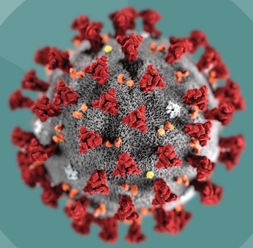
Supervised Learning

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

Programming Language: Python

Development Environment: Visual Studio Code & JupyterLab

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

jupyter

Data Preprocessing:

pandas NumPy

Data Visualization:

seaborn matplotlib

ML Algorithms:

scikit learn

Data Analysis & Preprocessing

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)

Preprocessing Techniques:

- Aggregation (ex: “ITCHY” Symptoms)
- Sampling (ex: Stratified Sampling)
- Dimensionality Reduction (ex: SelectKBest chi2)
- Feature Creation (ex: Symptoms per Person)

Algorithms

Test & Tuning approaches:

- StratifiedKFold Validation
- Parameter Tuning (ex: GridSearch)

Machine Learning algorithms:

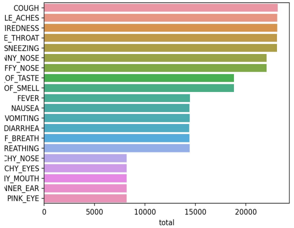
- Decision Trees
- Neural Networks
- K-NN (K-Nearest Neighbour)
- SVM (Support-vector machine)
- Random Forest

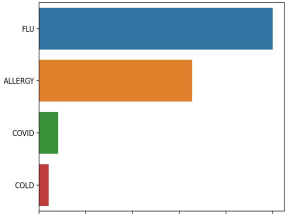
Evaluation metrics:

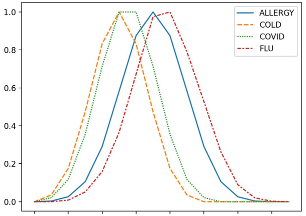
- Performance during learning
- Confusion matrix
- Precision, recall, accuracy
- F1 measure
- Time spent in train/test

Work already done

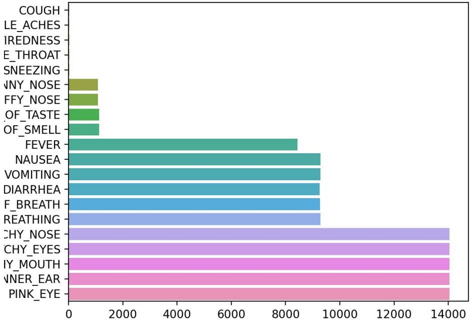
Symptoms and diseases counting

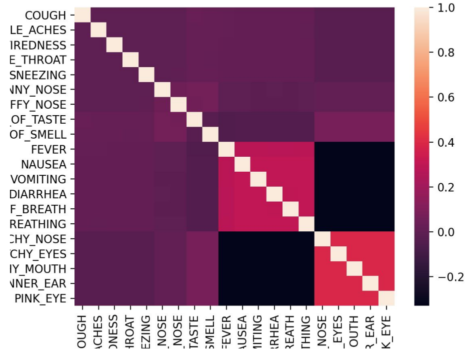




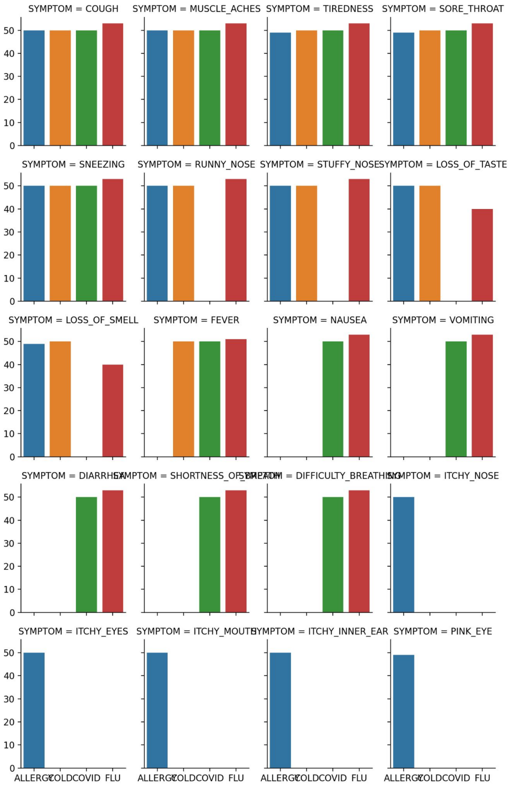


Symptoms relevance and correlation





Symptoms relations with diseases



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.