**ELIGIBILITY PREDICTION FOR LOAN** MODEL ON ELIGIBILITY PREDICTION FOR LOAN DATA SET :

**ABSTRACT:**

The prediction of loan eligibility is a crucial task in the lending industry, helping financial institutions determine the likelihood of approving loan applications. This research proposes a machine learning-based approach to predict loan eligibility, leveraging historical loan data and applicant information. The objective is to develop a reliable and accurate model that can assist lenders in making informed decisions efficiently.The study begins by assembling a comprehensive dataset containing various borrower attributes such as credit score, income, employment history, debt-to-income ratio, and demographic information. The dataset is then preprocessed to handle missing values, normalize variables, and eliminate any inconsistencies.

Next, a range of supervised machine learning algorithms, including logistic regression, decision trees, support vector machines, or neural networks, are employed to train predictive models using the preprocessed dataset. Multiple evaluation metrics, such as accuracy, precision, recall,F1 Score, are employed to assess the models' performance and identify the most effective algorithm.

In conclusion, this research presents a machine learning-driven approach for predicting loan eligibility. By leveraging historical loan data and applicant information, the proposed model provides lenders with a reliable tool to assess loan applications efficiently. The integration of this model into existing loan processing systems has the potential to revolutionize the lending industry by improving decision-making processes.

**Features:**

Loan\_ID ApplicantIncome

Gender CoapplicantIncome

Married LoanAmount

Dependents Loan\_Amount\_Term

Education Credit\_History

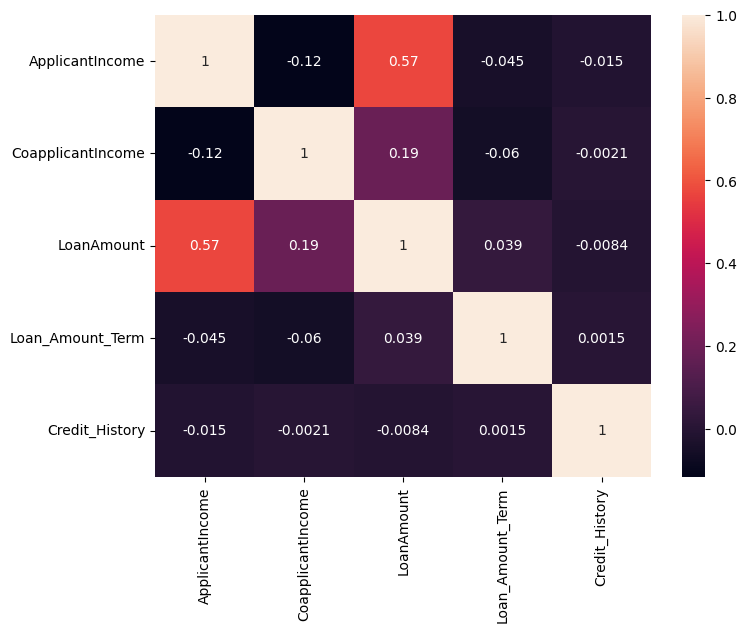
Self\_Empployed Property\_Area

Loan\_status

Diagram

Description automatically generated

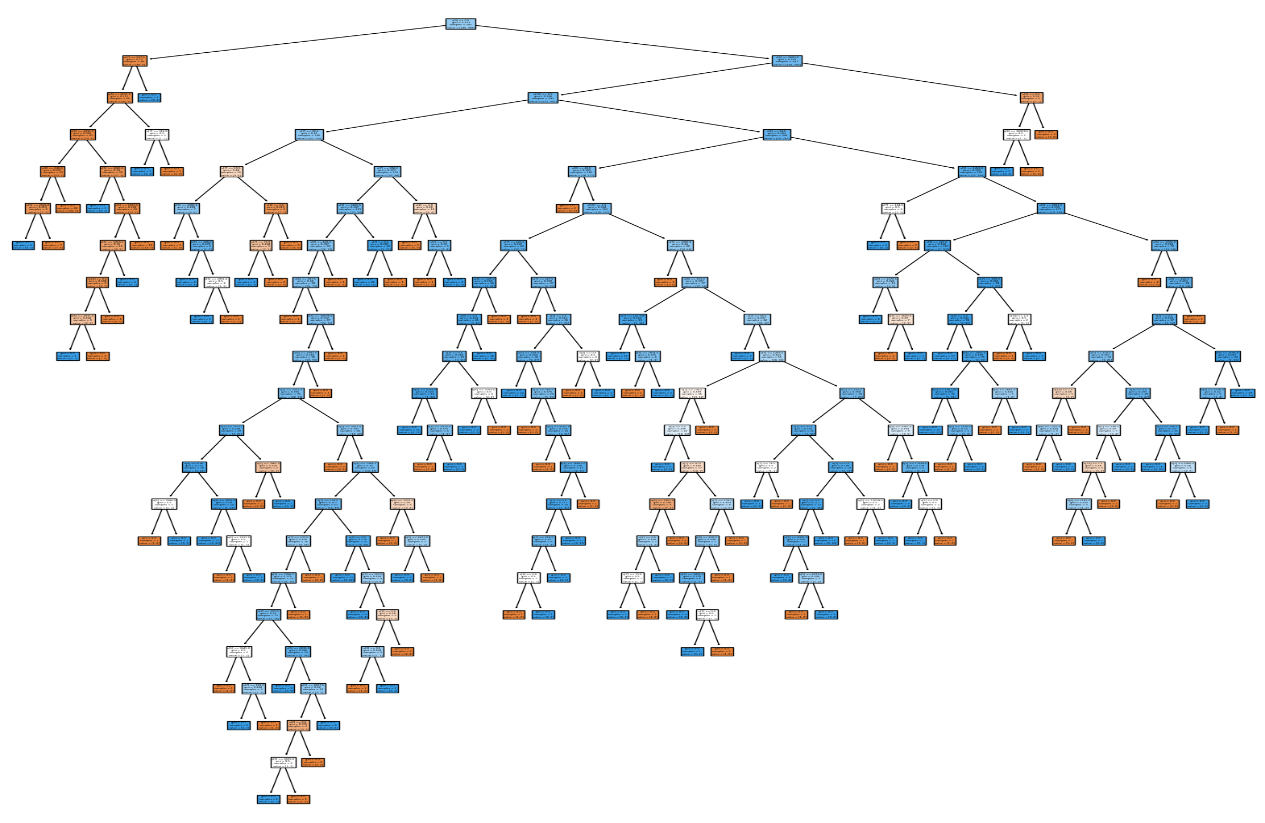
**Corelation matrix:** We can observe and say that no chance of dimensionality reduction since we cannot pick out any attribute.

****

**Algorithms Usage:**

### Decision Tree: Decision Tree algorithm in machine learning is one of the most popular algorithms in use today; this is a supervised learning algorithm that is used for classifying problems. It works well classifying for both categorical and continuous dependent variables. In this algorithm, we split the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

### **Accuracy = 0.846154 = 84%**



* **KNN (K- Nearest Neighbours) Algorithm:** This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It’s a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbours. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement. KNN can be easily understood by comparing it to real life.

For example, if you want information about a person, it makes sense to talk to his or her friends and colleagues.

Things to consider before selecting K Nearest Neighbours Algorithm:

1. KNN is computationally expensive.
2. Variables should be normalized, or else higher range variables can bias the algorithm.
3. Data still needs to be pre-processed.

Accuracy = 0.692308 = 69%

* **Naive Bayes Algorithm:** A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features are related to each other, a Naive Bayes classifier would consider all these properties independently when calculating the probability of a particular outcome. A Naive Bayesian model is easy to build and useful for massive datasets. It's simple and is known to outperform even highly sophisticated classification methods.
* Accuracy = 0.692308 = 69%.

### Random Forest Algorithm: A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

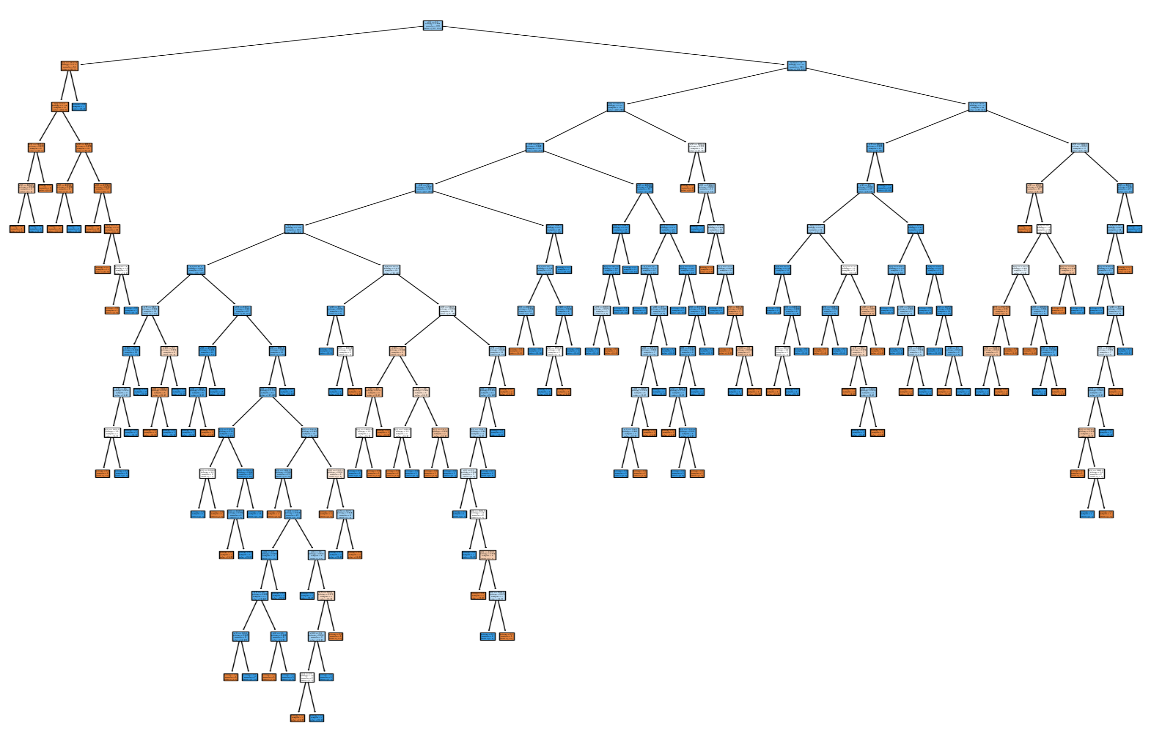
### Each tree is planted & grown as follows:

### If the number of cases in the training set is N, then a sample of N cases is taken at random. This sample will be the training set for growing the tree.

### If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M, and the best split on this m is used to split the node. The value of m is held constant during this process.

### Each tree is grown to the most substantial extent possible. There is no pruning.

### Accuracy = 0.849153 = 84%

****

### Logistic Regression: Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression.

### These methods listed below are often used to help improve logistic regression models:

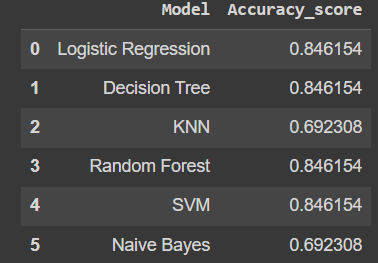
1. Include interaction terms.
2. Eliminate features.
3. Regularize techniques.
4. Use a non-linear model.

* Accuracy = 0.846154 = 84%
* **SVM (Support Vector Machine) Algorithm:** SVM algorithm is a method of classification algorithm in which you plot raw data as points in an n-dimensional space (where n is the number of features you have).
* Accuracy = 0.846154 = 84%

**CONCLUSION**:

The best model is to predict the eligibility for loan is that logistic regression ,decision tree,random forest and svm which is having the

accuracy of 84%.

****