Pune Institute of Computer Technology

Department of Computer Engineering

(2022- 2023)

“Titanic Survival Prediction”

Submitted to the

Savitribai Phule Pune University

In partial fulfilment for the award of the Degree of

Bachelor of Engineering

in

Computer Engineering

By

1)	**Abhijeet Jagtap**	**41427**
2)	**Jay Sonawane**	**41430**

Under the guidance of

Prof. Vaishali Kandekar

Problem Statement

Build a machine learning model that predicts the type of people who survived the titanic shipwreck using passenger data (i.e., name, age, gender, socio-economic class, etc)

Objective

To implement a ML model that can predict the passengers who survived the titanic shipwreck.

Theory

Machine Learning model:

A machine learning model is defined as a mathematical representation of the output of the training process. A ML model is like computer software designed to recognize patterns or behaviours based on previous experiences or data. The learning algorithm discovers patterns within the training data, and it outputs an ML model which captures these patterns and makes predictions on the new data.

Classification of ML models:

1) Supervised Learning
2) Unsupervised Learning
3) Reinforcement Learning

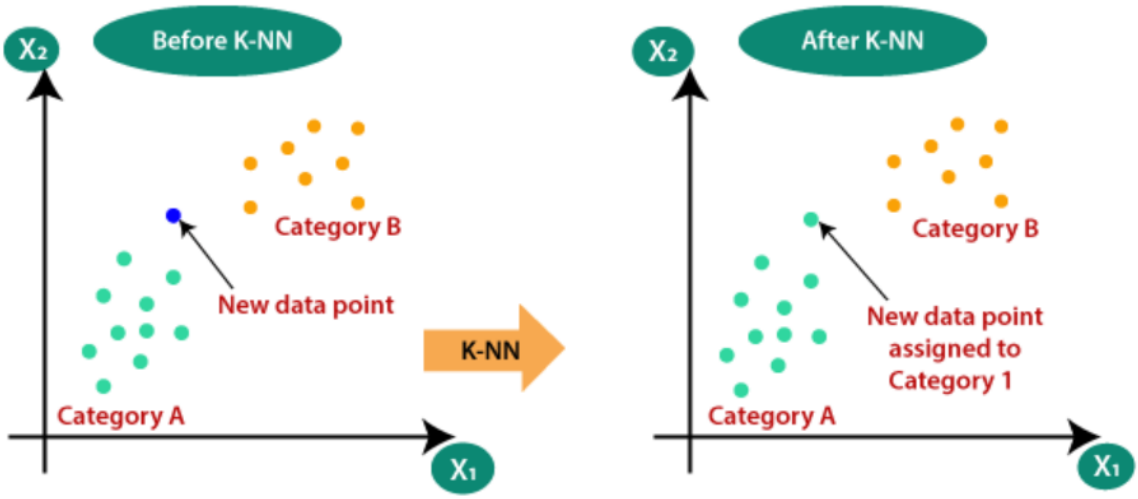
K-Nearest Neighbors (KNN):

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression problems. However, it is mainly used for classification problems in the industry.

Following are the some important points regarding KNN-algorithm.

- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much like the new data.



To implement KNN algorithm you need to follow following steps.

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of K number of neighbors
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

Logistic Regression:

Logistic Regression is basically supervised classification algorithm. In a classification problem, the target variable(or output), y , can take only discrete values for a given set of features(or inputs), X. The model builds a regression model to predict a probability that a given data entry belongs to the category numbered as “1”. Logistic regression models the data using the sigmoid function.

$$\theta(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$

Random Forest:

Random forest is a **Supervised Machine Learning Algorithm** that is **used widely in Classification and Regression problems**. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing **continuous variables** as in the case of regression and **categorical variables** as in the case of classification. It performs better results for classification problems.

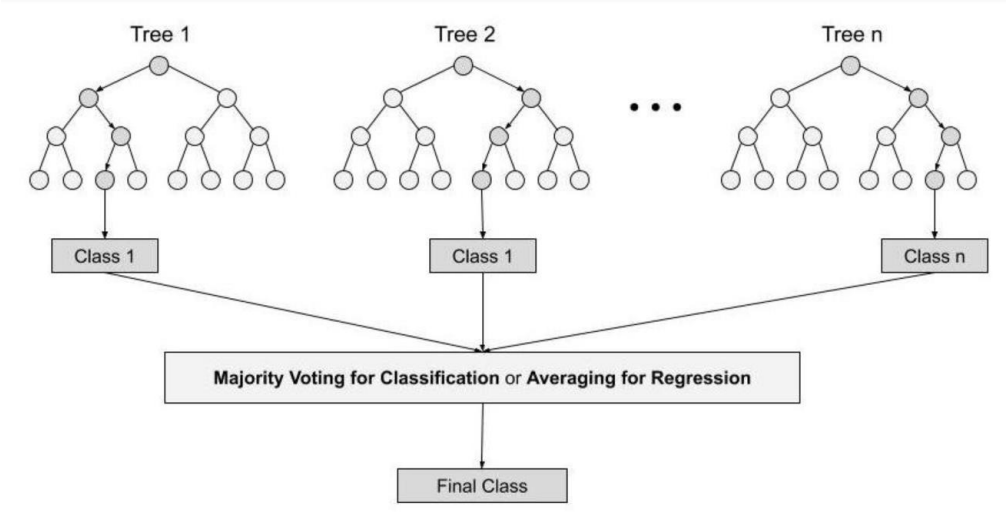
Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on ***Majority Voting or Averaging*** for Classification and regression respectively.



Naïve Bayes:

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

Bayes Theorem: $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is presence of one particular feature does not affect the other. Hence it is called naive.

Stochastic Gradient Descent:

Stochastic gradient descent is a very popular and common algorithm used in various Machine Learning algorithms, most importantly forms the basis of Neural Networks.

Gradient descent is an iterative algorithm, that starts from a random point on a function and travels down its slope in steps until it reaches the lowest point of that function.

The steps of the algorithm are

1. Find the slope of the objective function **with respect to each parameter/feature**. In other words, compute the gradient of the function.

2. Pick a random initial value for the parameters. (To clarify, in the parabola example, differentiate "y" with respect to "x". If we had more features like x1, x2 etc., we take the partial derivative of "y" with respect to each of the features.)

3. Update the gradient function by plugging in the parameter values.

4. Calculate the step sizes for each feature as : **step size = gradient * learning rate.**

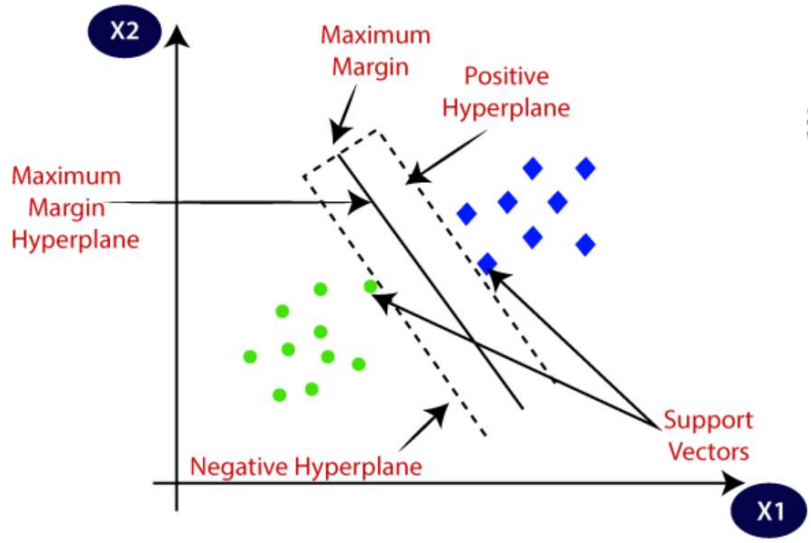
5. Calculate the new parameters as : **new params = old params -step size**

6. Repeat steps 3 to 5 until gradient is almost 0.

Linear Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.



Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Decision Tree:

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.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

Algorithm:

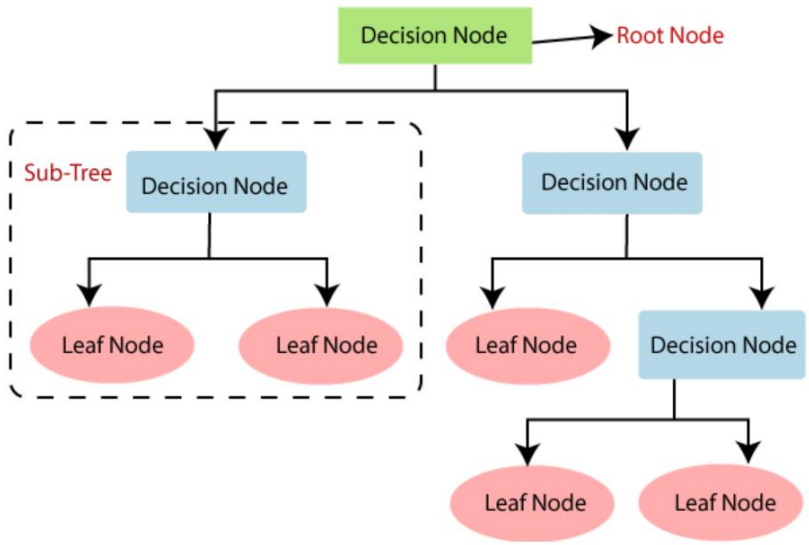
Step-1: Begin the tree with the root node, says S, which contains the complete dataset.

Step-2: Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**

Step-3: Divide the S into subsets that contains possible values for the best attributes.

Step-4: Generate the decision tree node, which contains the best attribute.

Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.



Conclusion

We have implemented various ML model and selected the best fit for the titanic dataset that produces high accuracy predictions of survival.