**CSE523 Machine Learning**

**Prof. Mehul Raval**

**Analyzing TCP Network Packets**

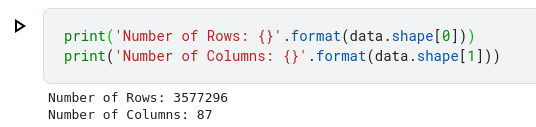
**Week 4 Report**

| **Name** | **Enrolment Number** |
| --- | --- |
| Nityam Dixit | AU2040140 |
| Tej Thakar | AU2040262 |
| Yagnesh Patel | AU2040025 |
| Kavan Gondalia | AU2040030 |

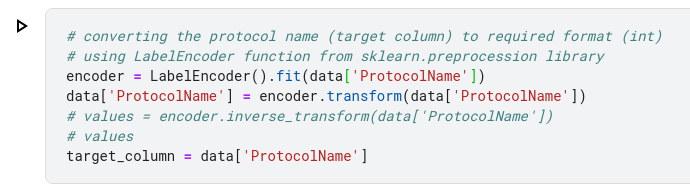
Kaggle Link: <https://www.kaggle.com/code/kavangondalia/predict-protocol-name>

Step by Step Procedures:

1. Started Coding in Kaggle. The link is provided above. Read some docs regarding the pandas library for importing and analyzing the data

* Dropping Columns that are not required for Correlation Matrix<https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.drop.html>  
    
    
  

1. Extracted the Target Column, i.e., ‘ProtocolName’, from the data frame and encoded it into numbers using LabelEncoder() function to build the Correlation Matrix.

* Sklearn Preprocessing: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>  
  

1. Finding the Correlation Matrix
   1. Reason: The Correlation matrix will help us with Dimensionality Reduction.  
      Dimensionality Reduction reduces the number of features for easy computing, leading to easy prediction.
   2. How? We selected columns that have a positive correlation with the target column. The higher correlation indicates they both are dependent, so a good choice in feature selection.  
      
   3. Following Links Referred:

* Reading docs about what is correlation <https://towardsdatascience.com/what-is-correlation-975ea899aaed>
* Understanding the difference between variance/covariance/correlation: <https://medium.com/geekculture/variance-vs-covariance-vs-correlation-what-is-the-difference-95adff96d542>
* Doc about how to choose Feature Selection in ML. This gave me a gist about correlation could be a good way to classify features. <https://medium.com/geekculture/feature-selection-in-machine-learning-correlation-matrix-univariate-testing-rfecv-1186168fac12>
* Getting the gist about features from the correlation matrix: <https://androidkt.com/find-correlation-between-features-and-target-using-the-correlation-matrix/>
* Plotting Heatmaps with heal of Seaborn library: <https://seaborn.pydata.org/generated/seaborn.heatmap.html>
* DIfferent Available Colormaps in Seaborn. We used ‘viridis’ <https://matplotlib.org/stable/tutorials/colors/colormaps.html>
  1. Reduced to **36 columns** from **87 columns** (features)

1. Implemented Random Forest Classification.

* Due to 3500k rows, **there was a time-out**.(: (30 mins idle time)
* So I reduced it to 100k rows and it worked.
* The accuracy **without dimensionality reduction** was **0.97535**
* The accuracy **with dimensionality reduction** was **0.98785**

