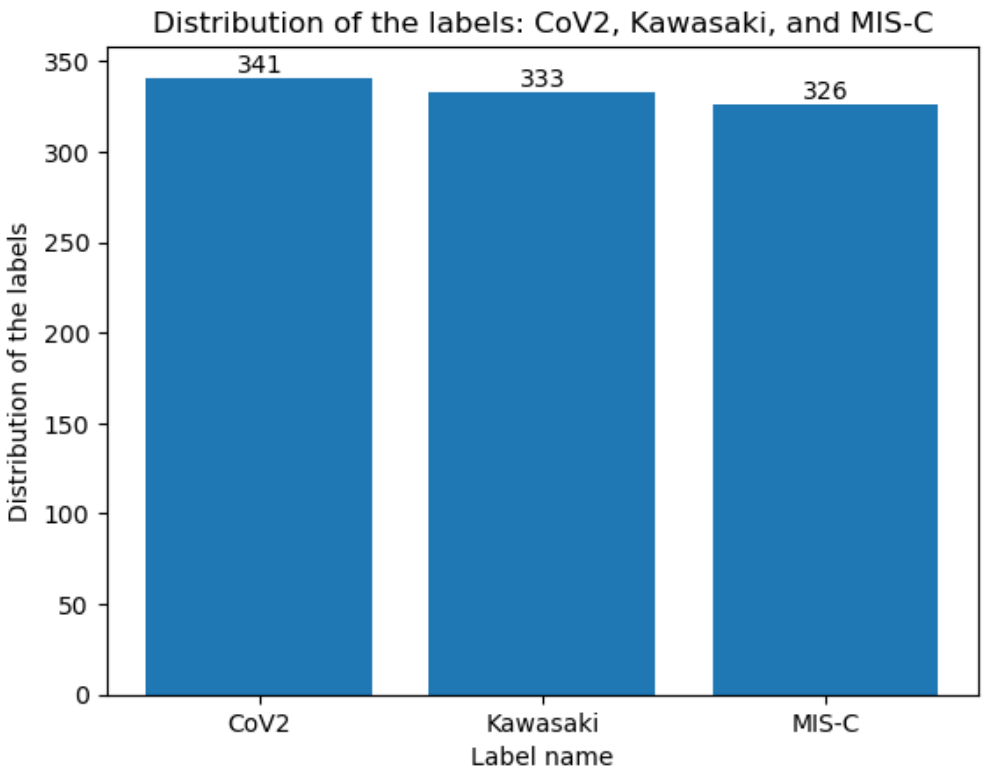
Anushka Sinha

02718 HW 2

Oct 10, 2023

1. **Check the distribution of the labels and generate a barplot to show the counts of the labels. Is this dataset balanced? (5 points)**



Yes, this dataset is balanced.

2. **The filter-based feature selection**

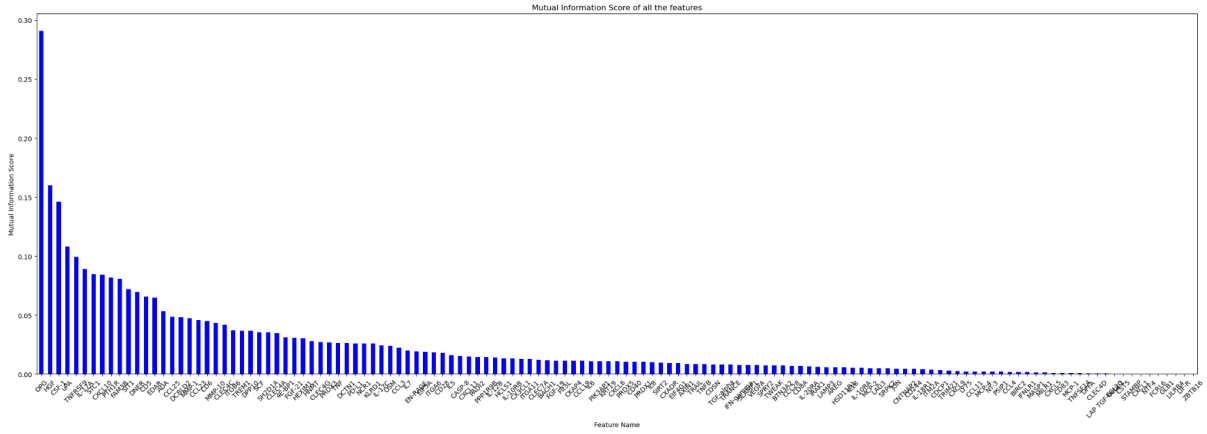
The final column in the file is a label that indicates which condition the patient has at the time of data collection. You will perform different feature selection methods to identify inflammatory biomarkers using multiple feature selection techniques that can distinguish between one of three conditions in children:

a) Infection by the SARS-CoV-2 virus

b) A rare, but potentially deadly complication of COVID-19, called Multisystem Inflammatory Syndrome in Children (MIS-C)

c) Kawasaki disease, a potentially deadly syndrome of unknown cause

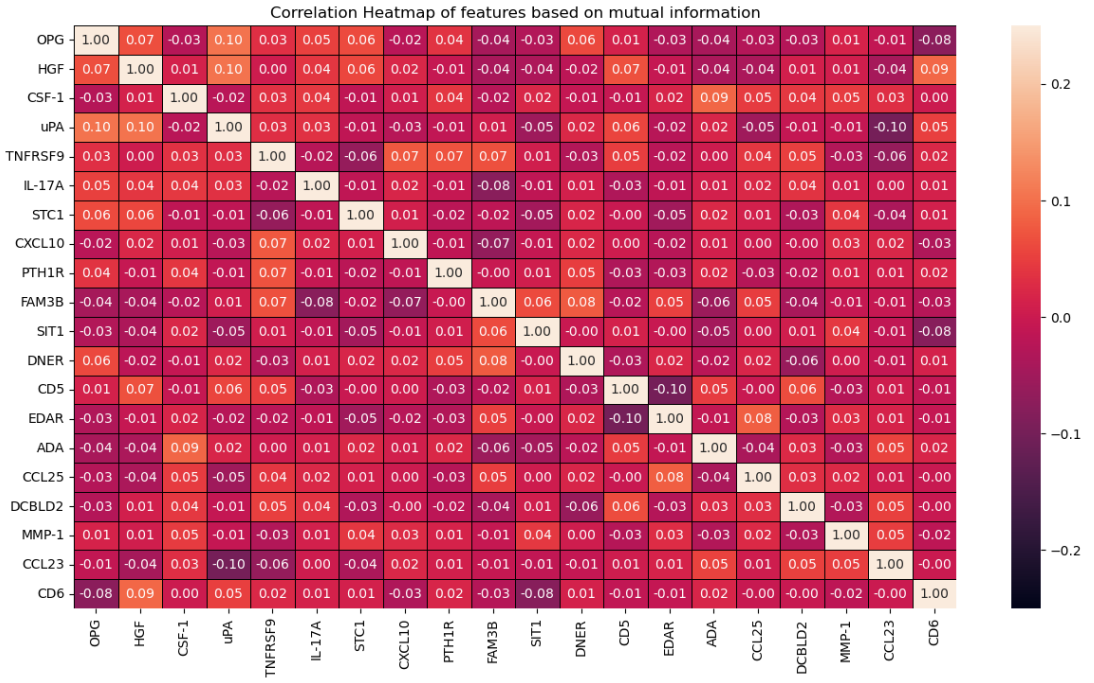
2.1 (10 points) First, apply two filter-based feature selection methods to the data using 10-fold cross validation. Write codes to implement Mutual Information and another method of your choice. Visualize the features based on their mutual information scores and paste your plot here.



2.2 (5 points) List at least two other filter-based selection methods other than Mutual Information and briefly introduce each method in one or two sentences.

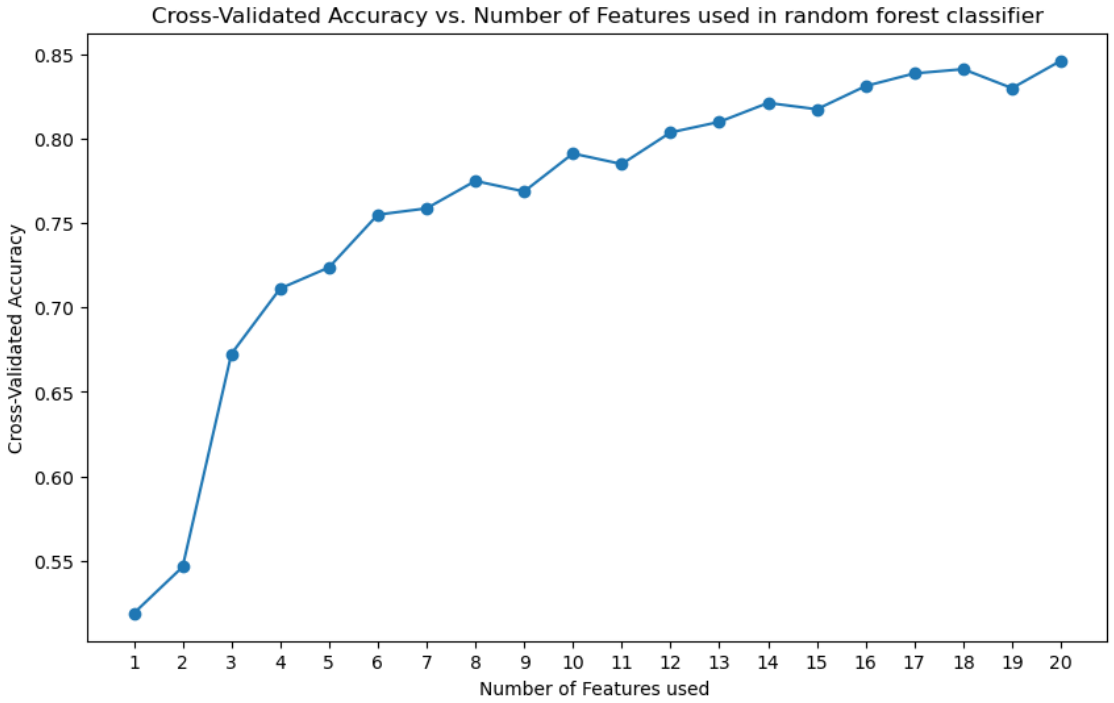
(i) ReleifF: the relief algorithm is used to evaluate the quality of the features by examining their ability to distinguish between conditions/instances that are close to each other. It calculates a score for every feature which is then used for ranking and can then later be used to select top scoring features for feature selection. (ii) t-statistic: t-test can be used to determine if a particular feature significantly differs between different conditions/labels.

2.3 (5 points) Create and plot a 20 by 20 correlation heatmap using the top 20 features based on Mutual Information. What is the average of the values in the heatmap?



Average obtained is 0.054

2.4 (5 points) Train classifiers using the top 1, 2, 3, …, 20 features based on the mutual information. Plot the 10-fold cross validated accuracy of the classifiers as a function of the number of features.



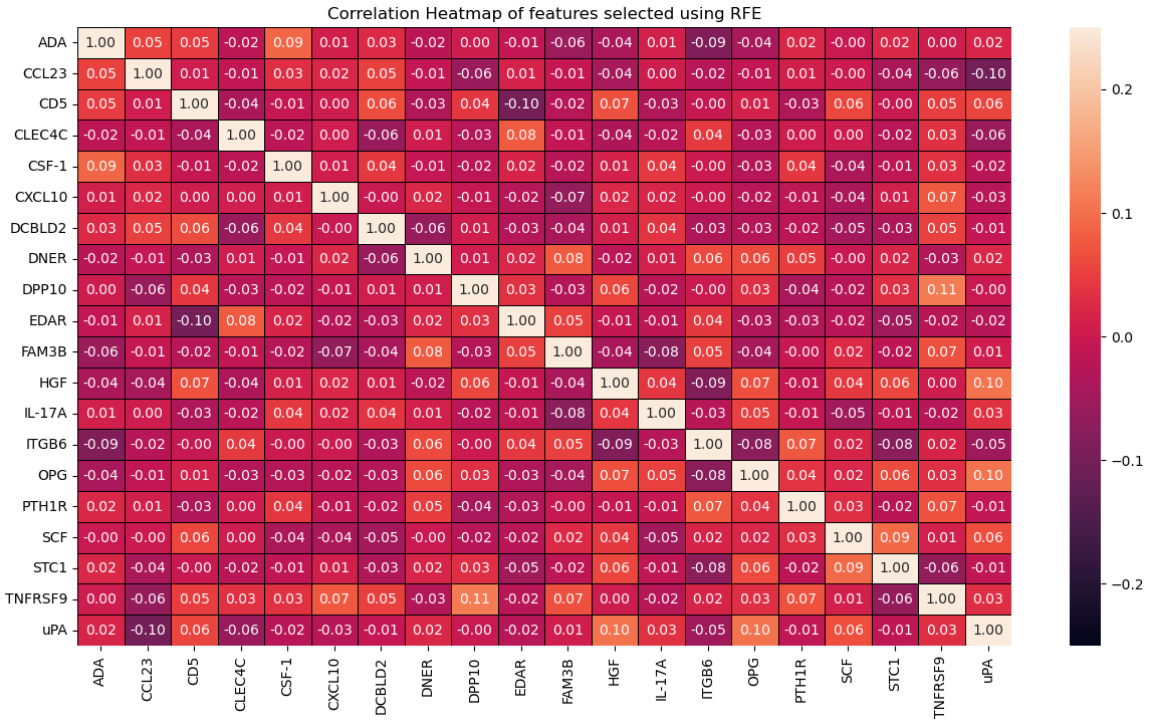
3. The wrapper-based feature selection

3.1 (10 points) Next, choose a wrapper-based feature selection method and apply it to the data. No need to answer Question 3.1 in the written part.

3.2 (5 points) List the top 20 features you selected using the wrapper-based feature selection method.

['ADA', 'CCL23', 'CD5', 'CLEC4C', 'CSF-1', 'CXCL10', 'DCBLD2', 'DNER', 'DPP10', 'EDAR', 'FAM3B', 'HGF', 'IL-17A', 'ITGB6', 'OPG', 'PTH1R', 'SCF', 'STC1', 'TNFRSF9', 'uPA'],

3.3 (5 points) Create and plot a correlation heatmap using the features from part 3.1. What is the average of the values in the heatmap?



Average obtained is 0.05176

3.4 (5 points) Train a classifier using the features from Question 3.1. Report the 10-fold cross-validated accuracy of the classifier.

The 10-fold-cross-validated accuracy of the classifier is 0.845.

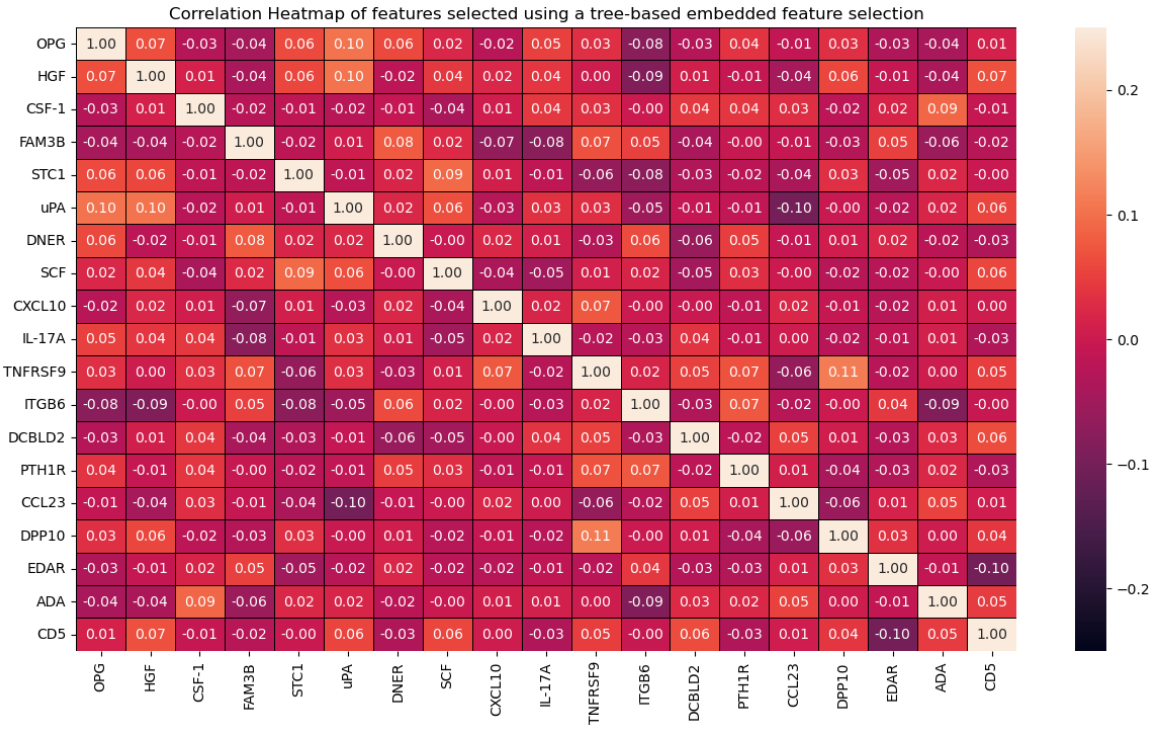
4. The embedded feature selection method

4.1 (10 points) Finally, implement a tree-based classification method to the data using 10-fold cross validation. No need to answer Question 4.1 in the written part.

4.2 (5 points) List the features that are selected in at least 8 out of 10 folds of the cross validation.

['OPG', 'HGF', 'CSF-1', 'FAM3B', 'STC1', 'uPA', 'DNER', 'SCF', 'CXCL10', 'IL-17A', 'TNFRSF9', 'ITGB6', 'DCBLD2', 'PTH1R', 'CCL23', 'DPP10', 'EDAR', 'ADA', 'CD5']

4.3 (5 points) Create and plot a correlation heatmap using the features from Question 4.1. What is the average of the values in the heatmap?



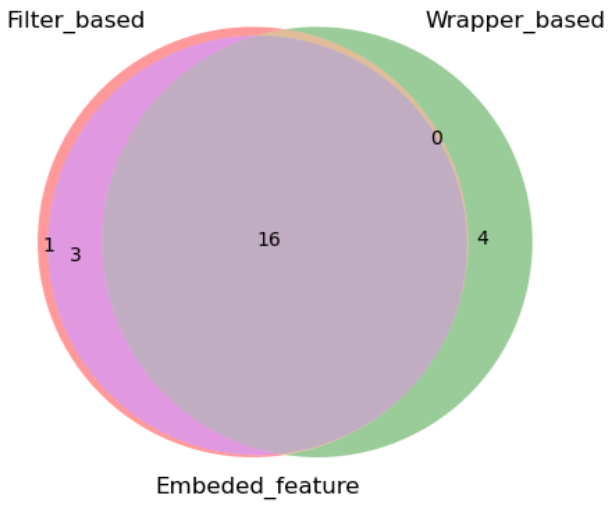
The average obtained is 0.0555

4.4 (5 points) Train a classifier using the features from Question 4.1. Report the 10-fold cross-validated accuracy of the classifier.

The 10-fold cross-validated accuracy of the classifier is 0.85.

5 Model Comparison and Interpretability

5.1 (10 points) Are there any features that are consistently selected across the filter-based (considering only the mutual information method), wrapper-based, and embedded feature selection methods? Illustrate the overlaps using a Venn diagram. If there are any, please list the names of those features.



There are 16 features that are consistently selected across the filter based, wrapper based, and embedded feature selection methods. Here are those 16 features:

{'uPA', 'HGF', 'FAM3B', 'CD5', 'STC1', 'EDAR', 'CCL23', 'CXCL10', 'PTH1R', 'TNFRSF9', 'CSF-1', 'DNER', 'IL-17A', 'ADA', 'OPG', 'DCBLD2'}

There are 16 features that are selected in both filter based and wrapper based methods. Here are those 16 features:

{'uPA', 'HGF', 'CD5', 'FAM3B', 'STC1', 'EDAR', 'CCL23', 'CXCL10', 'PTH1R', 'TNFRSF9', 'CSF-1', 'DNER', 'IL-17A', 'ADA', 'OPG', 'DCBLD2'}

There are 19 features that are selected in both filter based and tree based methods. Here are those 19 features:

{'SCF', 'CCL23', 'EDAR', 'TNFRSF9', 'CSF-1', 'DNER', 'OPG', 'uPA', 'HGF', 'CD5', 'STC1', 'CXCL10', 'ADA', 'FAM3B', 'ITGB6', 'PTH1R', 'DPP10', 'IL-17A', 'DCBLD2'}

There are 16 features that are selected in both wrapper based and tree based methods. Here are those 16 features:

{'uPA', 'HGF', 'CD5', 'FAM3B', 'STC1', 'EDAR', 'CCL23', 'CXCL10', 'PTH1R', 'TNFRSF9', 'CSF-1', 'DNER', 'IL-17A', 'ADA', 'OPG', 'DCBLD2'}

5.2 (5 points) From this list, select at least two features to discuss their potential clinical significance. Recall that you are trying to identify inflammatory biomarkers. If no features overlap between your methods, choose from the top-ranked features of any method.

OPG: It stands for Osteoprotegerin which is a biomarker for inflammatory bowel disease and gastrointestinal carcinoma. It is a part of the tumor necrosis factor receptor superfamily of proteins and plays an important role in inflammatory pathways and tumor cell survival. It is a clinically significant inflammatory biomarker because it influences cell turn-over, cell differentiation, cell death, and cell survival through extracellular pathways which is associated with a less favorable prognosis in inflammatory bowel disorders and a number of gastrointestinal carcinomas.

Reference used: https://pubmed.ncbi.nlm.nih.gov/26896745/

CD5: It is a lymphocyte surface marker and immunomodulator involved in the development, activation, differentiation, and survival of lymphocytes. It is also involved in the fine tuning of TCR signaling and may serve as a prognosis biomarker in resectable stages of non-small cell lung cancer (NSCLC). It is clinically significant because several research studies have shown that patients with higher CD5 expression had significantly increased overall survival.

Reference used: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7481598/

5.3 (5 points) Based on the results of your feature selection methods and subsequent classification performance, which method do you think is the best? Provide justifications for your choice. (Hint: You could consider classification accuracy, computational efficiency, interpretability, and stability of the selected features.)

Based on my results for feature selection methods and subsequent classification performance, I would recommend using the filter based selection method. Using a random forest classifier, I obtained classification accuracy of 0.85, 0.845, and 0.85 for filter-based, wrapper-based, and embedded tree-based feature selection methods respectively. The computational efficiency was very similar across different feature selection methods. However, there was potential for the wrapper based selection method to take a long time if SVC was used as an estimator. I also found the stability of the selected features in every feature selection method to be similar in the sense that if I run the code for selecting features multiple times, I would obtain similar results if not the same. The recommendation made here might not be entirely accurate because the classifier was trained by 19 features for the embedded tree based feature selection because we were particularly interested in obtaining features that were present in at least 8 out of the 10 folds. I recommend the filter-based feature selection method because I found it easier to interpret. It was easier for me to understand because I have come across concepts of mutual information gain and chi square test before.