**CHAPTER 5**

**EXPERIMENTAL SETUP**

## 5.1 SYSTEM ARCHITECTURE

Figure 5.1 System Architecture

## 5.2 MOTIVATION FOR USING CLASSIFICATION ML ALGORITHMS

Classification in machine learning is a supervised ML technique. It predicts group relationships for data instances in the dataset. It is a recursive process of recognizing, understanding and grouping the data objects using labels into pre-defined categories. Classification algorithms are used in Machine Learning to predict the class label of a given data point. It allows machines to learn and predict new data points, even when no class labels are known.

* Linear regression is utilized for regression tasks, while logistic regression helps accomplish classification tasks.
* Linear regression is used to predict the continuous dependent variable using a given set of independent variables whereas Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
* Since we are trying to predict a categorical value, we chose logistic regression instead of linear regression.

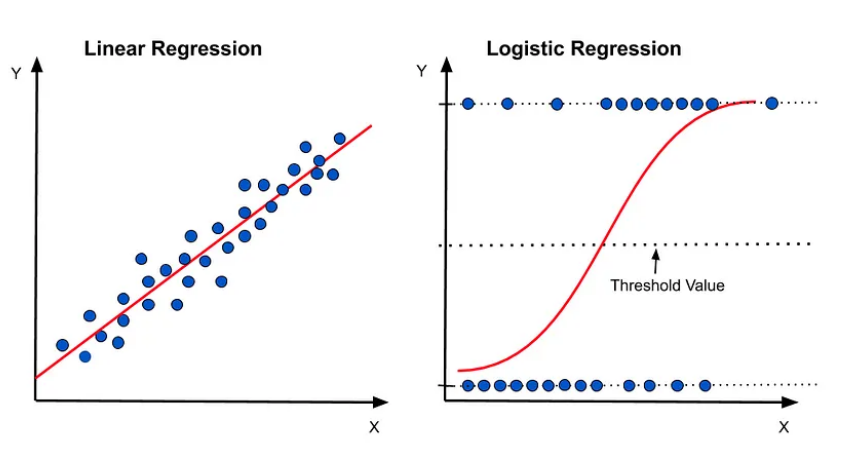


Figure 5.2 Comparing Logistic regression with Linear regression

## 5.3 MODEL BUILDING

## 5.3.1 STANDARDIZATION OF DATA

[Standardization of data is a technique used in machine learning to transform features so that they have a mean of zero and a standard deviation of one](https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/). [This is done to ensure algorithmic stability, prevent sensitivity to the scale of input features, improve optimization algorithms’ convergence and search efficiency, and enhance the performance of certain machine learning algorithms](https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/). [Standardization is used when a model has a regularization term or is otherwise sensitive to the scaling of input features](https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/). [Standardization is different from normalization, which scales features to a fixed range](https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/).

## 5.3.2 SPLITTING OF DATA

In order to test the model on unseen data, it is necessary to split the data and this should be done randomly to avoid bias in the training or the testing phase. Random split ensures that the model is trained on edges which belong to both the classes (1 and 0). We have used 70% of the data for training and the remaining 30% for testing the model.

## 5.3.3 BUILDING THE ML MODELS

## 5.3.3.1 LOGISTIC REGRESSION

Logistic regression gives us the probability of a certain input features belonging to a certain class. It uses the logistic function to model the binary dependent variable.

## 5.3.3.2 RANDOM FOREST

Random Forest is a collection of decision trees. It operates on the principle that; a large number of uncorrelated trees will outperform a single model. Each tree predicts a class and the class with most predictions is chosen as the model’s prediction

## 5.3.3.3 SUPPORT VECTOR MACHINE

In SVM, the aim is to find a hyperplane that best divides the data points into 2 classes. Predictions are made based on the position of a certain data point with respect to the hyperplane.

## 5.3.3.4 ADABOOST

AdaBoost stands for Adaptive Boosting. It is a statistical classification meta-algorithm that can be used in conjunction with many other types of learning algorithms to improve performance. [AdaBoost works by combining multiple weak learners (simple and weak classifiers) into a single strong learner](https://dsworld.org/what-is-an-adaboost-classification/). [The weak learners are trained on different subsets of the training data](https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/). [The final model is a weighted sum of the weak learners](https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/).

## 5.3.3.5 XGBOOST

XGBoost makes the best use of available resources like memory and time. It implements gradient boosting decision trees which trains multiple models in a gradual, additive and sequential manner