**CS6850 –Intro to Machine Learning**

**Project Report**

**(Bhavya Induri)**

In this project for the purpose of dataset classification , used different set of classification algorithms for different type of datsets.

**Language :** Python ,R

**Modules :**Pandas ,numpy,sklearn,MissForest,Mice ,Soft Impute

**Classification 1 :**Initially we normalize the values using feature scaling.Then PCA is performed to reduce dimensionality of the datasets. Then we use SVM linear classifier to perform classification.

SVM classifier : Support vector machine is another simple algorithm that every machine learning expert should have .Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks. The classifier separates data points using a hyperplane with the largest amount of margin

**Classification 2:** As the dataset given simple ,instead of using complex algorithms we simply use KNN to classify the dataset.

KNN classifier :  Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

**Classification 3:** Initially we fill the missing values by median of the feaures . Then we use SVM linear classifier to perform classification. Followed by gridsearch to perform the classification.

SVM classifier : Support vector machine is another simple algorithm that every machine learning expert should have .Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks. The classifier separates data points using a hyperplane with the largest amount of margin.

**Classification 4 & 5 :** Used Random Forest classifier for both the classification problems. For classification 4 we perform PCA before performing the classification.

Random Forest Classifier **Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks.** Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

**Classification 6 :** we use linear regression to classify the dataset.

Linear regression : Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

**Missindata Values 1:** we are using MICE algorithm.

Missing data are a common problem in psychiatric research. Multivariate imputation by chained equations (MICE), sometimes called “fully conditional specification” or “sequential regression multiple imputation” has emerged in the statistical literature as one principled method of addressing missing data. Creating multiple imputations, as opposed to single imputations, accounts for the statistical uncertainty in the imputations. In addition, the chained equations approach is very flexible and can handle variables of varying types (e.g., continuous or binary) as well as complexities such as bounds or survey skip patterns.

MICE, are very flexible and can be used in a broad range of settings. Because multiple imputation involves creating multiple predictions for each missing value, the analyses of multiply imputed data take into account the uncertainty in the imputations and yield accurate standard errors. On a simple level, if there is not much information in the observed data (used in the imputation model) regarding the missing values, the imputations will be very variable, leading to high standard errors in the analyses. In contrast, if the observed data are highly predictive of the missing values the imputations will be more consistent across imputations, resulting in smaller, but still accurate, standard errors\

**Missing Data 2:**

As the dataset given we use KNN to find missing values in the dataset.

KNN classifier :  Nearest Neighbor(KNN) is a very simple, easy to understand, versatile and one of the topmost machine learning algorithms. KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach. The assumption behind using KNN for missing values is that a point value can be approximated by the values of the points that are closest to it, based on other variables.A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

**Multi Label classification :** we are using one vs rest algorithm to perform multilabel classification.

One vs Rest : This strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label; discrete class labels alone can lead to ambiguities, where multiple classes are predicted for a single sample