

CSE 523 Machine Learning

Progress Report - 3

Section 1

Date of Submission: 23rd February 2022

Group Name: **Bug Smashers**

Group Details

| Roll No | Name |
| --- | --- |
| AU1940028 | Moksh Doshi |
| AU1940120 | Jaimik Patel |
| AU1940130 | Nandish Patel |
| AU1940164 | Jenil Bagadiya |

2022 Winter Semester

# Tasks Performed this week

* Updated [Google Colab](https://colab.research.google.com/drive/1sFlAk80RdzKh5-wykBfXMNb8BCkfNQup#scrollTo=B_x2jrWGphYn)
* Univariate and Multivariate Statistical Analysis to gain even deeper insights
* Finalize the features to be included in the model training
* Learn about different classification models which can be applied to this particular data set

# Outcomes of the tasks performed

# Here is our understanding of different algorithms:

1. Naive Bayes
   1. Naive Bayes is a classification technique that is based on the Bayes theorem.
   2. It is based on an assumption that predictors are independent of one another.
   3. In short, any particularly given feature is independent of the other features of the given classification.
   4. Bayes theorem simply provides a way to calculate the posterior probability [P(C|X)] where P(C) is Prior Probability of class, P(X|C) is Likelihood (Probability of Predictor given Class), P(X) is Prior Probability of Predictor
2. Logistic Regression
   1. Logistic regression is a classification algorithm that is used to estimate the output of discrete values which are based on independent variables of a given dataset. In our case, it will be used to predict whether the person has heart disease or not.
   2. It will simply predict the number of occurrences of an event that are going to happen. So, in simple terms, it will give the output probability and its value will lie between 0 to 1.
3. Support Vector Machine (SVM)
   1. Here the number of features is taken as dimensions and then the point is plotted
   2. Coordinates of the point will be support vectors
   3. We will draw lines to classify data into different groups which in our case will be whether disease occurs or not
4. k-nearest neighbors (kNN)
   1. It simply classifies a point with the same category as the majority of its k nearest neighbors.
   2. It has the challenge of deciding k and the optimum distance method to apply.
   3. It is computationally expensive as it has to store a lot of data but since it is a concern of life, if it helps to correctly predict, the issue can be dealt with.
5. Random Forest
   1. Random forest can be used for both regression and classification problems and it is implemented by using ensemble learning which is the process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. More trees in the forest prevent the problem of overfitting and lead to higher accuracy.
   2. For predicting the perfect correct output there are 2 assumptions for the given dataset.
      1. there should be some actual values in the features variable of the dataset
      2. the predictions from each tree must have very low correlations.
6. Decision Tree
   1. Decision tree is a supervised form of learning where certain parameters are divided into various data.
   2. It is represented mainly in 2 categories: decision nodes and leaves. The algorithm mostly works for both categorical and continuous dependent variables.
   3. The main agenda is to split the data into 2 or more homogenous sets which are done based on the most significant attributes to make as many distinct groups as possible.
   4. The decision tree generally mimics human thinking ability while making the decisions and compares the values of record attributes to root attributes along with sub-nodes and moves further.

# Tasks to be performed in the upcoming week

* 1. Start with the implementation of the algorithm
  2. Check accuracy of the model and tune the hyperparameters
  3. Learn about more libraries such as XGBoost