Pattern Recognition Assignment-1 Report

Data Preparation:

Important Note: In the dataset description document it is mentioned that the feature **'duration'** is **highly correlated** with the output class, hence this feature has been **removed** to make a **realistic predictive model**.

1. Encoding the data

- By analyzing the dataset we found out that out of 19 Features
    - 9 Features were Numerical Features
    - 4 Features were Ordinal Features
    - 6 Features were Nominal Features
- Now we had to convert Categorical Features(Nominal + Ordinal) into numerical features
    - **Ordinal Feature** entries are converted into numerical entries based on the mutual ordering among themselves.
    - Nominal Feature Entries are encoded in 2 ways
        i. Label Encoding
        ii. One Hot Encoding

NOTE: We now have 2 types of encoded data, one is the **label encoded data** whose dimensions are same as original data, another is **one-hot encoded data** whose dimensions are more than actual data. (i.e. One-hot encoded data has 55 Features, whereas original data has 19 features)

2. Making the Data Symmetric

- The final output (i.e. class) can either be 'yes' or 'no'. In the given dataset 87% of tuples belong to the class 'no'. Therefore the data is **highly Skewed**.
- To make the data symmetric, we took the tuples belonging to minority class(i.e. Class 'yes') and duplicated them until both the classes have an equal number of tuples in them.

3. Splitting 'pdays' feature into 2 features

- The entries of 'pdays' feature are as follows:
    - More than 90% of the entries are 999. (i.e. Client was never contacted)
    - Remaining entries are in the range 0 - 20.
- Because of this, when we standardize the data, the entries corresponding to 999 will become 1 and remaining entries will be very close to zero(they all will be almost the same).
- To avoid this we split the feature into 2 features where one feature contains information about whether the client was contacted before or not(binary feature), and another feature will contain information about how long ago the client was contacted(if the client was never contacted we put 30 in this field instead of 999).

4. Standardizing the data

- This step was done :
    - For faster convergence.
    - To ensure that variables measured at different scales do not contribute differently for analysis.

Classification

In this assignment we tried out the following classifiers:

- Logistic Regression(with Linear and Polynomial Features)
- SVM Classifier
- Random Forest Classifier

Here is a table comparing the accuracy, precision, recall, and f1_score of all the models.

Classifier	Accuracy Score	Precision Score	Recall Score	F1 Score
Logistic Regression (Linear Features)	0.751260	0.639889	0.819149	0.718507
Logistic Regression (Poly Features of degree = 2)	0.824553	0.844875	0.809735	0.826932
Logistic Regression (Poly Features of degree = 3)	0.922584	0.984303	0.875205	0.926554
SVM Classifier	0.799817	0.848060	0.726685	0.782695
Random Forest Classifier	0.958554	0.999459	0.924072	0.960288

Note: Logistic Regression and **SVM classifier** gave better results with **One Hot Encoded data**, whereas **Random Forest Classifier** gave similar results for both **One Hot Encoded data** and **Label Encoded data.**

PCA

We now have 2 types of encoded data:

1. One Hot Encoded Data (55 Features)
2. Label Encoded Data (20 Features)

By Applying PCA on both the datasets we observe that to capture 80% of the variance:

1. We need only **17 features** out of 55 features in **One Hot Encoded Data**.
2. We need only **3 features** out of 20 features in **Label Encoded Data**.

Note:

- It has been observed that we can achieve similar accuracy scores as mentioned above by just using features enough to capture 90% of the variance.
- Using PCA we found that we just need 24 features out of 55 features(in One Hot Encoded Data) and 5 features out of 20(in Label Encoded Data) to capture 90% of the variance.
- Therefore using PCA we can reduce the number of features used drastically(hence reduce the computational resources used and time taken to train) and still obtain the same accuracy.

Done By:

T. Bharat Bhushan Reddy - B160198CS

G. Bharath - B160653CS

K. A. Siva Vardhan Reddy - B160333CS

Valeti Manoj - B160091CS