Division of Computing, Engineering and Mathematical Sciences
Department of Computing, University of Kent *in Canterbury*

Practical Data Analysis Using Python

Written By:

Lukasz Tomaszewski (lrgt2@kent.ac.uk)

Victor Maurin (vm288@kent.ac.uk)

Module CO832: Data Mining and Knowledge Discovery

Computer Science in Artificial Intelligence MSc

LaTeX Word Count: 490

April 2022

The dataset given contains two classes of data, H1N1 (swine flu) and COVID 19 which contains various attributes: sex, symptom, etc. Utilising a decision tree classifier to split the class into a true or false statement to allowing for more accurate treatment of the attributes to be learnt.

The first step is to pre-process the data, first removing all the p" to have strings that python could process by selecting all the object type columns. Then transforming all the columns with strings into two columns for each to have 1 or 0 for the algorithm. The get dummies function from pandas was used and removed the columns where the values were "?". For columns containing numbers, fill NaN to replace all null values with the average of the column.Then, to get the target the algorithm aims, transforming the strings "COVID" and "FLU" into 1 and 0. After processing the data, the data is split, the target and the features. These two groups of data will be used by the algorithm so that it can know what to target by taking some features as input. To learn the decision tree for the algorithm, the scipy package was used. A decision tree was utilised since the data contained numerous columns with true or false values. Having boolean data allowed to have good branching and accuracy in my decision tree.

Analysing the results shown in fig. 1 and fig. 2, in which a 10-fold cross-validation score is implemented on 70% training and 30% test data, outputted the decision tree shown in fig. 3. The tree has the primary internal nodes that split the H1N1 and COVID data into various attributes until it reaches the leaf nodes, the data filters out the number of samples of the attributes in each node that learns onto the next. With an overall accuracy score of 91% of the testing data, with an overall error of 1% error, tabulated in table 2 and the 10-fold cross validation score in table 1, plotting this data in fig. 1 and fig. 2, which highlights the error of the training data onto the testing data.

The decision trees nodes learn the most sampled attributes first and then splits into other attributes such as a symptom first and then learns the male and female attributes of that individual symptom regardless of the class. The tree splits true or false for the H1N1 and COVID 19 classes, however the most learnt attribute is age more so on the H1N1 class, this indicates that COVID 19 is broad in age brackets that H1N1. Regardless of class however, the 'Age' attribute is important as it's fast identifier to those who are more susceptible to get either H1N1 or COVID 19, whereas the attributes: diabetes, serumlevelsofwhitebloodcells and lymphocytes are not as fast to identify, this is why the tree learns these attributes first as they aren't unique attributes compared to the rest of the data.

Appendix

Score of 10-fold cross-validation										
n-fold	1	2	3	4	5	6	7	8	9	10
Score	0.911	0.933	0.911	0.911	0.8	0.911	0.864	0.9318	0.8182	0.8182

Table 1: Tabular results of the score of 10-fold cross-validation

Accuracy Score	0.90807174887892377
Test Score	0.90807174887892377
Train Score	0.988450433108758
Test Error	0.09192825112107628
Train Error	0.011549566891241536

Table 2: Tabular results of the overall scores of the algorithm

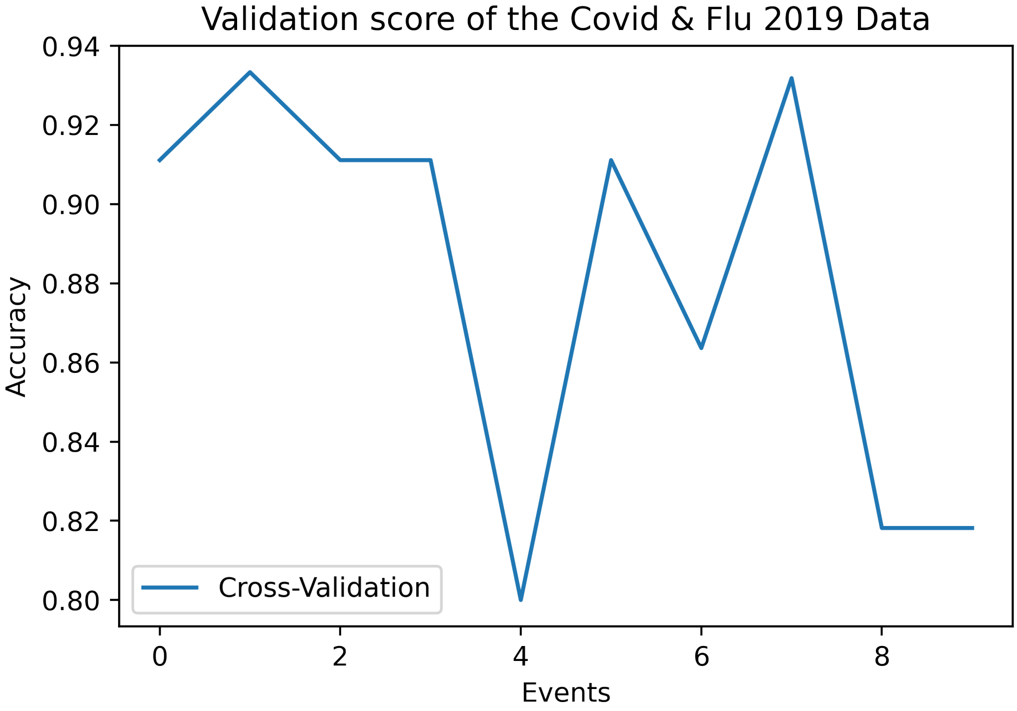

Figure 1: A figure displaying the score validation of the training algorithm using samples from the Covid and Flu 2019 data census.

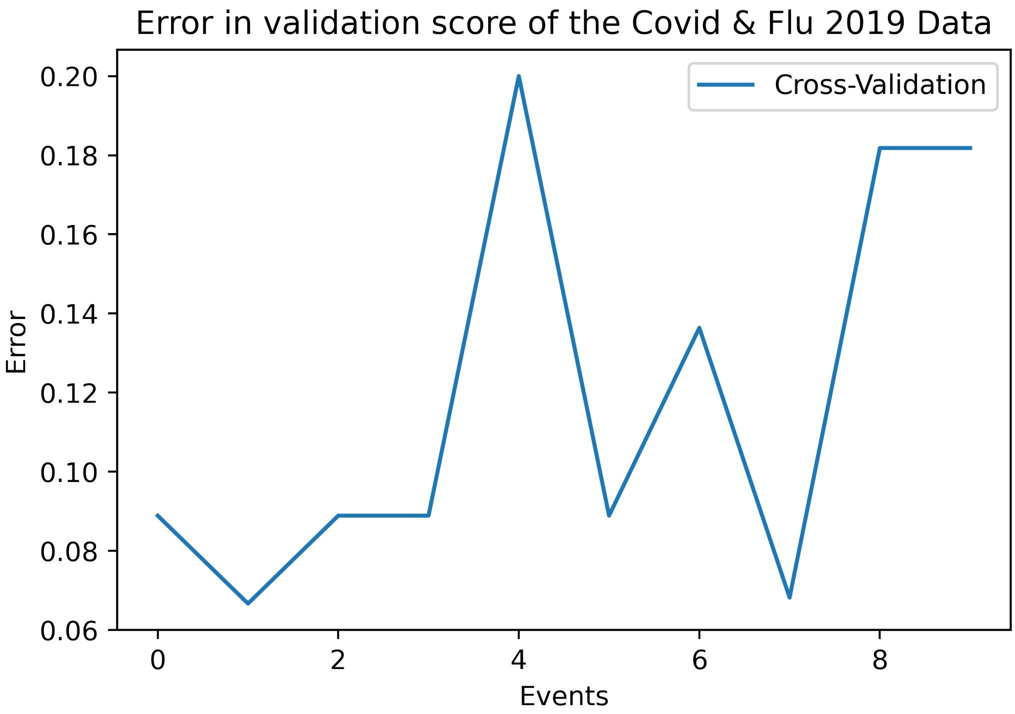

Figure 2: A figure displaying the error in the score validation of the training algorithm using samples from the Covid and Flu 2019 data census.

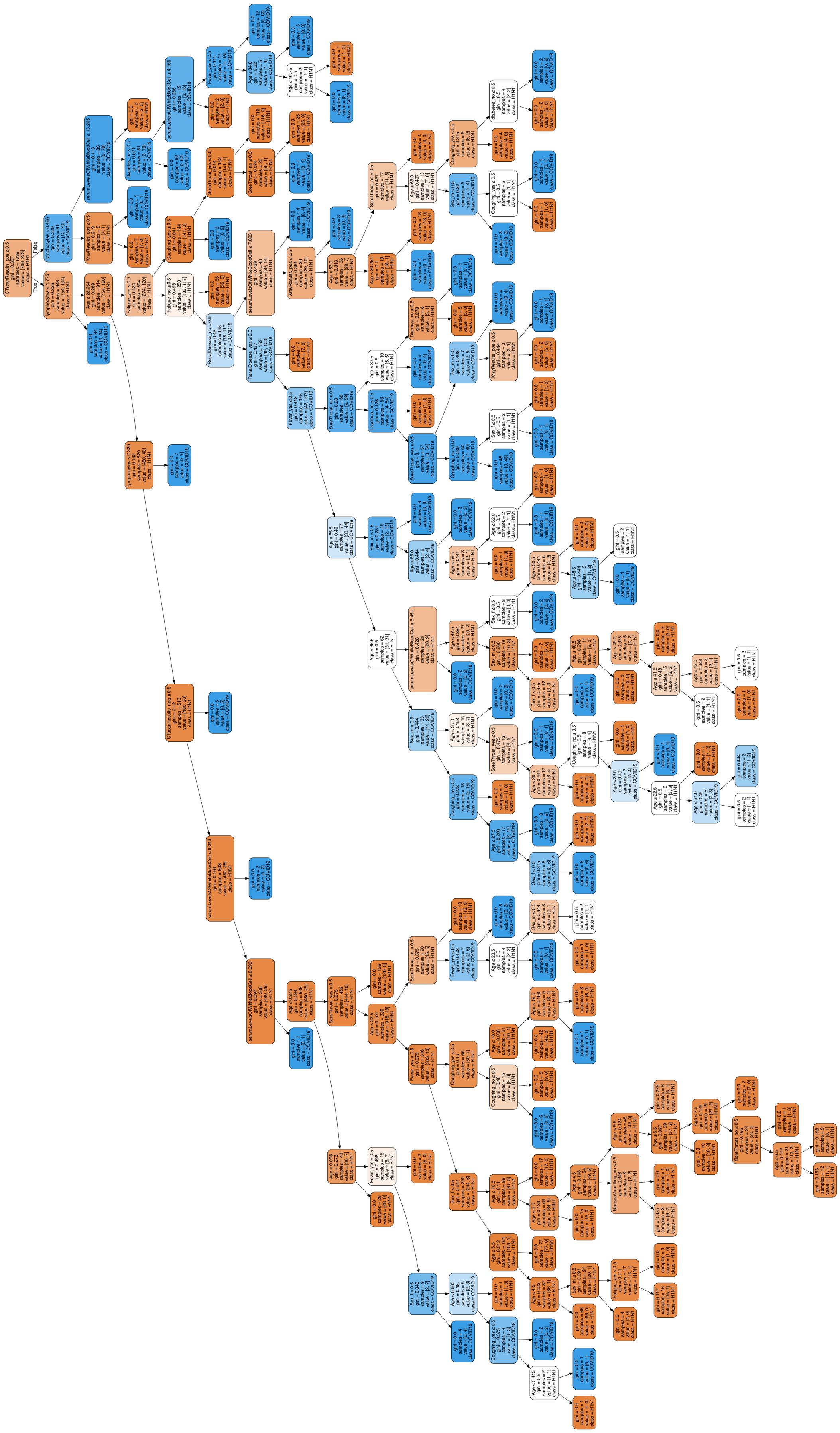

Figure 3: A generated decision tree based off of the algorithm using samples from the Covid and Flu 2019 data census.