.

**Machine Learning Algorithms:**

**1)Logistic Regression:**

It is one of the most simplest and popular machine learning model.



This model takes the input values as x and gives the output values as f(x) i.e. 0 or 1. If I need to built a machine learning model then each data point of independent variable will be x (i.e. sum of x1 \* w1 + x2 \* w2 . . . .so on ) and this will give a value that is between 0 to 1. If I consider that 0.50 as deciding value or threshold. Then any result above 0.5 would be taken as 1 and below that as 0.

**2.Decision Tree**

The first step is to split the labelled data into train and test data.The tree begins with the root node which consists of the entire training data.The best attribute is found using the Attribute selection measure.Entropy: (chances of being incorrect).Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. When the data is highly impure, or highly pure, the entropy is 0. Entropy can be calculated as:

Entropy(s)= -P(Yes) \* log2 P(Yes) - P(No) \* log2 P(No)

Entropy(S) = 1 when P(Yes) = P(No) = 0.5.

Entropy(S) = 0 when P(Yes) = 1 or 0

Information Gain:

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.It calculates how much information a feature provides us with a class.According to the value of information gain, we split the node and build the decision tree.A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain = Entropy(S)- [(Weighted Avg) \*Entropy(each feature)]

Select the feature with maximum information gain and divide the dataset. Recursively make new decision trees using the subsets of the dataset created in above step.Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

**3.Random Forest**

A random forest algorithm is a supervised learning algorithm that can be used for both classification and regression tasks. It is an ensemble learning method, which means that it combines the predictions of multiple individual models to produce a more accurate prediction.

Random forest algorithms work by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

Random forest algorithms are more robust to overfitting than individual decision trees because they average the predictions of multiple trees. This is because each tree in the forest is trained on a different bootstrap sample of the training data and a different random subset of features.

**Hardware Requirements:**

• CPU: A modern multicore processor, such as an Intel Core i5 or i7, is recommended for training and deploying machine learning models.

• RAM: A minimum of 8GB of RAM is recommended, although more may be required for larger datasets.

• GPU: A powerful graphics processing unit (GPU) can significantly speed up training and inference of deep learning models. Nvidia GPUs are the most commonly used for machine learning tasks.

• Storage: A large amount of storage is required to store and process large datasets. Solid-state drives (SSDs) are recommended for fast access to data.

**Software Requirements:**

• Operating System: Most machine learning frameworks are compatible with popular operating systems such as Windows, macOS, and Linux.

• Python: Python is the most commonly used programming language for machine learning, and most machine learning frameworks are built using Python. • Machine Learning Frameworks: There are many machine learning frameworks available, such as TensorFlow, PyTorch, Keras, and scikit-learn. These frameworks provide tools and APIs for building, training, and deploying machine learning models.

• Development Tools: IDEs such as PyCharm, Jupyter Notebook, and Visual Studio Code are commonly used for machine learning development.

• Libraries: Libraries such as NumPy, Pandas, and Matplotlib are commonly used for data manipulation, analysis, and visualization.

In summary, to get started with your project, you will need a computer with a powerful CPU and GPU, a minimum of 8GB of RAM, and large storage capacity. You will also need to install Python and various machine learning libraries and frameworks.

**Existing System:**

Borrowers use a loan application to qualify for a mortgage. The model can anticipate outcomes and is quickly adaptable to a wide range of inputs. Also, this strategy saves the banking industry and its staff a significant amount of time.Existing systems for Loan Approval Systems typically rely on traditional rule-based systems. These systems are manually designed by Bank officials to decide whether to provide loan to a candidate or not based on a set of predefined rules.In the present system ,the loan is provided to the loan seeker if he has high Cibil score, decent assets value,good education level and standard source of Income.Data Science Algorithms and techniques are not used in Existing system. Rule-based systems are not effective in all real world scenarios, while they have a number of limitations. First, they can be difficult to maintain and update as new Loan schemes emerge all the time.

**Proposed system:**

The proposed system for Loan approval system will use machine learning to develop a model that can accurately predict whether to provide loan to a loan seeker or not. The model will be trained on a dataset of historical loan details data which includes Education level,Cibil score,Annual income,additional income,loan amount demanded by the candidate,loan term in years,residential asssets value,commercial assets value and employment status.Once trained, the model will be able to predict the probability to provide loan to the loan seeker such as Bank and its Loan policies.The proposed system offers a number of benefits over existing systems, including:

* Improved accuracy: The proposed system is expected to be more accurate in detecting fraudulent transactions than rule-based systems, as it is able to learn from data and adapt to changing fraud patterns.
* Reduced false positives: The proposed system is expected to produce fewer false positives than rule-based systems, as it is able to consider a wider range of factors when evaluating a transaction.
* Scalability: The proposed system is scalable to handle large volumes of transaction data.
* Flexibility: The proposed system can be customized to meet the specific needs of a financial institution or payment processor.

Machine Learning algorithms are implemented in the Loan approval system project where we train multiple models and predict the output.We evaluate multiple evaluation metrics like test accuracy,false positives,false negatives,bias,variance,f1 score,recall score,precision value,cohen kappa score,ROC AUC score.

**Conclusion:**

The ML model which has high test accuracy,low bias,low variance,least false negatives,ideal precision,recall,f1 score and ROC-AUC score close to 1 is the best model for our project. Random Forest algorithm is the most effective model.