Introduction:

Machine Learning (ML) pertains to the ability of data-driven models to “learn” information about a system directly from observed data without predetermining mechanistic relationships that govern the system. ML algorithms are able to adaptively improve their performance with each new data sample and discover hidden patterns in complex heterogeneous and high dimensional data. In different engineering domain ML offers predictive models, such as Decision Trees (DTs), Random Forests (RFs), Support Vector Machines (SVMs), etc. Which are able to map highly non-linear heterogeneous input and output patterns even when physiological relationships between model variables could not be determined due to complexity, pathologies, or lack of biological understanding. They cope with missing values and are able to combine heterogeneous data types into a single model, whilst also performing an automatic principal feature selection. Combining multiple Decision Trees (DTs) in a Random Forests (RFs) maintains this interpretability, but offers state-of-the-art prediction accuracies. Machine learning is a branch of artificial intelligence that enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed. Two types of learning technique are there in machine learning Supervised Learning and Unsupervised Learning.

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. Decision tree consists of two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

As in the past couple of decades, the decision making for the financial lending has been very much influenced by the information sharing and technological advancements. The technique of credit scoring is to evaluate different credit attributes by analyzing and classification to an individual and enterprise profile to assess the credit decision or to estimate the creditworthiness Only credit scoring is not sufficient for the financial lending because of such a massive number of loan defaulters. As financial analysts not only rely on the credit scores but also on their experience regarding the historical successful and unsuccessful cases as well for better decision making. Moreover, with such tremendous growth of the financial lending and to improve the credit defaulter ratio, advanced statistical methods were introduced to fill the gap of underperforming credit scoring models. These advanced statistical models provided the alternative from the previous traditional statistical models which were based on the logistic regression and discriminate analysis In the more recent years, different researchers have also employed different data mining techniques for the loan defaulter predictions. We have proposed a study regarding the comparison of Decision tree classifier and Random Forest algorithm towards recommendation engine against the financial lending request. The purpose of this study is to provide a comprehensive research and to develop a model to predict the loan defaults. This kind of models becomes inevitable as the issue of bad loans is very much critical in the financial sector especially in micro financing banks of various underdeveloped and developed countries. To cope up with this problem a comprehensive literature review was done to study the significant factors that lead to this issue. Moreover, these reviewed studies were critically focused towards applying data mining techniques for the prediction and classification of the loan defaults.