COVID, FLU, COLD and ALLERGY Symptoms

Grupo 42:

- Diogo Rosário (up201806582)
- Henrique Ribeiro (up201806529)
- Tiago Silva (up201806516)

Specification

The objective of this project is to figure out which of the four illnesses (covid, flu, cold and allergy) a patient has, given a set of symptoms.

Consulted Material

Libs

- MatplotLib
- Seaborn
- Scikit-learn
- Imblearn

Helpful Materials

- Label Encoding
- Guide on feature selection
- Data science cheat sheet
- Smote

Tools used

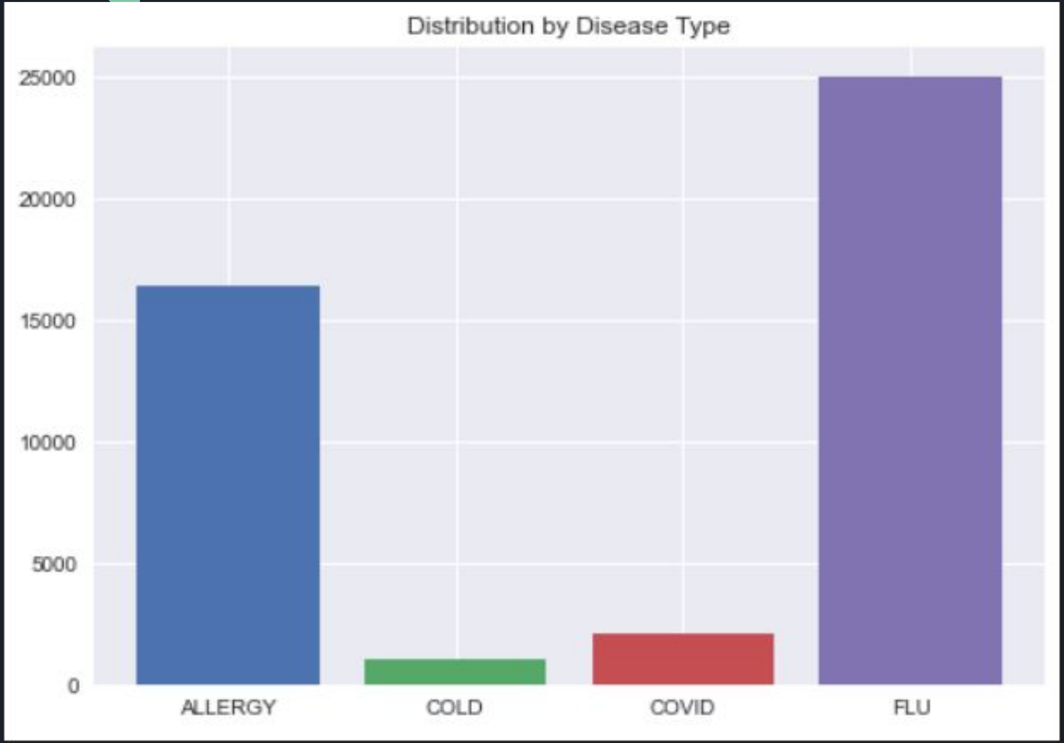
Pandas: to analyse data;

Scikit-Learn: a machine learning focused library with functions that help with various algorithms like SVM and random forest;

Seaborn/Matplotlib: to plot data;

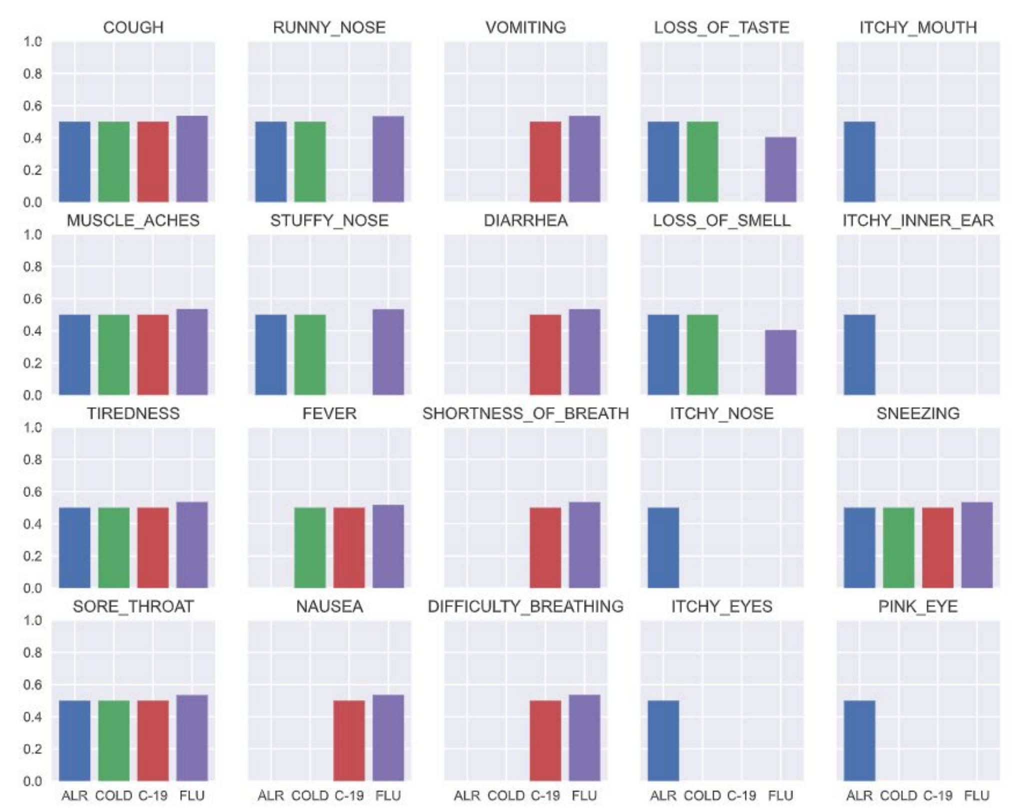
Imblearn: to deal with the unbalanced dataset;

Initial data distribution



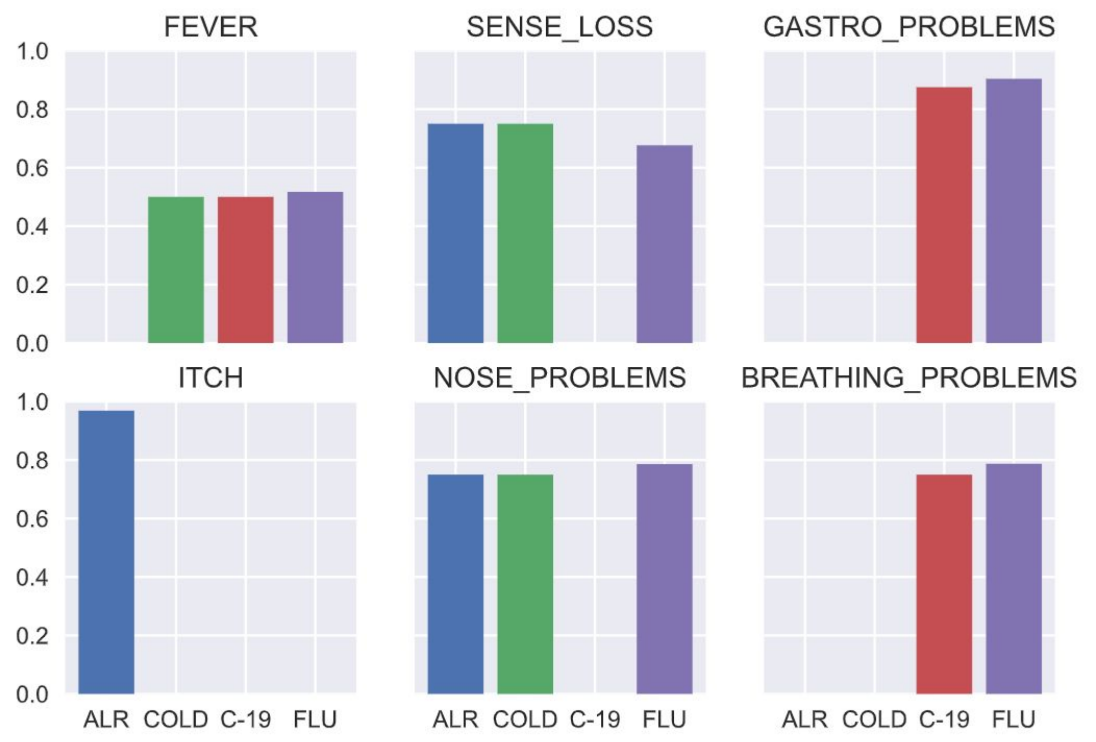
- Unbalanced dataset;
- Requires data treatment like “SMOTE”.

Filtering data - initial data



- All itches and Pink Eye can be merged into one column;
- Cough, Muscle Aches, Tiredness and Sore Throat do not give relevant information as they are present 50% of the times;
- Vomiting, Diarrhea, and Nausea can also be merged into one column;
- Shortness of breath and difficulty breathing can also be merged into one column;
- The Loss of Taste and Smell can also be merged into one column;
- Same goes for Runny Nose and Stuffy Nose;

Final data



- ⅕ number of columns
- 100% of the people with an itch have an allergy
- Gastro Problems and Breathing problems are very common for both Covid and Flu
- There are no people with covid that have sense loss nor nose problems

Algorithms used

- Random Forest: ensemble learning method for tasks that operates by constructing multiple decision trees during training.

- SVM (support-vector machines): one of the most robust prediction methods, being based on statistical learning frameworks.

- KNN: non-parametric classification method. Consists of the k closest training examples in the Dataset.

More data treatment

- Label Encoding:
  - Allergy was swapped by a 0
  - Cold was swapped by a 1
  - Covid-19 was swapped by a 2
  - Flu was swapped by a 3
- Dataset analysis (dataset head and describe)
- Dataset correlation analysis
- Oversampling ~ Smote

Algorithms used

Grid Search For:

- Random Forest
- SVM
- KNN

Classification with Smote For:

- Random Forest
- SVM
- KNN

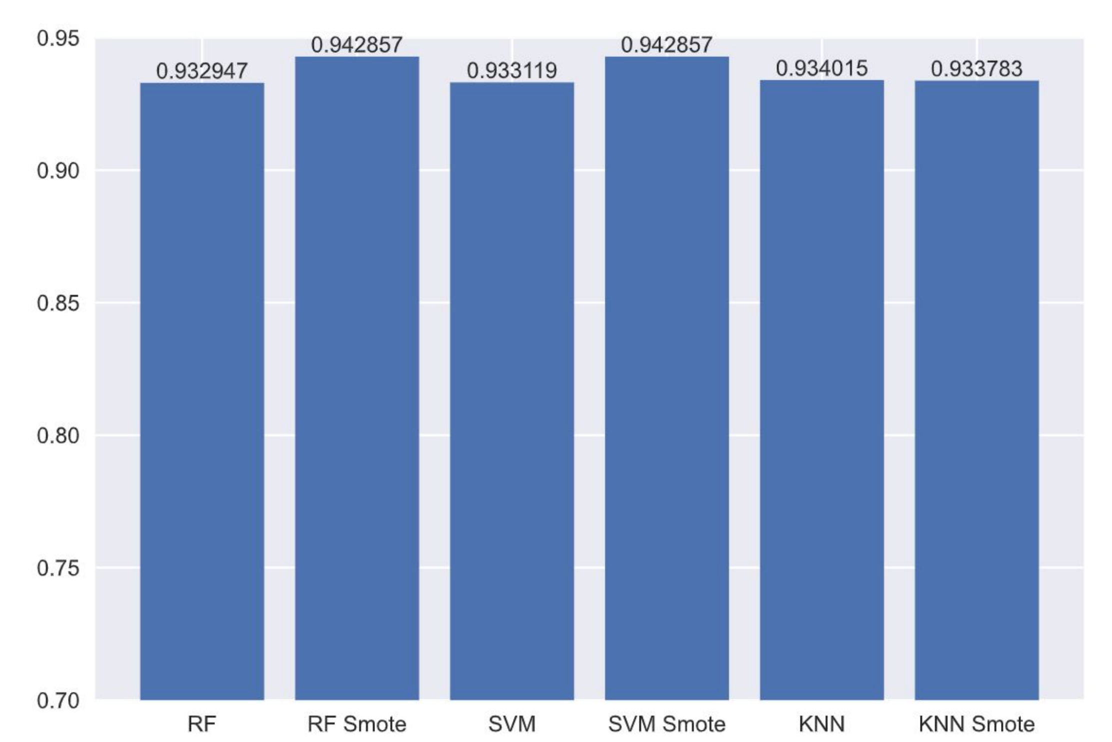
Cross Validation For:

- Random Forest
- SVM
- KNN

Data analysis and treatment
Data plotting

Values obtained for each algorithm

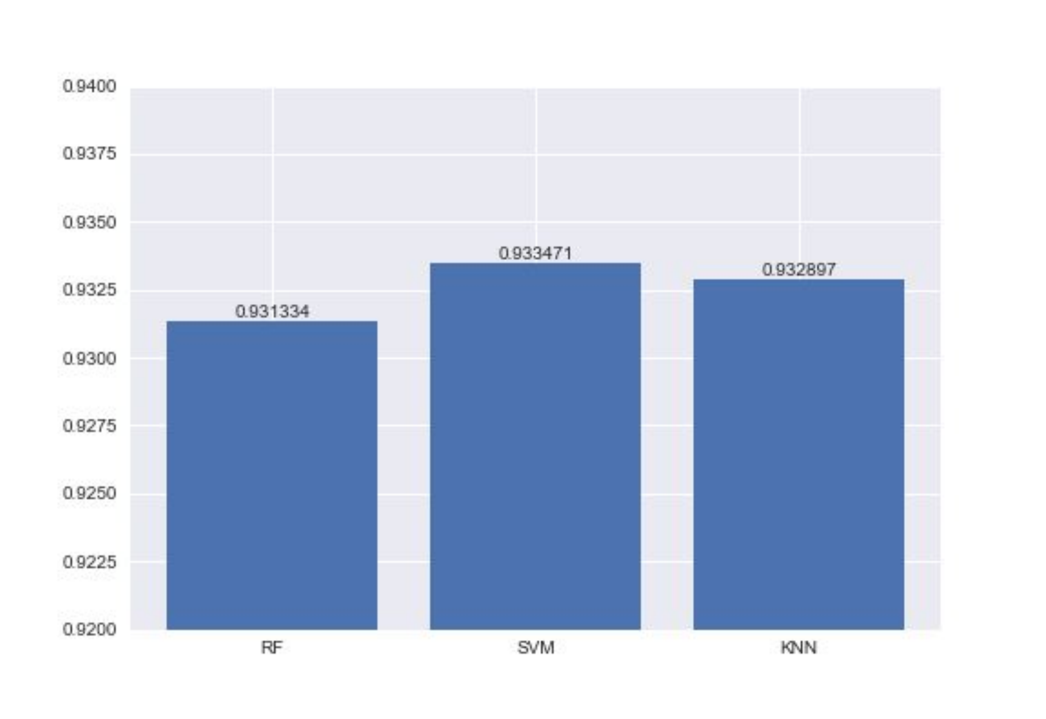
Graph represents f1_score by algorithm



- Best default algorithm is Random forest with SMOTE as well as SVM with smote

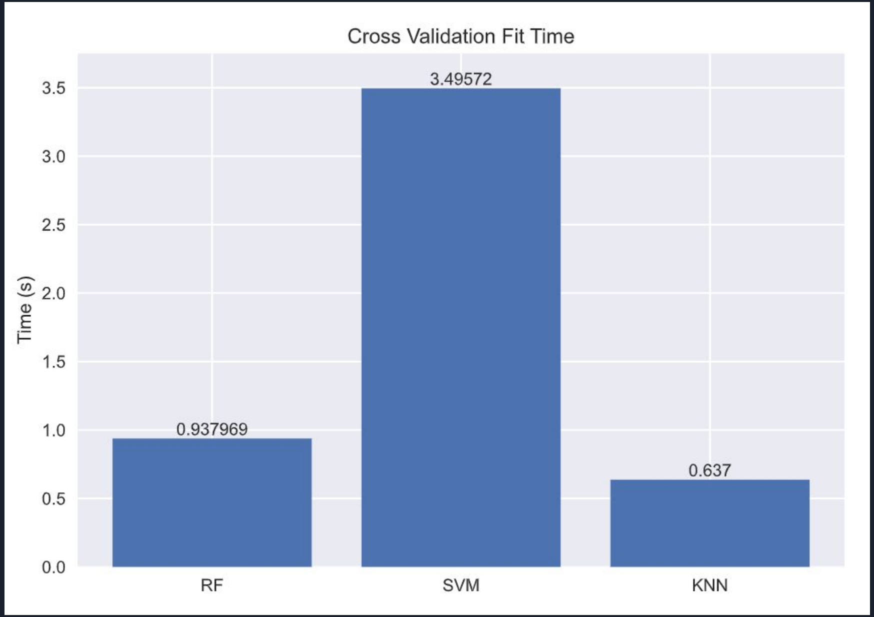
- KNN is not influenced by SMOTE

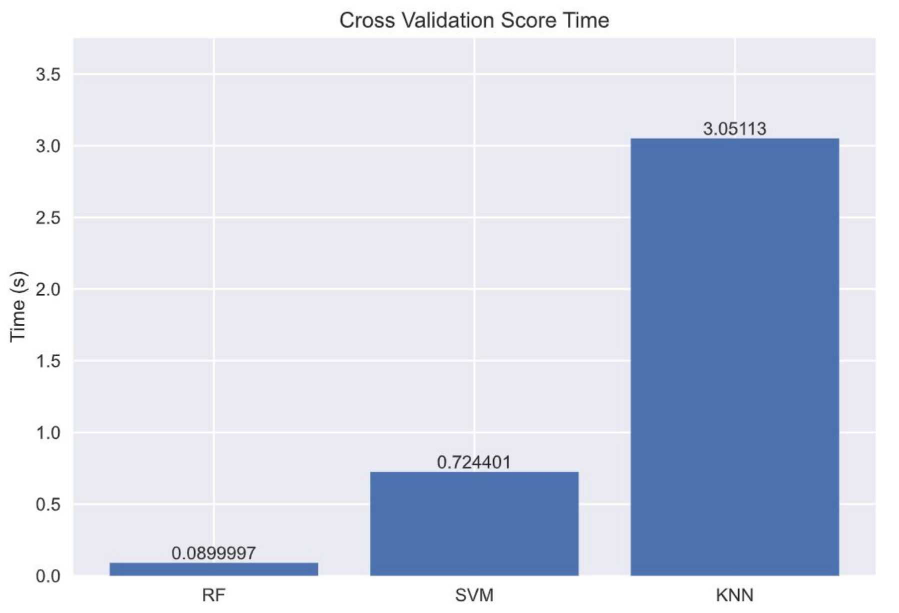
Results obtained by the Grid Search



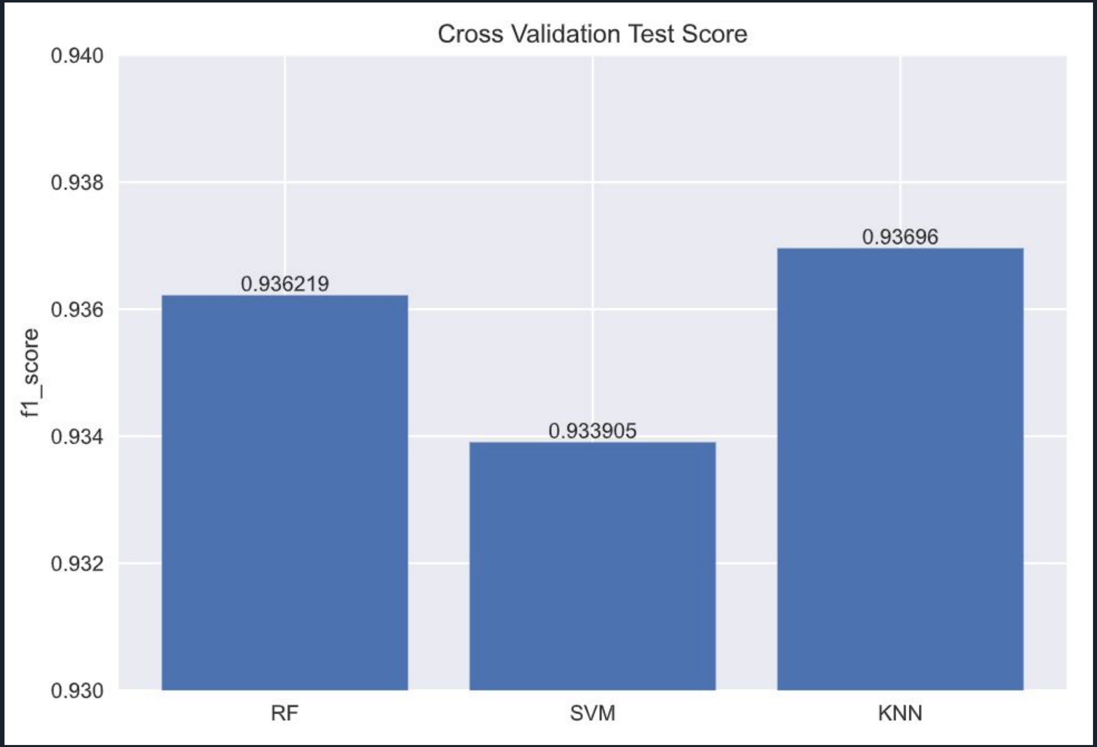
- SVM is the best scoring one
- Small difference between all of them

Results obtained by the Cross Validation





Results obtained by the Cross Validation


Cross Validation Test Score
f1_score
RF
SVM
KNN