# Report 2: Task (CSCI 527)

**Siva Sai Reddy Kaliki (202004650),**
**Kranthika Gurram (202005785),**
**Hemanth Chekka (202006238),**
**Sai Teja Reddy Palam(202106549)**
St. Francis Xavier University

## 1 Task Design

We have selected the Wine Quality Dataset [1] from the Kaggle website. We aim to predict the quality of a given wine based on the attribute values. In this task, we are going to assign two classes that are either "good" or "bad" for the target variable of the given wine. This means that the output will be of classification type.

Since we know the expected result from our task, we have prepossessed our dataset to remove any inconsistencies. In the dataset, we have found null values using the isnull() method in some of the records of the following columns such as Fixed Acidity, Volatile Acidity, Residual Sugar, Chlorides, Citric Acid, pH, sulphates which can be seen in Figure 1.

```
type                   0
fixed acidity         10
volatile acidity       8
citric acid            3
residual sugar         2
chlorides              2
free sulfur dioxide    0
total sulfur dioxide   0
density                0
pH                     9
sulphates              4
alcohol                0
quality                0
dtype: int64
```

Figure 1: Before Prepossessing

After identifying the columns with the null values, the mean of each column is calculated. After that, the null values in each column are replaced by their respective column's mean. After prepossessing, the dataset has no more null values in any of the columns which can be seen in Figure 2.

```
type                   0
fixed acidity          0
volatile acidity       0
citric acid            0
residual sugar         0
chlorides              0
free sulfur dioxide    0
total sulfur dioxide   0
density                0
pH                     0
sulphates              0
alcohol                0
quality                0
dtype: int64
```

Figure 2: After Prepossessing

On the prepossessed dataset, we have transformed the data before splitting them into training and testing sets by following the steps below:

- The column 'type' is dropped from the dataset.
- The array items are divided into several bins using the Pandas cut() method which is used for scalar data statistical analysis.
- For the Pandas cut() method, the following variables are passed: the dataset with only the quality attribute, bins with values (2, 6.5, 8), and labels as ['bad', 'good'].
- Finally, the preprocessed dataset without the 'type' attribute, is divided into two datasets as 'x' and 'y' where 'x' will consist of the dataset without the 'quality' attribute, on the other hand 'y' will only have 'quality' attribute values in the dataset.

For training and testing the models, the dataset is split across the 'x' and 'y' datasets in the proportion of 80 percent and 20 percent with some randomization which gives us the x_train, x_test, y_train, y_test sub-datasets by using the train_test_split() method by passing 'x' and 'y' datasets.

## 2 Models Used

Since the above task can be considered as both predictive and classification type, we have considered classifier models. Classifier Models come under the category of Supervised Machine Learning Models because they involve training the model with labeled input and output values.

For our task, we have used Logistic Regression and Random Forest Classifier models. Training data is fitted into these models for training the model. After fitting, we obtain the predicted output values possible from each classifier model when the input label of testing data is given. Predicted output is used along with test output values to check the model performance on the given dataset. The working of each model on the dataset is explained below:

### 2.1 Random Forest Classifier

A random forest fits a number of decision tree classifiers on various sub-samples of the dataset. It averages the outputs and improves the predictive accuracy. It also controls over-fitting. We have set n_estimators parameter value to 50 which means we will have 50 trees in our classifier model. We trained a random forest classifier by passing the x_train and y_train as input parameters to the fit() method. Then, we used predict() method to obtain the predicted values of the random classifier (random_pred) by passing test data input (x_test). Finally, test-output (y_test) and random_pred (predicted value) are sent as input to accuracy_report() method to obtain performance measures of the random forest classifier.

```
The Accuracy of Modle-1: Random Forest Classifier is:  0.8853846153846154
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.97      0.93      1041
           1       0.80      0.56      0.66       258
           2       0.00      0.00      0.00         1

    accuracy                           0.89      1300
   macro avg       0.57      0.51      0.53      1300
weighted avg       0.88      0.89      0.88      1300
```

Figure 3: Random Forest Performance

### 2.2 Logistic Regression

Logistic regression is made on the logistic model in statistics. It models the probability of an event taking place. LogisticRegression() constructor creates the default logistic classifier. Similar to the random forest classifier, we used fit() method with x_train and y_train as input values to train the logistic classifier model. We use predict() method with x_test (test-input) on the model and store predicted logistic regression model values in the logistic_pred variable. Performance values are observed using accuracy_report with y_test and logistic_pred values.

```
The Accuracy of Modle-2: Logistic Regression is:  0.8215384615384616
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.96      0.90      1041
           1       0.62      0.25      0.36       258
           2       0.00      0.00      0.00         1

    accuracy                           0.82      1300
   macro avg       0.49      0.41      0.42      1300
weighted avg       0.80      0.82      0.79      1300
```

Figure 4: Logistic Regression Performance

## 3 Results Obtained

Based on the calculations performed using statistical measures like accuracy, f1-score, and precision on these classifier models, the Random Forest classifier is found to give better accuracy than the Logistic Regression Model on our dataset.

| Model | Accuracy |
|---|---|
| Random Forest | 88.5384 |
| Logistic Regression | 82.1538 |

## References

[1] Wine quality — kaggle dataset. https://www.kaggle.com/datasets/rajyellow46/wine-quality.