U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

# Covid, Flu, Cold Symptoms ...

Artificial Intelligence
2021, May 24th
3MIEIC02, group 24

Inês Silva, up201806385@fe.up.pt
Mariana Truta, up201806543@fe.up.pt
Rita Peixoto, up201806257@fe.up.pt

# Problem definition

Often, flu, colds and allergies may be mistaken for COVID. To help set the difference between these problems, the data set used in this project was created with this intent.

The goal of this project is to study the way to obtain the most accurate diagnosis with a set of given symptoms.

The data set provided consists of a set of 20 attributes corresponding to the symptoms. Each entry is described using a boolean, meaning, having or not having the given symptom.

# Related work

**Documentation used as guidance:**

- ➔ https://scikit-learn.org/stable/

**Relevant websites:**

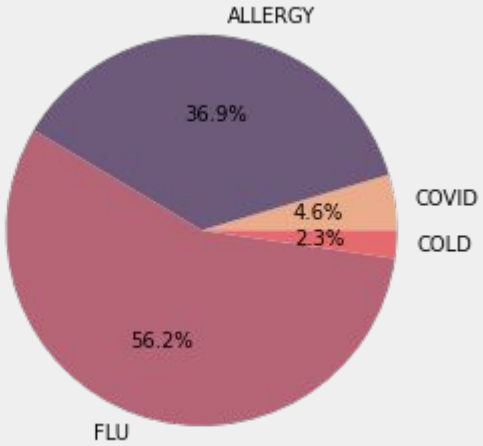
- ➔ Problem data: https://www.kaggle.com/walterconway/covid-flu-cold-symptoms
- ➔ Comparison of Machine Learning Classifiers: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4684714/
- ➔ https://www.dovepress.com/predicting-environmental-allergies-from-real-world-data-through-a-mobi-peer-reviewed-fulltext-article-JAA

# Imbalanced dataset

Our dataset is imbalanced which means that the class distribution is not equal or close to equal.

Initially, we verified that the algorithms performed poorly in the covid and cold cases when using the original dataset, since there were very few occurrences of these diseases.

Thus, to balance the training set, we oversampled the minority class and undersample the majority class.



# Data Preprocessing

In terms of preprocessing, there wasn't much we could find useful in our case because there were no missing values, no way we could classify any data as duplicate, no outliers and no attributes that were correlated enough to aggregate.

Therefore, in the preprocessing stage, we switched the data type of 'type' from object to a category and standardized our data scale so all the data was used in the same scale and contribute equally to the model fitting, as it is a good practice to these algorithms.

```
df['TYPE'] = df['TYPE'].astype('category')
```

# Data Processing

- ➔ Undersampling and oversampling data
- ➔ Train and test split data
- ➔ Scaler: we standardized our data scale so all the data was used in the same scale and contribute equally to the model fitting, as it is a good practice to these algorithms.

# Developed Algorithms

### Decision Tree Classifier

This classifier uses the Decision Tree algorithm, whose goal is to predict the value of a target variable by learning simple decision rules inferred from the data features.

### Support Vector Classification

This classifier uses the Support Vector Machines algorithm, whose learning problem is formulated as a convex optimization problem and is robust to noise.

### K-Nearest Neighbors

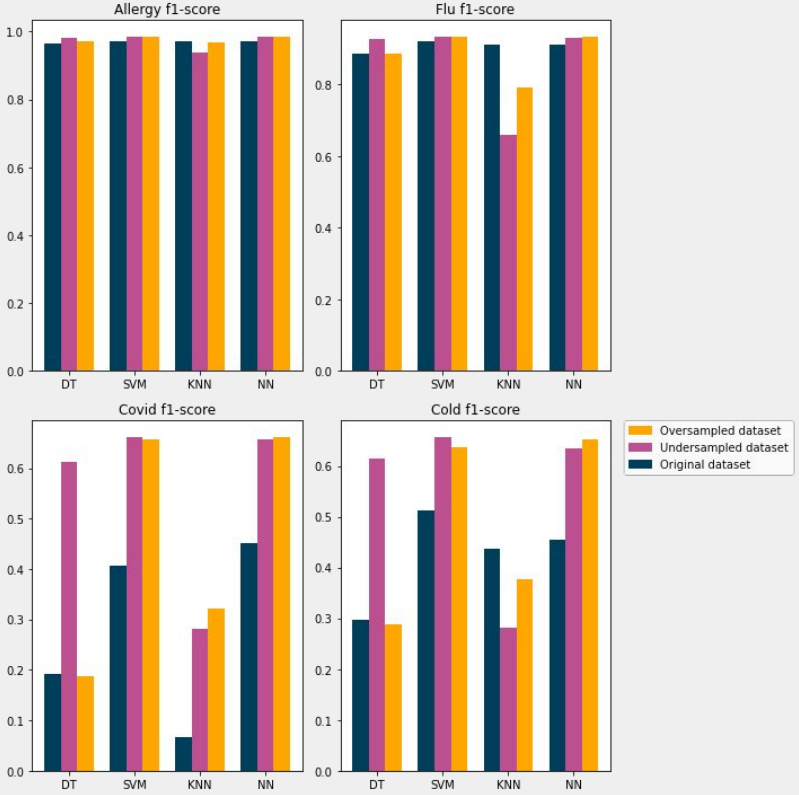
K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

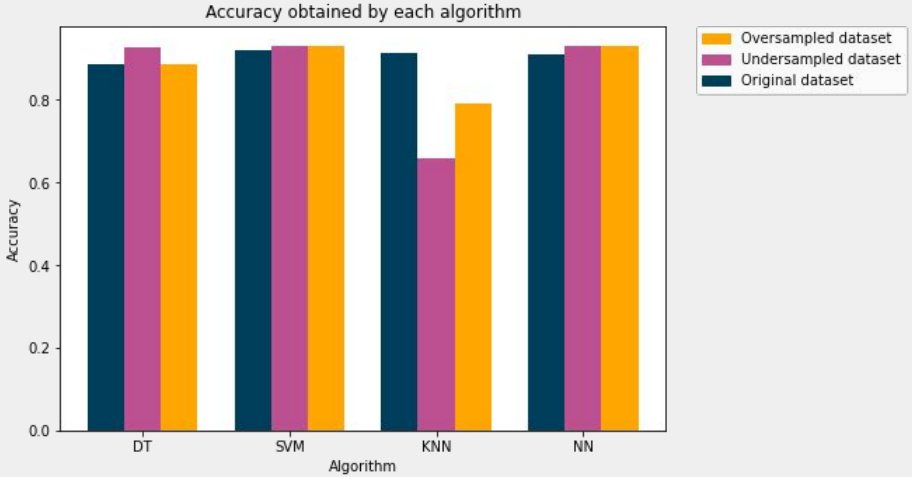
### Neural Networks

Neural networks learn by processing examples, each of which contains a known "input" and "result," forming probability-weighted associations between the two, which are stored within the data structure of the net itself.

**Note:** We applied these algorithms for the original dataset, an undersampled dataset and an oversampled dataset.

# Developed Algorithms: evaluation and comparison



# Conclusion

With this project we were able to have an introduction to Machine Learning, allowing us to understand how to implement the algorithms explored plus what to do to try to improve its results.

Also, we were able to understand the influence of imbalanced data in the algorithms when we started to take this into account, improvements were clear.

Using the obtained classification reports we also plotted the algorithms' performance for each disease type. We can see how important the unbalanced data sampling methods were to improve the score for the covid and cold cases, where the algorithms performed poorly when using the original dataset since there were very few occurrences of these diseases.

Unfortunately, we weren't able to run all the algorithms we planned because they did not execute in doable time.