**DS Lab**

**Exp - 10**

**Shashwat Tripathi**

**D20A Roll no: 64**

**Aim**: Analysis and comparison of different Machine learning and Deep learning algorithms

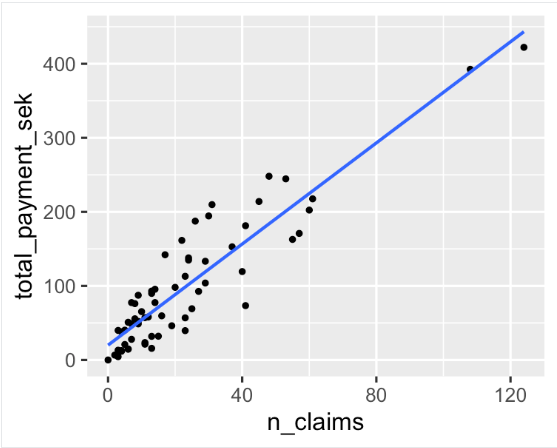
.

**Theory:**

### **Supervised Machine Learning Algorithms**

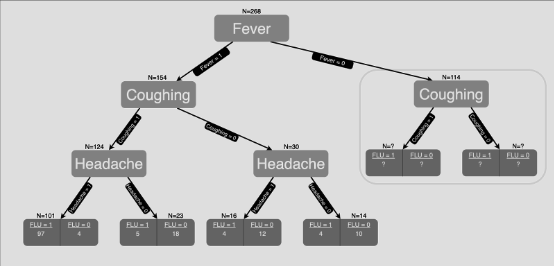
#### **1. Linear Regression**

A simple algorithm models a linear relationship between one or more explanatory variables and a continuous numerical output variable. It is faster to train as compared to other machine learning algorithms. Its biggest advantage lies in its ability to explain and interpret the model predictions. It is a regression algorithm used to predict outcomes like customer lifecycle value, housing prices, and stock prices.



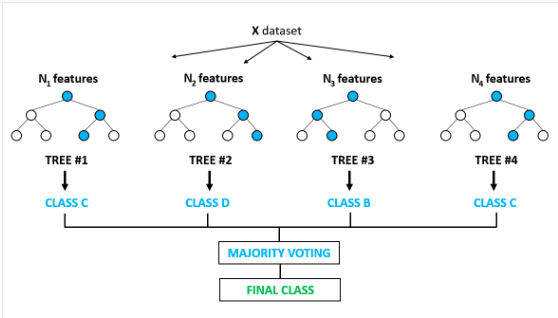
#### **2. Decision Trees**

A decision tree algorithm is a tree-like structure of decision rules that are applied to the input features to predict the possible outcomes. It can be used for classification or regression. Decision tree predictions provide a good aid for healthcare experts as it is straightforward to interpret how those predictions are made.



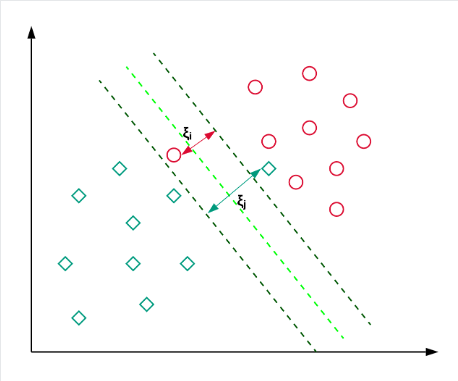
#### **3. Random Forest**

It is arguably one of the most popular algorithms and builds upon the drawbacks of overfitting prominently seen in the decision tree models. Overfitting is when algorithms are trained on the training data a bit too well, and where they fail to generalize or provide accurate predictions on unseen data. Random forest solves the problem of overfitting by building multiple decision trees on randomly selected samples from the data. The final outcome in the form of the best prediction is derived from the majority voting of all the trees in the forest.



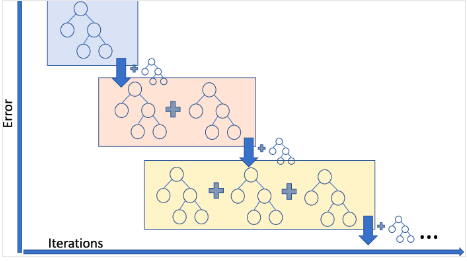
#### **4. Support Vector Machines**

Support Vector Machines, commonly known as SVM, are generally used for classification problems. As shown in the example below, an SVM finds a hyperplane (line in this case), which segregates the two classes (red and green) and maximizes the margin (distance between the dotted lines) between them.



#### **5. Gradient Boosting Regressor**

Gradient Boosting Regression is an ensemble model that combines several weak learners to make a robust predictive model. It is good at handling non-linearities in the data and multicollinearity issues.

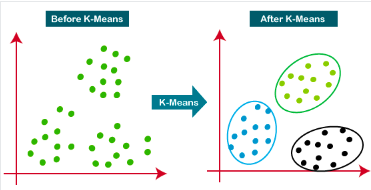


If you are in a ride sharing business and need to predict the ride fare amount, then you can use a gradient boosting regressor.

### **Unsupervised Machine Learning Algorithms**

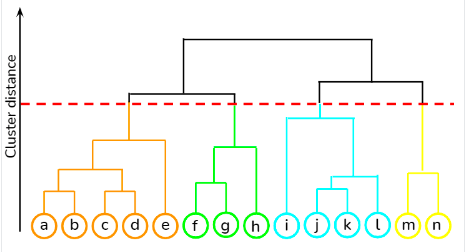
#### **6. K-means Clustering**

K-Means is the most widely used clustering approach—it determines K clusters based on Euclidean distance. It is a very popular algorithm for customer segmentation and recommendation systems.



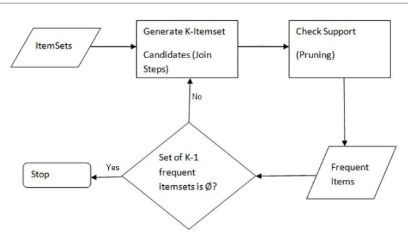
#### **7. Hierarchical Clustering**

It is a bottom-up approach where each data point is treated as its own cluster, and then the closest two clusters are merged together iteratively. Its biggest advantage over K-means clustering is that it does not require the user to specify the expected number of clusters at the onset. It finds application in document clustering based on similarity.



#### **8. Apriori Algorithm**

A rule-based approach that identifies the most frequent itemset in a given dataset where prior knowledge of frequent itemset properties is used. Market basket analysis employs this algorithm to help behemoths like Amazon and Netflix in translating the heaps of information about their users into simple rules of product recommendations. It analyses the associations between millions of products and uncovers insightful rules.

.

### **Ensemble learning**

### Ensemble learning is a machine learning technique that combines the predictions from multiple individual models to obtain a better predictive performance than any single model. The basic idea behind ensemble learning is to leverage the wisdom of the crowd by aggregating the predictions of multiple models, each of which may have its own strengths and weaknesses. This can lead to improved performance and generalization.

### Ensemble learning can be thought of as compensation for poor learning algorithms that are computationally more expensive than a single model. But they are more efficient than a single non-ensemble model that has passed through a lot of learning. In this article, we will have a comprehensive overview of the importance of ensemble learning and how it works, different types of ensemble classifiers, advanced ensemble learning techniques, and some algorithms (such as random forest, xgboost) for better clarification of the common ensemble classifiers and finally their uses in the technical world.

### Ensemble Learning Techniques

### **Gradient Boosting Machines (GBM):** Gradient Boosting is a popular ensemble learning technique that sequentially builds a group of decision trees and corrects the residual errors made by previous trees, enhancing its predictive accuracy. It trains each new weak learner to fit the residuals of the previous ensemble’s predictions thus making it less sensitive to individual data points or outliers in the data.

### **Extreme Gradient Boosting (XGBoost):** XGBoost features tree pruning, regularization, and parallel processing, which makes it a preferred choice for data scientists seeking robust and accurate predictive models.

### **Random Forest Variants:** They introduce variations in tree construction, feature selection, or model optimization to enhance performance.

### 

### **Overall Insights**

| **Algorithm** | **Definition/Description** | **Use Cases** |
| --- | --- | --- |
| **Linear Regression** | Predicts a continuous target variable based on linear relationships with input features. | Regression tasks, trend analysis. |
| **Logistic Regression** | Predicts binary outcomes based on linear relationships. | Binary classification, medical diagnosis. |
| **Decision Trees** | Tree-like model that splits data into branches based on feature values. | Classification, regression. |
| **Random Forest** | Ensemble method using multiple decision trees for improved accuracy and reduced overfitting. | Classification, regression, feature importance. |
| **Bagging** | Combines multiple models trained on different subsets of data to improve accuracy. | Reducing variance in predictions. |
| **Boosting** | Sequentially combines weak learners to create a strong learner, focusing on errors. | Improving model accuracy, especially in imbalanced datasets. |
| **Support Vector Machines (SVM)** | Finds the optimal hyperplane to separate classes in high-dimensional space. | Classification, outlier detection. |
| **K-Nearest Neighbors (KNN)** | Classifies instances based on the majority class of their nearest neighbors. | Classification, recommendation systems. |

**Conclusion**:The experiment demonstrated the effectiveness of AdaBoosting in focusing on misclassified instances to improve model accuracy, while Random Forests showed its strength in reducing overfitting by averaging multiple decision trees. Both algorithms successfully highlighted the power of ensemble methods in supervised learning for classification tasks.