Coursework Summary

Fraidoon Omarzai
230073955
MSC AI

1. INTRODUCTION

This summary report talks about the analysis we did for our Data Mining project. We had to look into epitope data and try to guess epitopes for Trypanosoma cruzi. In this document, I explain the steps I took using a Jupyter Notebook to complete this project.

2. Data Loading and EDA

First, I load the dataset which was provided and performed the following operations to get insight the data and get familiar with the given dataset:

- Load the dataset
- Display the first five rows
- List the columns name
- Check the info
- Check the mathematical description of the dataset
- Check the shape of the dataset
- Remove id columns
- Check for null values
- Check the balance of the target feature
- Get only categorical columns and go deeper
- Get only numerical features and explore it
- Numerical variables are usually of 2 types
    - Continuous variable
    - Discrete Variables
- Check the distribution of continuous variables of 20 features only to get some idea about the distribution, whether normal distributed or not

Note: Based on the observation we need to do:

1. Get read of null values

2. Convert feature `Info_AA` from categorical to numerical

3. Balance the dataset

4. Most of the features are normal distributed

3. FEATURE ENGINEERING

I performed the below operations on feature engineering sections:

- Handling missing values using imputer library where I load from sklearn
- Converting Categorical features into Numerical using label encoder
- Handling imbalance dataset using imbalance library

4. FEATURE SELECTION

I used sklearn library for feature selection, and I used SelectFromModel API where I selected the Lasso model and out of 1286 features, the feature selection method extracted only the 6 most important features. I used several different techniques and hyper parameter which I decided to select the final techniques due to getting higher accuracy.

Note: I also split the dataset to train and test in order to select the best model.

5. MODEL TRAINING AND EVALUATION

In this section I used several models such as:

- RandomForest
- KNN
- SVM
- Logistic Regression

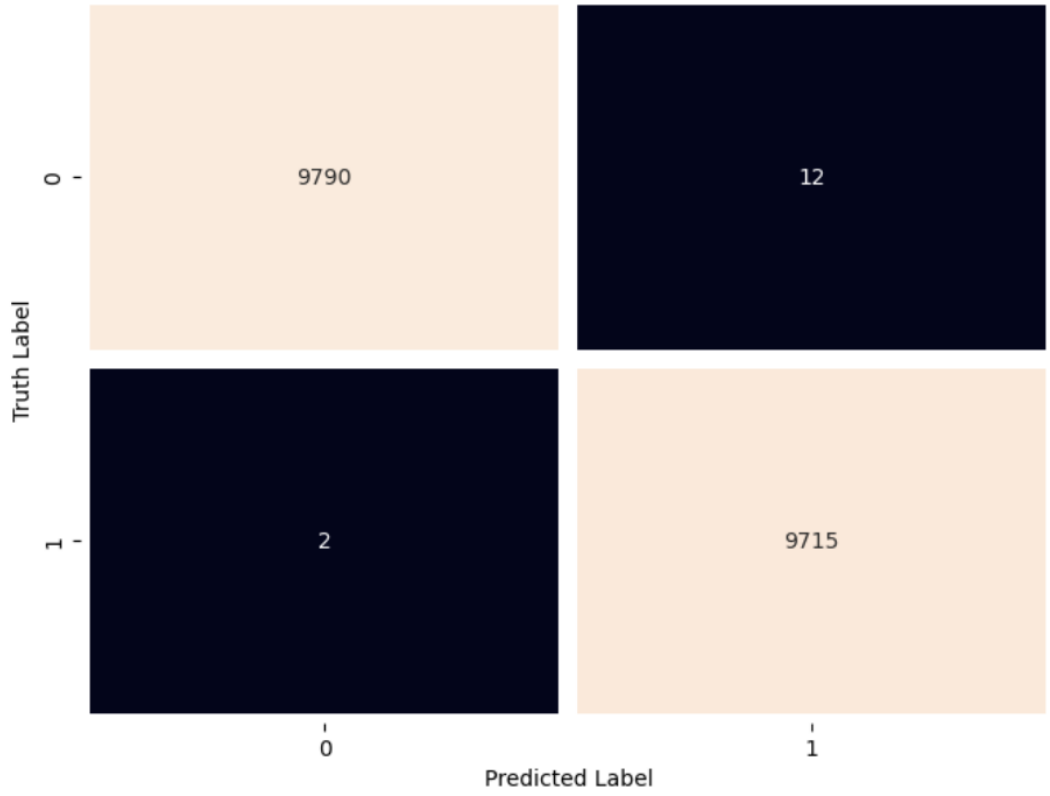
Where I got the highest accuracy of 99.9% using Randomforest, so I selected that model.

Below is the accuracy which I got:

Accuracy-Score: 1.00

	precision	recall	f1-score	support
-1	1.00	1.00	1.00	9802
1	1.00	1.00	1.00	9717
accuracy			1.00	19519
macro avg	1.00	1.00	1.00	19519
weighted avg	1.00	1.00	1.00	19519

Below is the image of Confusion Metrix:



6. FINALLY

Finally, I load the df_holdout.csv file and perform the same operations which I performed on the given dataset for training.

Then, I trained the model on given dataset and store the prediction on submission.csv file.