Aston University

BIRMINGHAM UK

CS4850 Data Mining

Kenny Kei Yun Sum

Student ID : 230215740

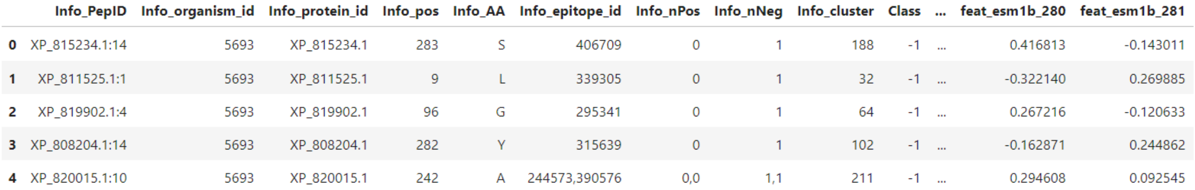
Data Mining Coursework

Analysis Summary

Introduction

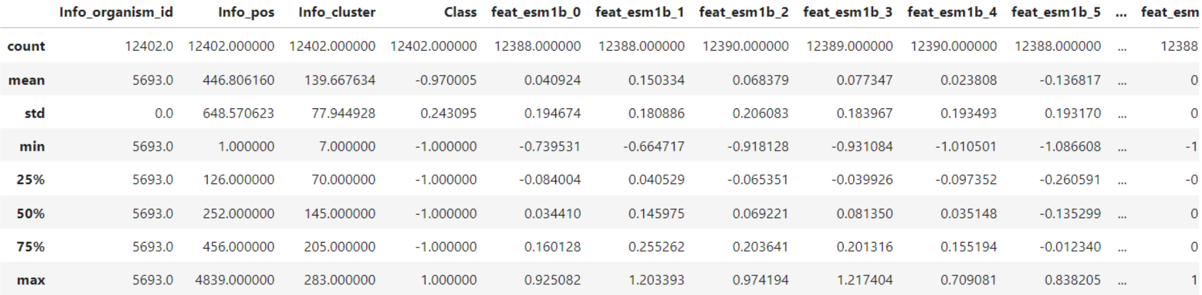
The purpose of this summary is to analysis the data mining pipeline used to efficiently predict new, previously unknown epitopes in the proteins of this virus. This is done by analysing the data given to us by various databases and evaluating the methods used.

Exploratory data analysis



	Info_PepID	Info_organism_id	Info_protein_id	Info_pos	Info_AA	Info_epitope_id	Info_nPos	Info_nNeg	Info_cluster	Class	...	feat_esm1b_280	feat_esm1b_281
0	XP_815234.1:14	5693	XP_815234.1	283	S	406709	0	1	188	-1	...	0.416813	-0.143011
1	XP_811525.1:1	5693	XP_811525.1	9	L	339305	0	1	32	-1	...	-0.322140	0.269885
2	XP_819902.1:4	5693	XP_819902.1	96	G	295341	0	1	64	-1	...	0.267216	-0.120633
3	XP_808204.1:14	5693	XP_808204.1	282	Y	315639	0	1	102	-1	...	-0.162871	0.244862
4	XP_820015.1:10	5693	XP_820015.1	242	A	244573,390576	0,0	1,1	211	-1	...	0.294608	0.092545

Figure 1 – df.head()



	Info_organism_id	Info_pos	Info_cluster	Class	feat_esm1b_0	feat_esm1b_1	feat_esm1b_2	feat_esm1b_3	feat_esm1b_4	feat_esm1b_5	...	feat_esm
count	12402.0	12402.000000	12402.000000	12402.000000	12388.000000	12388.000000	12390.000000	12389.000000	12390.000000	12388.000000	...	12388
mean	5693.0	446.806160	139.667634	-0.970005	0.040924	0.150334	0.068379	0.077347	0.023808	-0.136817	...	0
std	0.0	648.570623	77.944928	0.243095	0.194674	0.180886	0.206083	0.183967	0.193493	0.193170	...	0
min	5693.0	1.000000	7.000000	-1.000000	-0.739531	-0.664717	-0.918128	-0.931084	-1.010501	-1.086608	...	-1
25%	5693.0	126.000000	70.000000	-1.000000	-0.084004	0.040529	-0.065351	-0.039926	-0.097352	-0.260591	...	-0
50%	5693.0	252.000000	145.000000	-1.000000	0.034410	0.145975	0.069221	0.081350	0.035148	-0.135299	...	0
75%	5693.0	456.000000	205.000000	-1.000000	0.160128	0.255262	0.203641	0.201316	0.155194	-0.012340	...	0
max	5693.0	4839.000000	283.000000	1.000000	0.925082	1.203393	0.974194	1.217404	0.709081	0.838205	...	1

Figure 2 – df.describe()

Investigation and analysis were done on the df_reduced dataset. To understand the data given, multiple functions were used to investigate the big data. Functions such as df.shape, df.head(), df.columns, df.dtypes and df.describe() were used. Results of these functions are shown in figure 1 & 2.

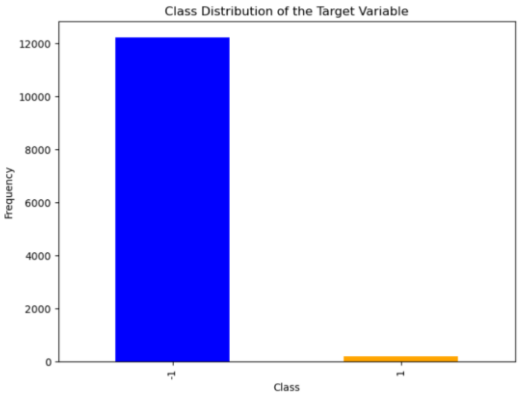


Figure 3 – Class Distribution between class 1 and class -1

Further investigation shows that the data is separated into two types of classes (class 1 and class -1). A histogram plot shows that the dataset contains data with the majority in class -1 with a value count of 12216, compared to class 1 of value count of 186 (figure 3). This is done to examine the differences in the number of counts between classes.

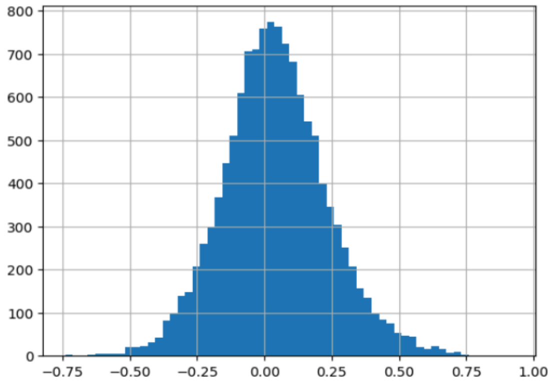


Figure 4 – Histogram plot of the mean value distribution between counts.

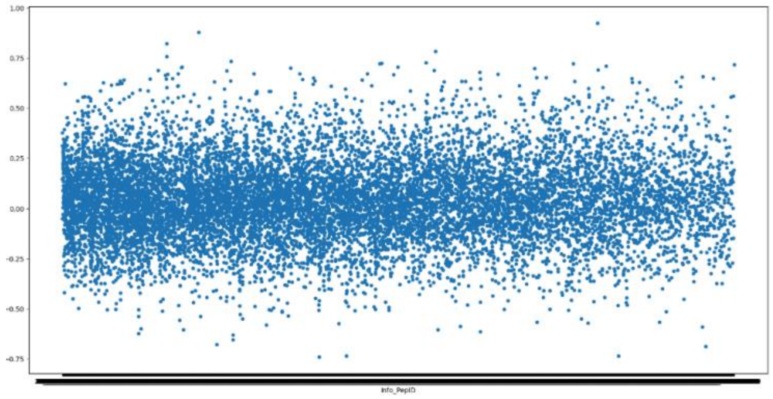


Figure 5 – Scatterplot of the mean value distribution between counts.

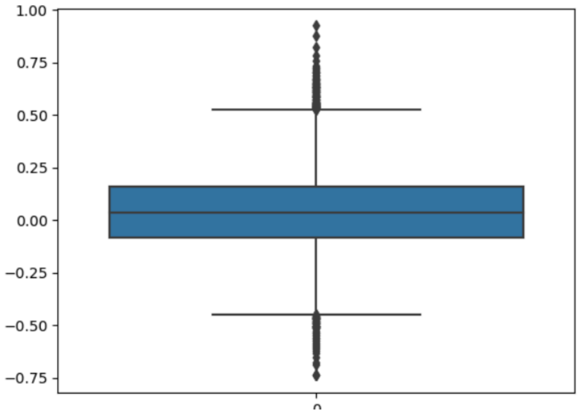


Figure 6 – Boxplot of the mean value distribution between counts.

In addition, a function was used to find the mean value distribution between counts of each column (figure 4, figure 5 and figure 6). The purpose of this examination is to understand the relationship between feats. The plots shown in figure 4, 5 & 6, reveals that most of the mean values of counts are distributed between -0.25 and 0.25 with the number of mean counts decreasing outside of this range. The plot also shows extreme outliers around the values of 0.9 and -0.9 on the y- axis (figure 5).

Data Preprocessing

As shown on figure 3, the plot shows that the dataset contains two distinct class values (class = 1 and class = -1). In additional, it shows that the plot shows an extreme difference in number of counts between both class (class =1 containing 186 and class = -1 containing 12216). Because of these differences the dataset was split into two categories known as class 0 (containing class = -1 values) and class 1 (containing class = 1 values) during the data preprocessing stage. This was done to examine if there is a difference between the two classes, which will be done later during the data processing stage.

Afterwards, a code was created to find the missing values of both groups that are greater then 20 for each column. As a result, these columns were removed due to the number of missing values which may lead to inaccurate analysis and distort the true pattern and relationship within the dataset. Another code was created to fill in any missing values for each column by the mean value for both groups. To further clean the data, an interquartile range method was used to detect outliers outside the first and third quartile for each column and was then removed. Outliers were removed to reduce noise, enhance accuracy of data and allow better generalization.

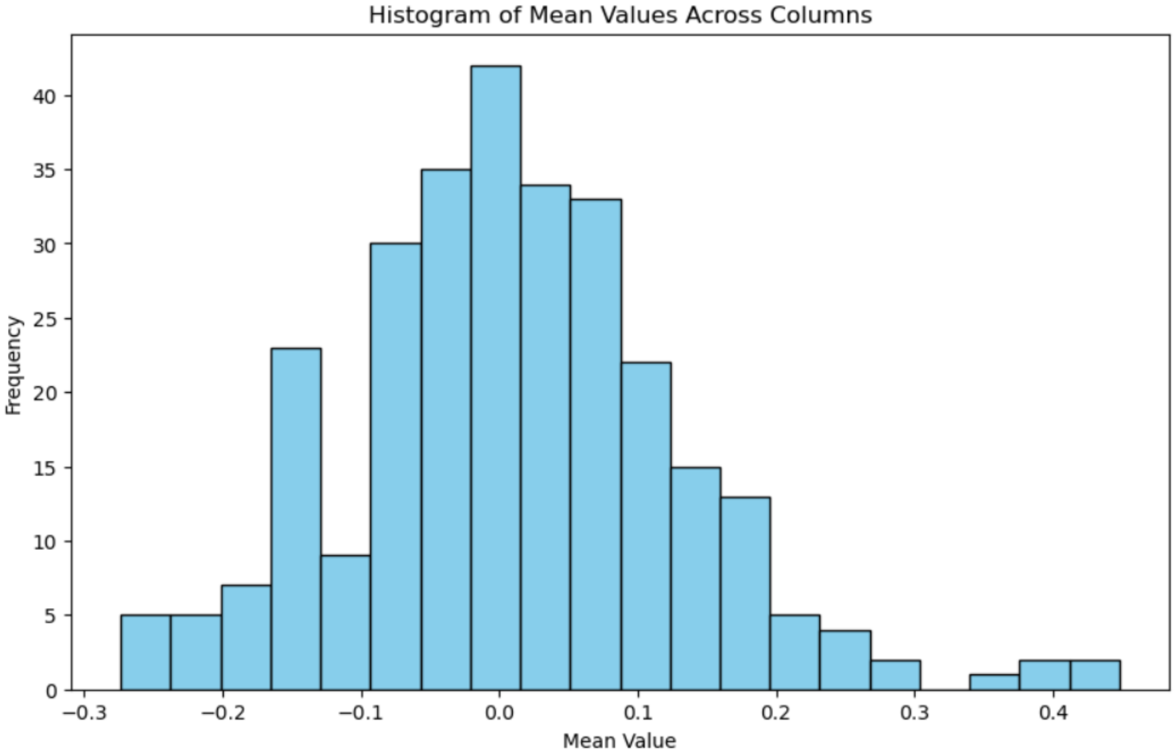


Figure 7 – Mean value distribution between counts for class 0 (class = -1).

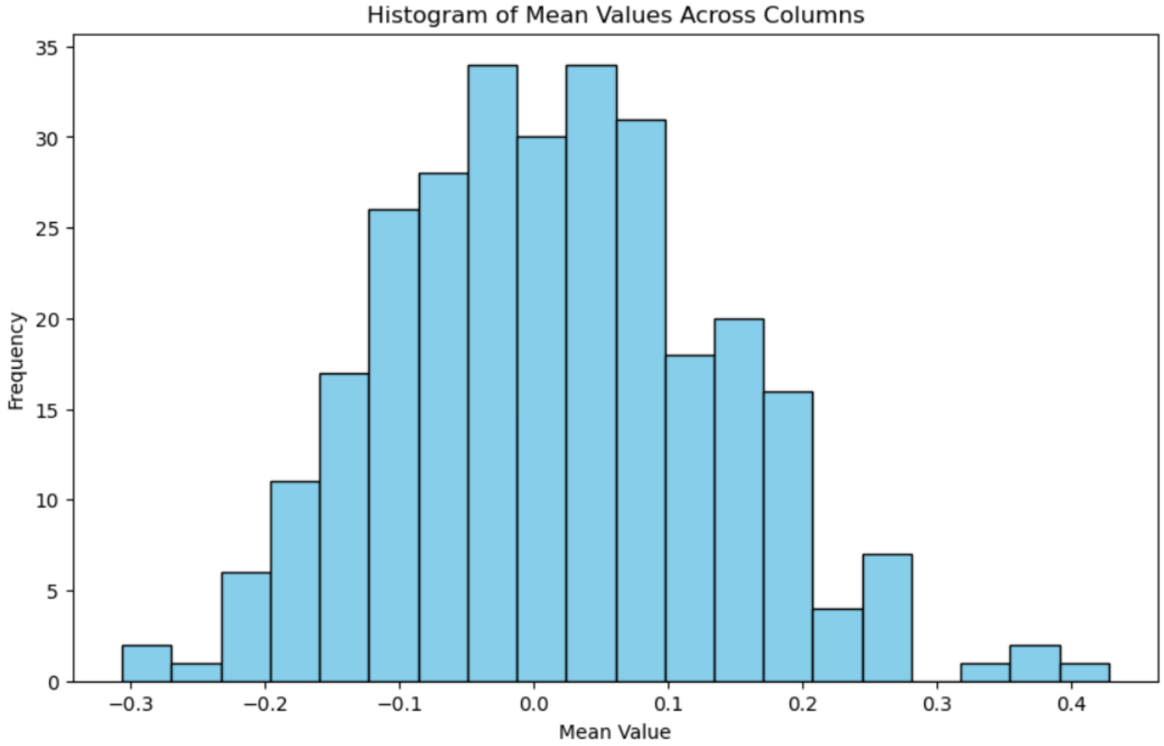


Figure 8 – Mean value distribution between counts for class 1 (class = 1).

A histogram plot was created for both classes (class 0 and class 1), showing the mean value distribution between counts. This was done to compare the difference in distribution of mean values between both classes (figure 7 & figure 8). Both plots show that distribution of mean values are very similar within the range between -0.25 and 0.25.

Resampling

As shown on figure 3, there is a large difference in number of counts between class 0 (class = -1) and class 1 (class = 1). Class 0, containing 12216 values compared to class 1 which only contains 186. After filtering each class, the number of counts was reduced. For class 0 (class = -1), the number of counts was reduced from 12216 to 4927 and for class 1 (class = 1), the number of counts was reduced from 186 to 39. An up-sample function was therefore used to randomly increase the number of counts for class 1 (class =1) to match the number of counts for class 0 (class =-1). This was done to balance the two classes and reduce bias towards class 0, since it contains a significant number of counts in comparison. Both datasets were concatenated together and was then split for training and testing the model.

Random Forest

The random forest regressor was used for predicting continuous outcomes. Random forest was chosen due to its high accuracy, robustness to overfitting and being capable of handling high-dimensional data.

Evaluate the Model

An evaluation of the model was also done which gave a mean squared error: 0.03, mean absolute error: 0.13, root mean squared error: 0.16 and R^2 score: -0.06.

Pipeline

A pipeline was also created for automating and streamlining the sequence of data processing steps and model training processes.