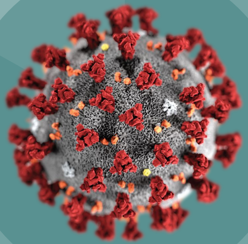
Supervised Learning



Disease Classification based on symptoms (Covid-19, Flu, Cold, Allergy)

Professor: Luís Paulo Reis

Group members:

Eduardo Brito, up201806271
Paulo Ribeiro, up201806505
Rita Silva, up201806527

Specification

Covid-19, the common **Cold**, **Seasonal Allergies** and the **Flu** have many similar signs and symptoms. These common problems are often mistaken for Covid-19 and this project will help provide a distinction between them.

Based on a data set with information about some patients' diagnosis and the experienced symptoms, our goal is to associate them and understand their relationship in order to help diagnose new patients.

Hereupon, we identify this as **a single label multiclass classification problem**, with 21 attributes:

- 20 distinct symptoms, with a value of 1 if the patient suffers from it and 0 otherwise
- 1 diagnose with 4 possible outcomes (Covid-19, Cold, Allergies and Flu).

SORE THROAT	TIREDNESS	RUNNY NOSE	LOSS SMELL	...	ITCHY NOSE	SNEEZING	PINK EYE	TYPE
0	1	1	0	...	1	0	1	ALLERGY
0	1	0	0	...	1	1	0	COVID
1	1	0	0	...	1	0	0	COLD
1	0	1	1	...	0	1	0	FLU
0	1	1	1	...	0	1	1	ALLERGY

Tools & Resources

jupyter

Programming Language: Python

Development Environment: Visual Studio Code & JupyterLab

Data Set: https://www.kaggle.com/walterconway/covid-flu-cold-symptoms

Data Preprocessing:

pandas NumPy

Data Visualization:

seaborn matplotlib

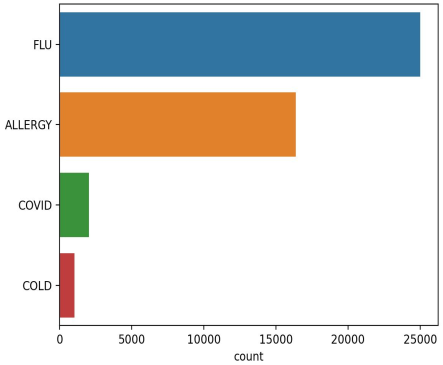
ML Algorithms:

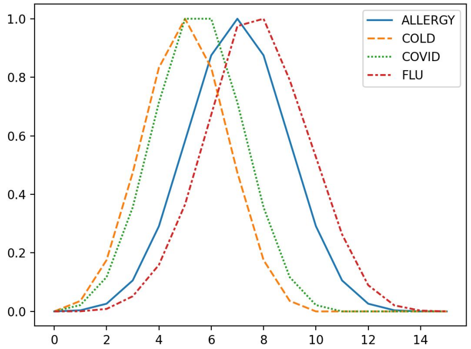
scikit learn

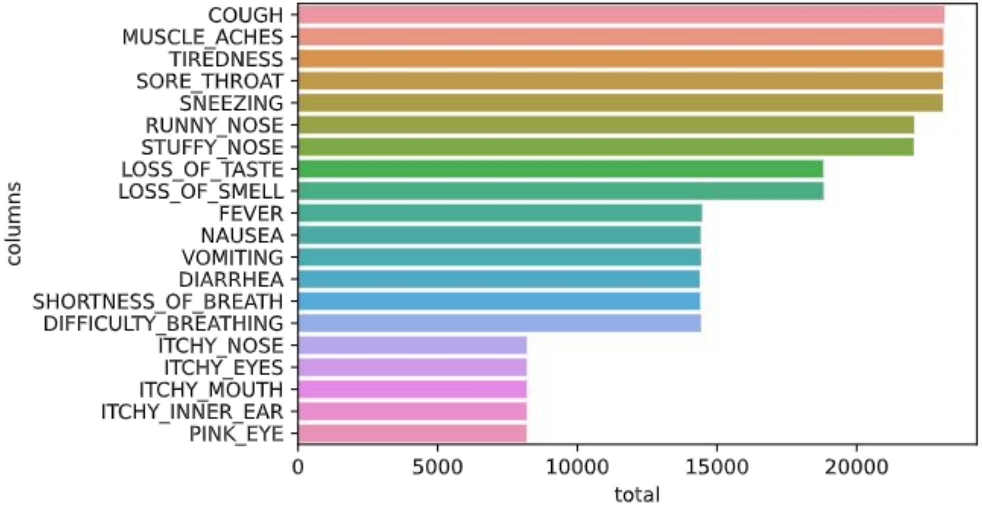
Data Analysis

Our problem presents some properties:

- Nominal and Discrete binary attributes
- Dimensionality = 21 attributes
- Size = 44k records
- Type = Data Matrix
- No meaningful outliers
- No missing or duplicate Data
- Similarity of around 55% (Hamming Distance)





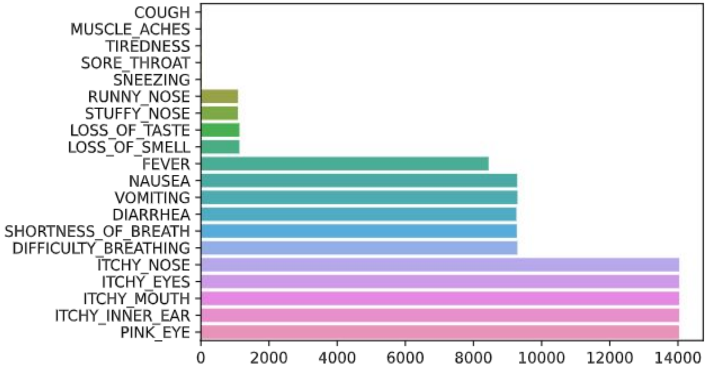


Data Analysis

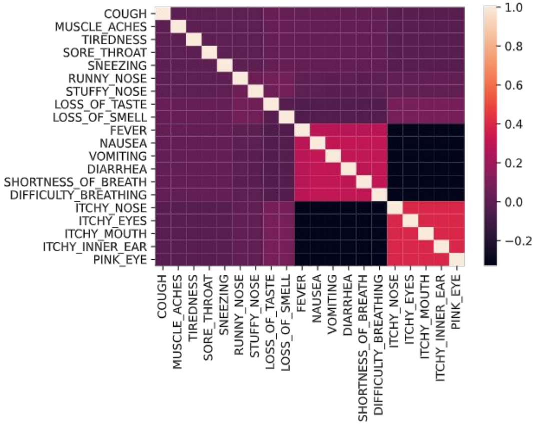
- The first graphic presents the relevance of each symptom in relation to the class column, obtained using the **SelectKBest** algorithm, with the **chi2** as score function (best for boolean attributes).

- The second graphic shows the **correlation** for each pair of symptoms, where the brightest colour correspond to higher correlation levels and the darkest to the lower ones.

Symptoms relevance

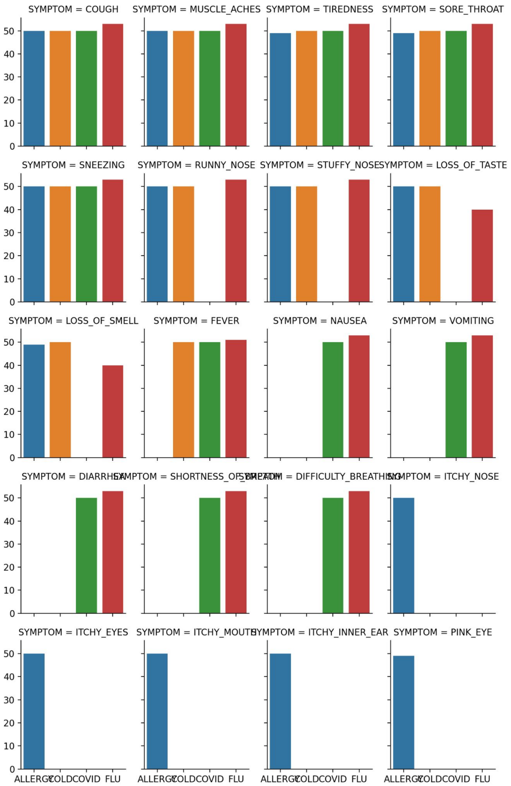


Symptoms correlation



Data Preprocessing

Symptoms relations with diseases



By analysing the following graphic, we conclude that there are some symptoms with similar relations with the diseases, which could be aggregated into one single attribute, to reduce the dimensionality of the problem.

Techniques:

- **Aggregation:**

  The new columns will be the result of a *Logical OR* applied to the aggregated columns. We decided, based on the plots, to aggregate the NAUSEA & VOMITING related symptoms as well as all the ITCHY ones into new columns. This action helped us reducing the total number of features from 21 to 13.

- **Encoding:**

  Using the *LabelEncoder* function from scikitlearn library, we also encoded the *class* column, mapping the strings to representative numbers, to correctly feed the supervised learning algorithms.

- **Train and Test split:**

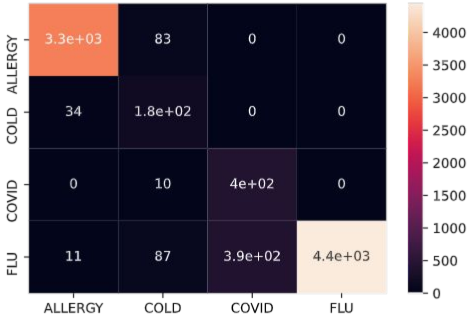
  After the data preprocessing, it is finally ready to be splitted into Train and Test sets, with a respective percentage of 80/20 of the total rows.

Decision Tree & Random Forest

As tuning approaches we decided to use the scikitlearn functions **StratifiedKFold** and **GridSearchCV** on our training process.

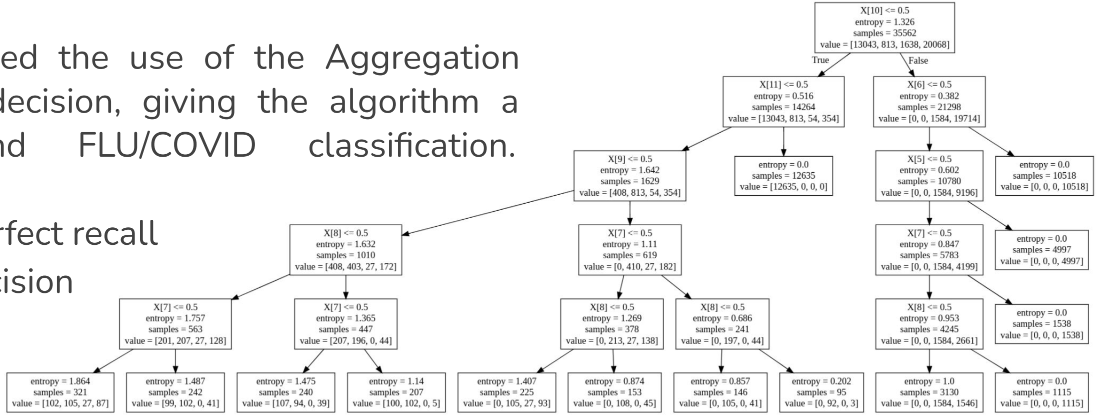
The best parameters found were:

- criterion: entropy
- max_depth: 5
- max_features: 12
- splitter: best



Based on the tree generated, we noticed the use of the Aggregation columns for the first depth levels of decision, giving the algorithm a quicker answer for ALLERGY and FLU/COVID classification.

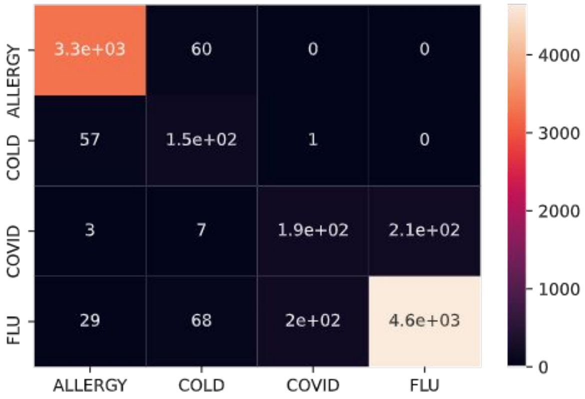
The confusion matrix shows an almost perfect recall for COVID classification and a perfect precision detecting FLU and ALLERGY.



Support Vector Machine (SVM)

The best parameters found for this algorithm were:

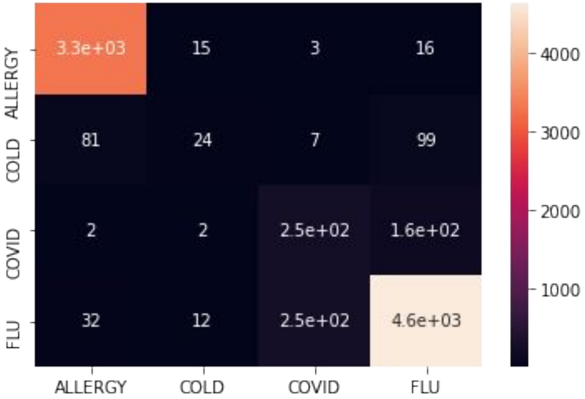
- kernel: linear



K-Nearest Neighbors (k-NN)

The best parameters found for this algorithm were:

- algorithm: brute
- n_neighbors: 5
- weights: uniform



Although this methods' accuracy was high, their results were not the best, since they were really bad at detecting COVID/COLD cases, due to the low precision. Plus, these two methods took a lot longer to execute.

Neural Networks

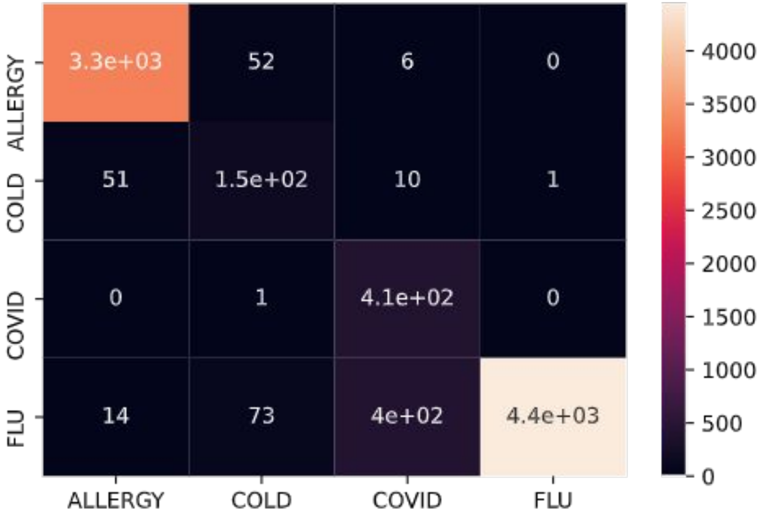
As tuning approaches we decided to use the scikitlearn functions **StratifiedKFold** and **GridSearchCV** on our training process.

The best parameters found for this algorithm were:

- activation: tanh
- hidden_layer_sizes: (3, 5, 8, 13, 21, 34)
- solver: adam

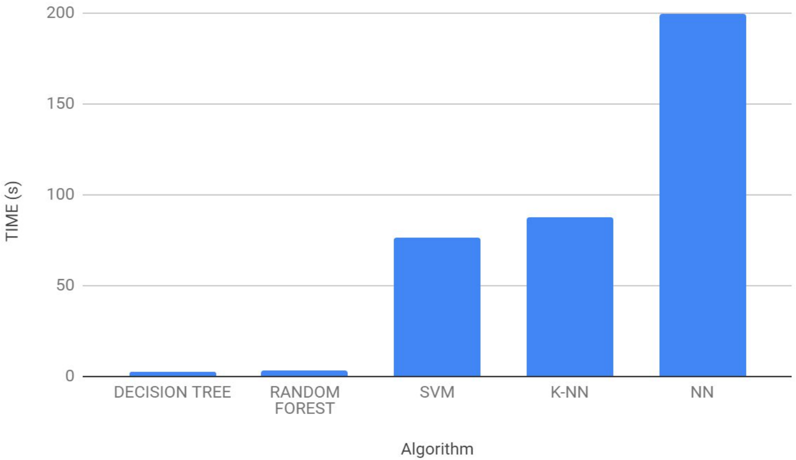
This was our best model for detecting COVID with a recall of 100%. FLU also got a maximum precision, although a lot of patients with that disease were mislabelled with COVID.

One disadvantage of this algorithm is the time taken when executing, which is 200x greater than a simple Decision Tree.

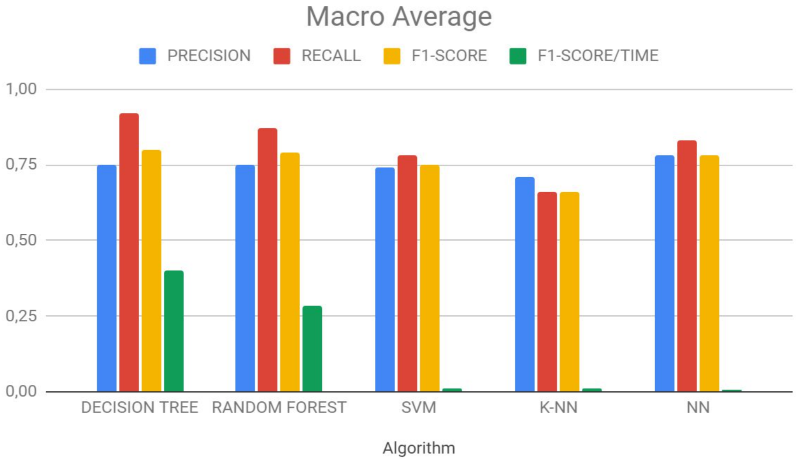


Algorithms Comparison

Time Spent (Training)



Evaluation Metrics



By analysing these graphics, we conclude that the most efficient algorithm is the DECISION TREE, since it achieves the best results in the shortest time. This might be due to the nature of our problem, which can be represented by a simple set of Yes/No decisions, easily represented by a tree.

Bibliography

Kaggle Data Set. **COVID, FLU, COLD Symptoms.** April, 2021.

Mayo Clinic. **COVID-19, cold, allergies and the flu: What are the differences?** May 06, 2021.

Walter Conway. **Symptom Generator.** April, 2021.

Edureka. **Classification Algorithms.** November, 2020.

DataScience StackExchange. **When should I use Gini Impurity as opposed to Information Gain (Entropy)?** February, 2016.