**Thesis Final Report**

1. **Data preprocessing**
   * **Kyoto**
2. Drop all Columns with NAN values ( total 121 rows )
3. Change Categorical to numerical:
   * Service
   * Flag
   * Protocol Type
4. Remove unneeded columns (6 cols) :
   * IDS\_detection, Mashula\_detection, Malware\_detection
   * Source and IP destination
   * Start time

D. Min Max normalization

* + **NSL**

1. Convert label to 5 categories ( normal , probe , u2r , r2l , dos )
2. Categorical to numerical:
   * Protocole type
   * Service
   * Flag
   * **ISCX**
3. Split the 4 GB dataset into multiple 70MB files
4. Convert to CSV **ERROR !!**
5. **Grid Search**
   * **Kyoto**
6. KNN
7. MLP
   * 'alpha': 0.01,
   * 'hidden\_layer\_sizes': (80, 80),
   * 'learning\_rate': 'adaptive',
   * 'solver': 'adam'
8. DT
   * **NSL**
9. KNN
   * 'n\_neighbors': 1
10. MLP
    * 'alpha': 0.001
    * 'hidden\_layer\_sizes': (20, 20, 20)
    * 'learning\_rate': 'adaptive'
    * 'solver': 'adam'
11. DT
    * 'max\_depth': 19
    * 'min\_samples\_split': 10

* + **ISCX**

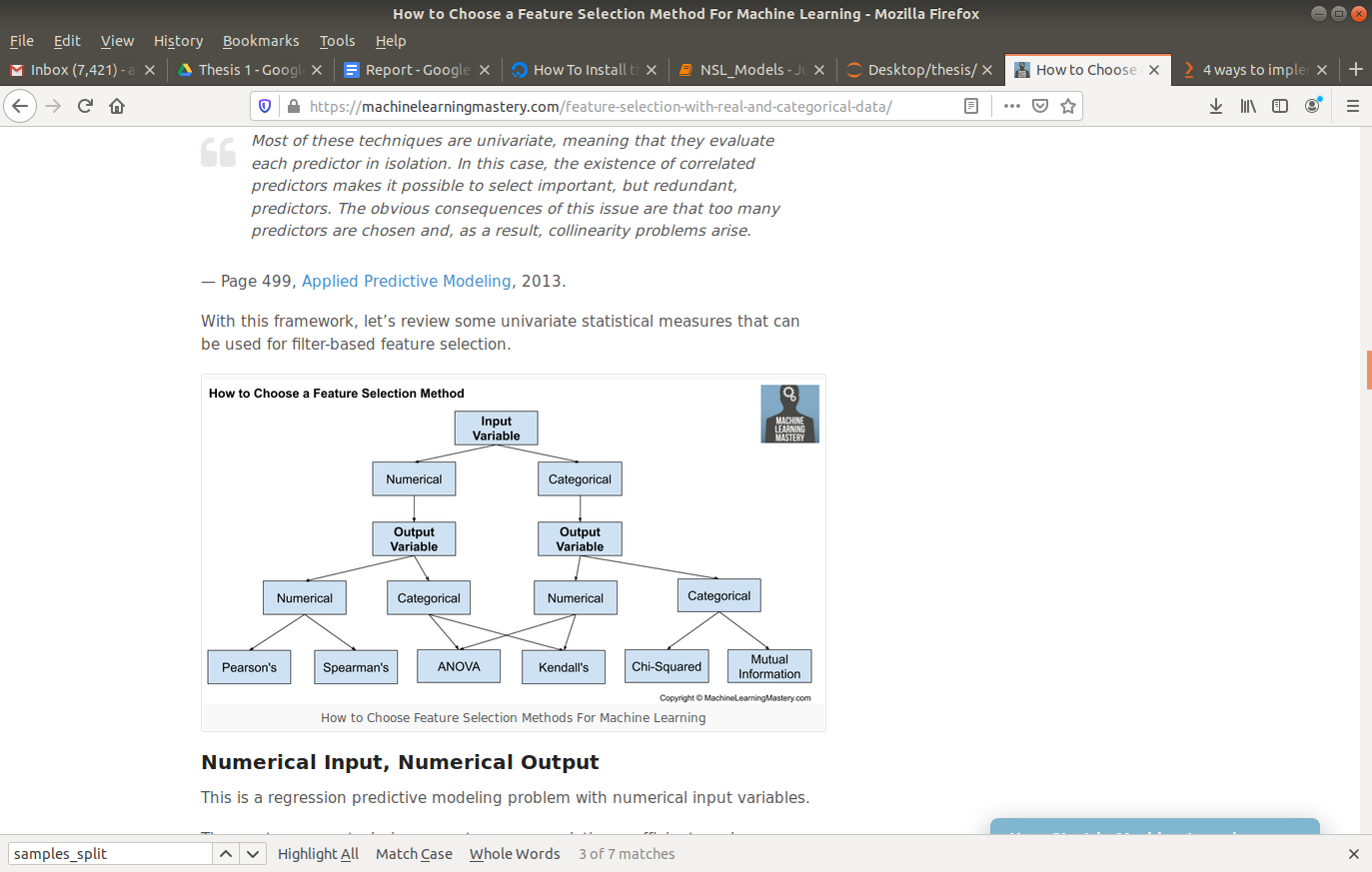
N/A

1. **Feature Selection**

There are mainly 4 methods for feature selection:

1. Univariate selection

Statistical tests can be used to select those features that have the strongest relationships with the output variable.



In our Case we have numerical input and Categorical output which is a classification predictive modeling problem.

The Feature selection in this section will be performed using **ANOVA F** measure via the **F\_classif()** function.

1. Recursive Feature Elimination (RFE)
2. Choosing important features (feature importance)
3. **Results**
   * **Kyoto**
4. Entire Dataset

|  |  |
| --- | --- |
|  | **Accuracy** |
| **SVM** | 89.1 |
| **Random Forest** |  |
| **Decision Tree** | 98.3 |
| **MLP** | 94.02 |
| **KNN** | 97.2 |

1. Recursive feature selection

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Number of Features** | **Features Selected** | **Accuracy** |
| **Random Forest** | 11 | Source\_bytes,  Destination\_bytes,  Srv\_serror\_rate,  Dst\_host\_count,  Dst\_host\_serror\_rate,  Label,  Source\_Port\_Number,  Service\_http,  Flag\_RSTRH,  Flag\_S2,  Flag\_SH |  |
| **SVM** |  |  |  |
| **MLP** |  |  |  |

* + **NSL**

1. Entire Dataset

|  |  |
| --- | --- |
|  | **Accuracy** |
| **SVM** | 93.5 |
| **Random Forest** | 95.7 |
| **Decision Tree** | 99.7 |
| **MLP** | 97.9 |
| **KNN** | 99.6 |

1. Recursive feature selection

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Number of Features** | **Features Selected** | **Accuracy** |
| **Random Forest** | 16 | Src\_bytes , dst\_bytes , logged\_in , count , srv\_serror\_rate , same\_srv\_rate , diff\_srv\_rate , diff\_srv\_rate , dst\_host\_srv\_count , dst\_host\_diff\_srv\_rate , dst\_host\_diff\_srv\_rate , dst\_host\_diff\_srv\_rate , dst\_host\_same\_src\_port\_rate  , dst\_host\_serror\_rate , dst\_host\_srv\_serror\_rate , dst\_host\_rerror\_rate , protocol\_type\_icmp , flag\_S0 , flag\_SF |  |
| **SVM** |  |  |  |
| **MLP** |  |  |  |

1. Feature Importance:

Decision Tree

Random Forest

* + **ISCX**