**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning |
| **Assessment Title:** | CA1 Project |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Daniela Mariano Barreto |
| **Student Number:** | 2023278 |
| **Assessment Due Date:** | 26/11/2023 |
| **Date of Submission:** | 26/11/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Table of Contents

[**Introduction** 3](#_Toc151828496)

[*Word count* 3](#_Toc151828497)

[**Data Description** 3](#_Toc151828498)

[**Data Cleaning and Preprocessing** 4](#_Toc151828499)

[**Machine Learning Models** 4](#_Toc151828500)

[***Outcomes*** 5](#_Toc151828501)

[*Random Forests* 5](#_Toc151828502)

[*Random Forests – SMOTE* 8](#_Toc151828503)

[*Principal Component Analysis* 9](#_Toc151828504)

[*Random Forests - PCA* 11](#_Toc151828505)

[*Artificial Neural Networks* 13](#_Toc151828506)

[***Assessment*** 15](#_Toc151828507)

[**Conclusion** 16](#_Toc151828508)

[**Reflective journal** 16](#_Toc151828509)

[**References** 16](#_Toc151828510)

# **Introduction**

Since 2020, more than 7 million people have lost their lives around the world to a lethal virus known as SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), caused by a disease named COVID-19 (coronavirus disease 2019) (World Health Organization, 2023). This pandemic has had a substantial impact on global public health, the economy, and people's lives, which may have possible severe consequences for this disease that are still unknown.

Many scientists worldwide have mobilized to develop vaccines, which are currently considered the fastest vaccines made to date (Khuroo et al., 2020, p.1). Although most people are presently vaccinated, as with any other vaccine, the COVID-19 vaccines have been reported to have several side effects, from mild to severe, like fever and fatigue to cardiac problems and, in some cases, resulting in death.

Machine learning (ML) has been applied in healthcare sectors to simulate and predict outcomes, evaluate medicines, and diagnose and prognose many diseases (Bansal et al., 2021). Based upon the reports from VAERS (Vaccine Adverse Event Reporting System) (Garg, 2023, p.1), in this project, I aim to apply two ML models to evaluate which model can achieve the most excellent accuracy, precision and recall in the prediction of people's death after COVID-19 vaccination and identify which feature can contribute more to the death risk after the vaccination, such as the presence of allergies or illnesses pre-existed.

## *Word count*

Introduction: 227

Data description: 77

Data preparation and preprocessing: 236

Machine learning models: 240

Total: 920

Outcomes: 140

Assessment: 266

Conclusion:

Reflective journal:

Total:

# **Data Description**

The dataset used is from the Kaggle repository (Garg, 2023, p.1). This dataset contains the adverse events reported by individuals after the COVID-19 vaccine from January/2021 to March/2021 in American States. It has more than 35 features and more than 34 thousand records. I focused on some features that I believe may help me to identify which feature can contribute more to the death risk after the vaccination for example the presence of allergies and illnesses pre-existed.

# **Data Cleaning and Preprocessing**

Previously, I filtered the data by ‘STATE’ to focus on California State, a cosmopolitan place with people from different ethnicities, which may reduce the probability of bias. Then, I checked for duplicates, missing and null values, and NaN, used to fill blank spaces, as predetermined by VAERS to represent non-occurrence. Thus, the NaN values were replaced with zeros. I dropped some features, leaving only those that might be essential to answer my questions, such as the presence of pre-existing illnesses and allergies.

I replaced all sentences reported in 'CUR\_ILL', 'HISTORY', and 'ALLERGIES' with blank space, meaning the absence of the occurrence, 'U' when it was not informed, and 'Y' in the case in which the patient related any occurrence (from mild to several), thus, in this study, I will not make distinguish of the degree of illnesses or allergies.

Afterward, I replaced the letters with numbers because machine learning models work efficiently with numeric representations, and it is possible to make standardization and normalization, which is essential for some models like Neural Networks (Müller and Guido, 2017 p.114).

I also applied scaling in the data and used the Synthetic Minority Over-sampling Technique (SMOTE) to address the class imbalance from the minority, which is 15 times smaller than the majority class. Since ML learns the decision boundary for the majority class more efficiently than the minority class, I used this method to deal with it.

Data dictionary:

A screenshot of a computer

Description automatically generated

# **Machine Learning Models**

In this project, I applied Random Forests (RF) and Artificial Neural Networks (ANN), supervised learning models used for classification to predict whether a person died or not due to some circumstances (pre-existing illness).

RF is a robust algorithm that combines multiple machine learning models (Müller and Guido, 2017, p.83), and that is why I chose this model because the predictions are based on the median of many random trees, which can be an advantage for imbalanced data and can avoid overfitting. I also obtained the most important feature, that can answer my question about which features can significantly contribute to the death risk after vaccination.

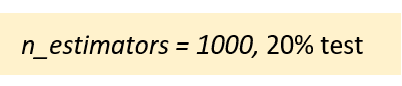
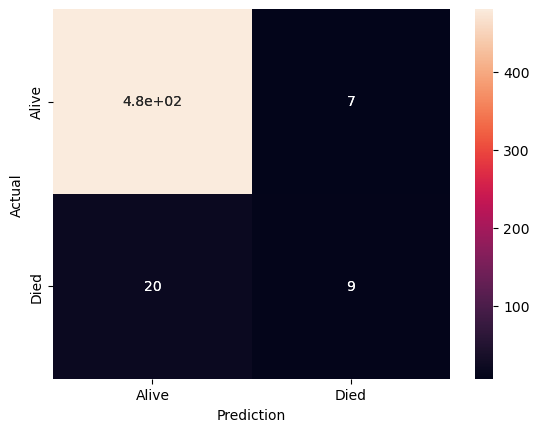
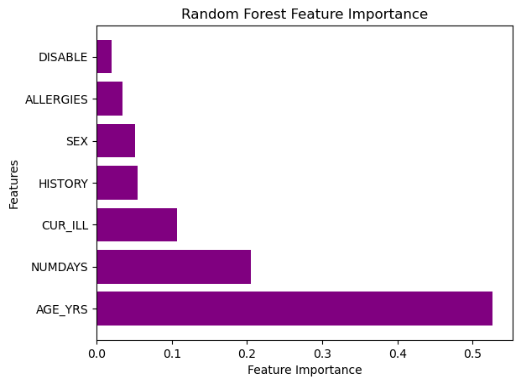
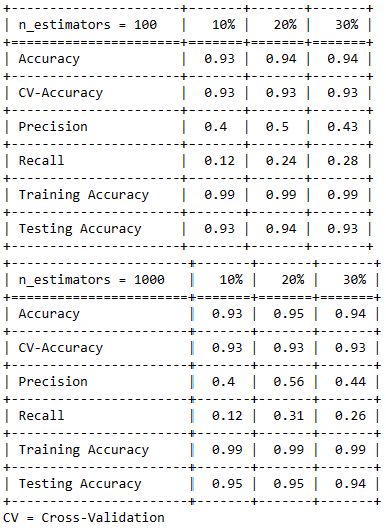
As ANN is powerful for classification tasks and works best with features with the same meaning (Müller and Guido, 2017, p.117-118). This model suits my goals as I have homogeneous features and binary target, ideal for the classification method. I also scaled the data as required.

I used the Principal Component Analysis (PCA), a dimension reduction method that extracts principal features essential to explain the data (Bishop, 2006, p.561), removing redundant features and avoiding overfitting. I used it to improve the model performance and noise reduction, focusing on the most significant pattern.

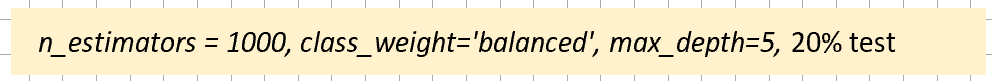
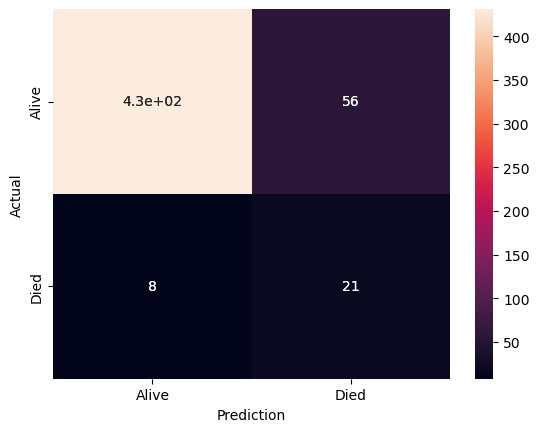
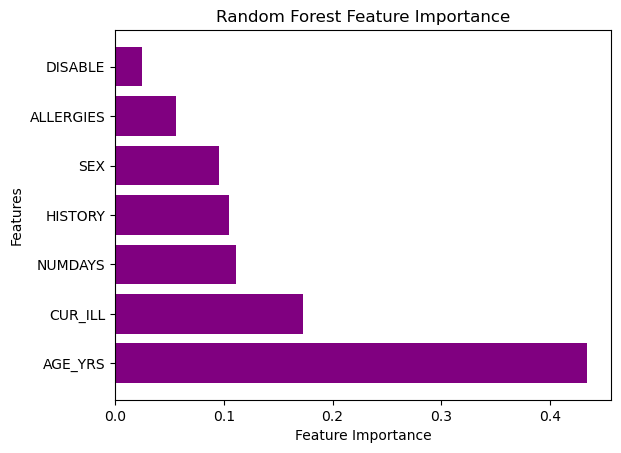
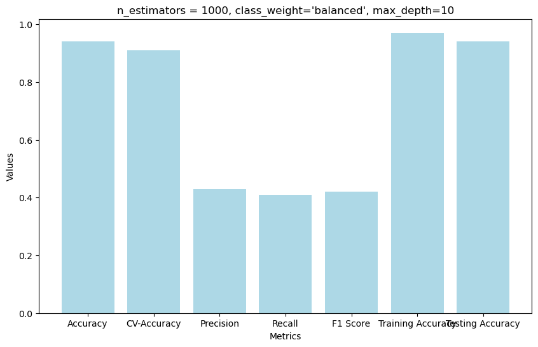
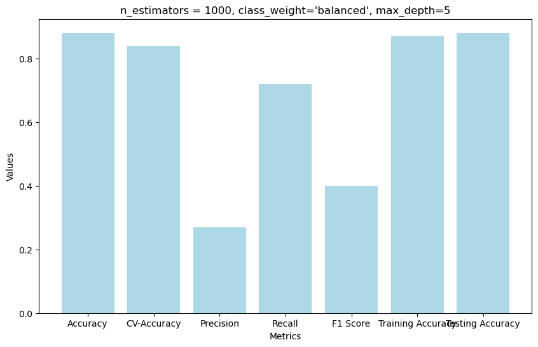
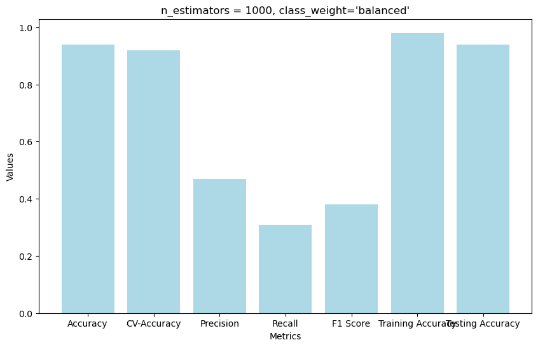
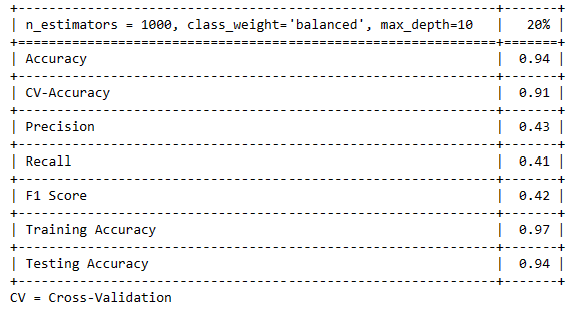
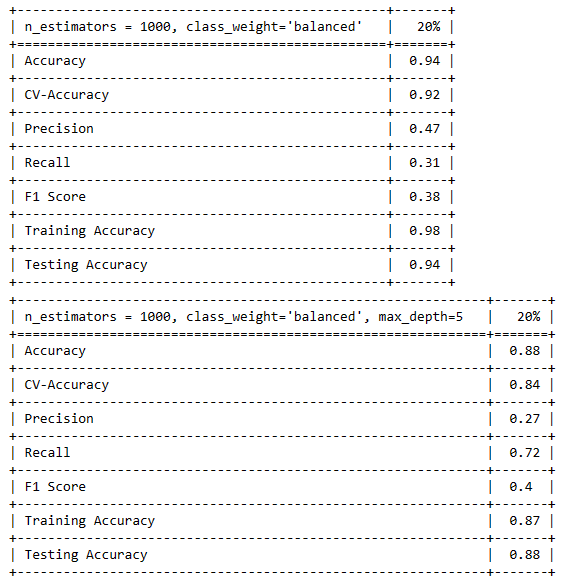
## ***Outcomes***

### *Random Forests*

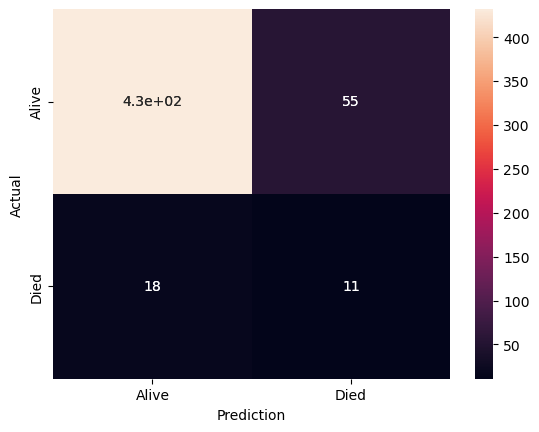
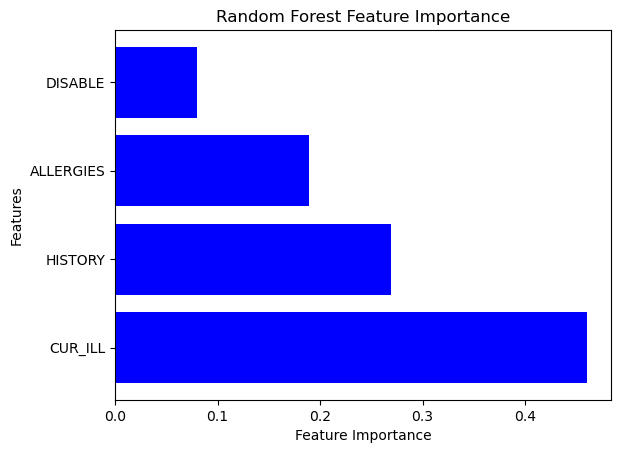
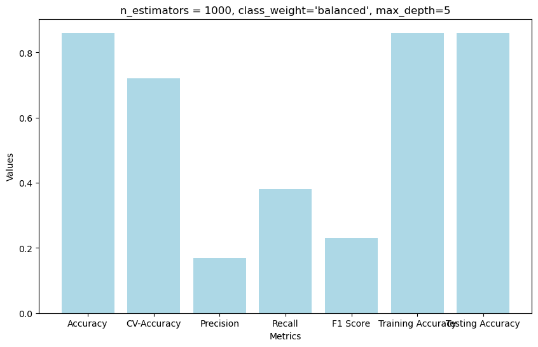
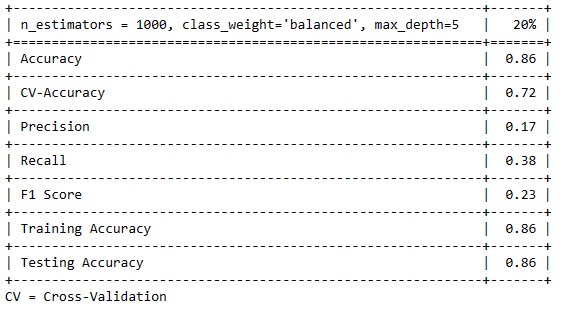
Table and charts below represent the results for RF model in which I set the number of trees (n\_estimators) and the split percentage.



Results of hyperparameters set are below with focus on the 20% test which presented greater performance anteriorly.

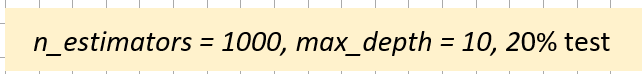
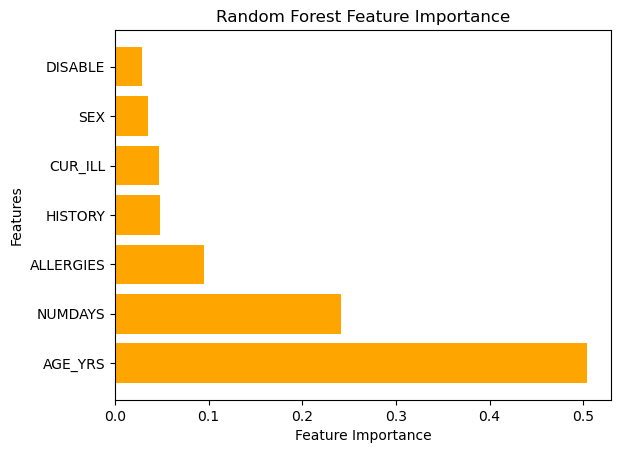
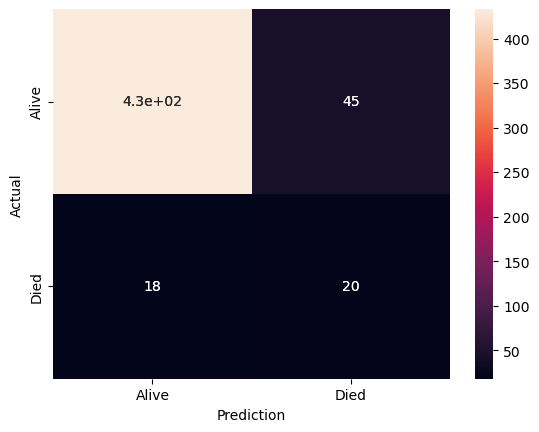
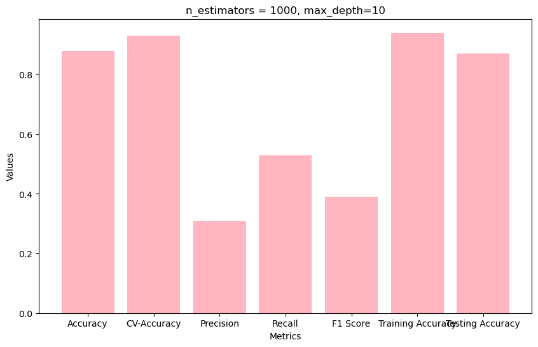
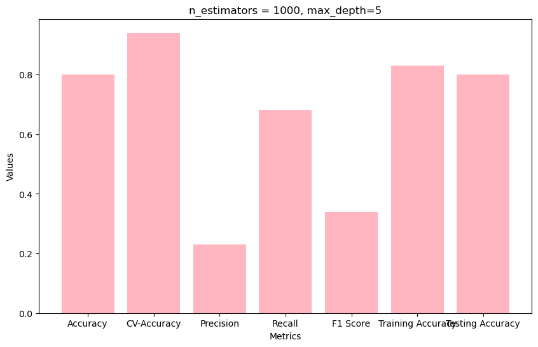
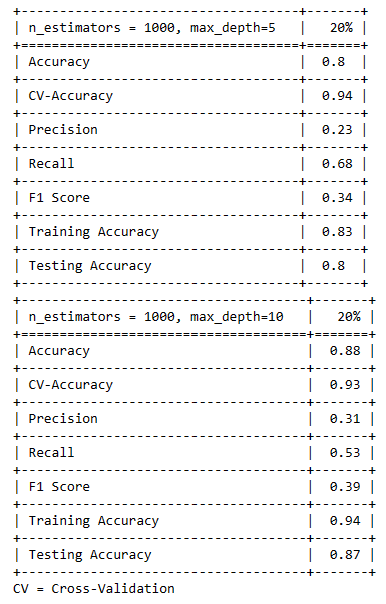


This is the result for four features in focus to identify which can contribute more to the death risk after the vaccination. I used the best hyperparameters found earlier.



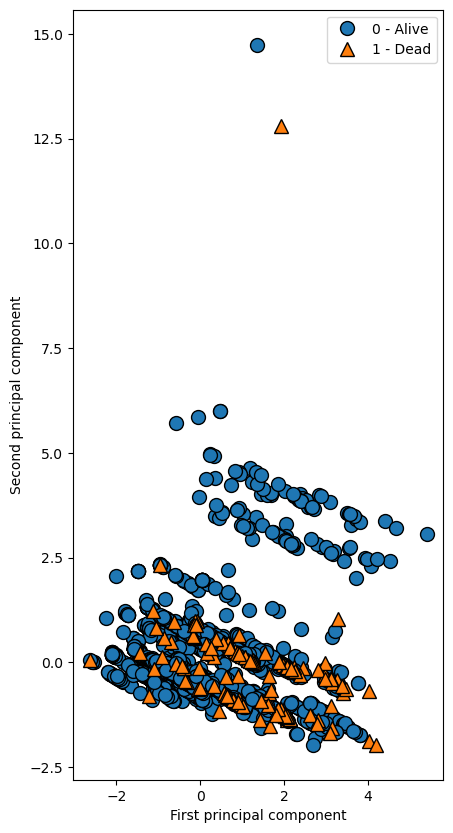
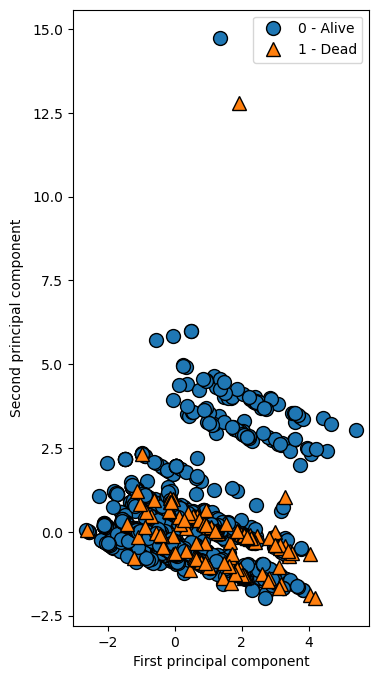
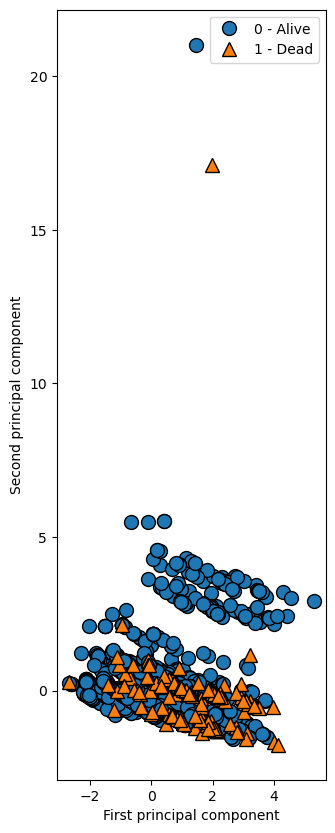
### *Random Forests – SMOTE*

Results for SMOTE applied on the training set then RF performance with hyperparameters adjustments.



### *Principal Component Analysis*

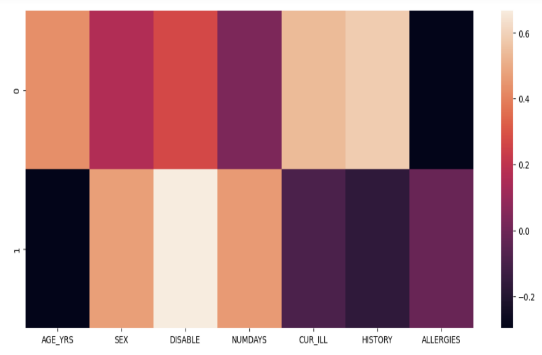
Charts below show the principals components for each sample split and the heatmap the feature contribution for the target class.



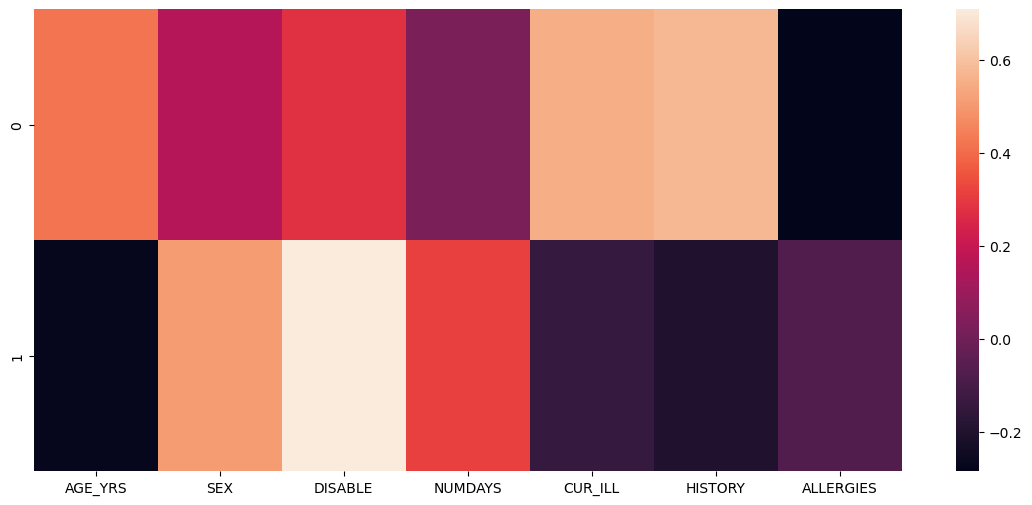
20% test

30% test

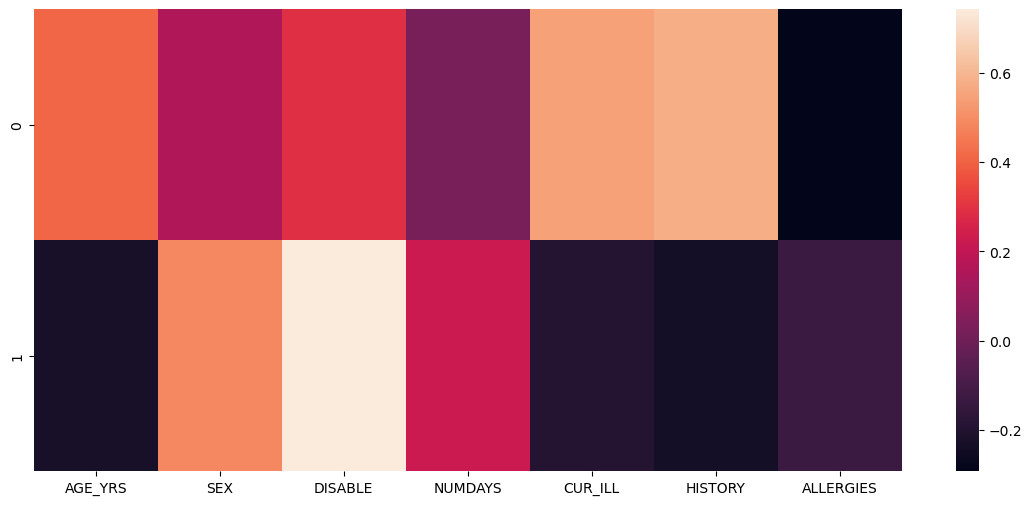
10% test



10% test



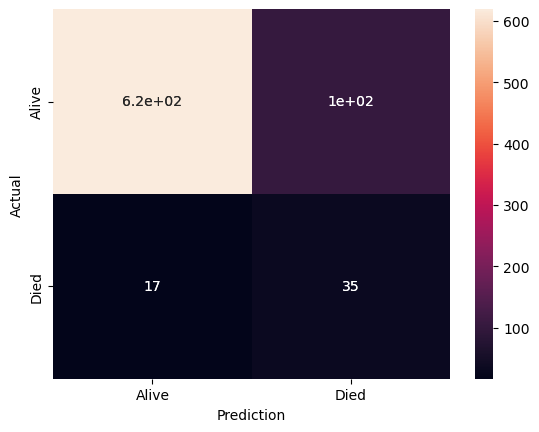
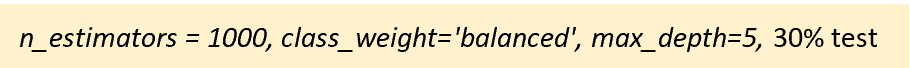
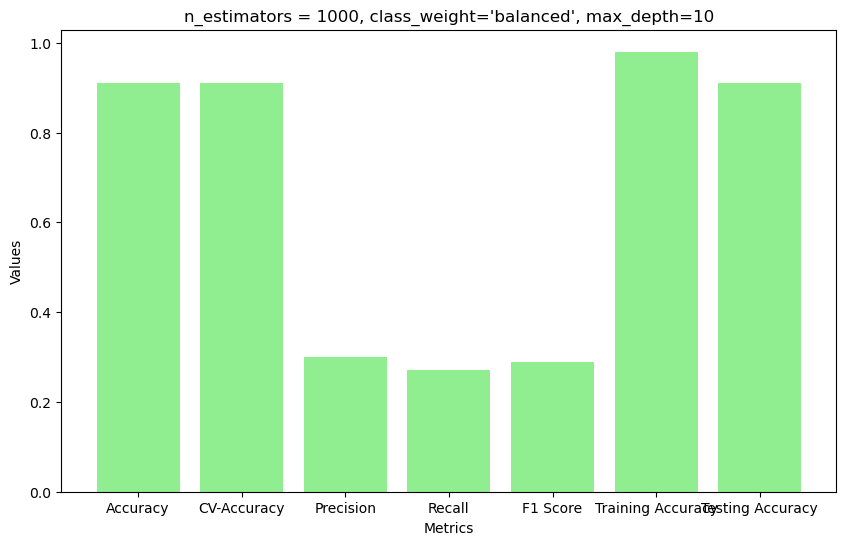
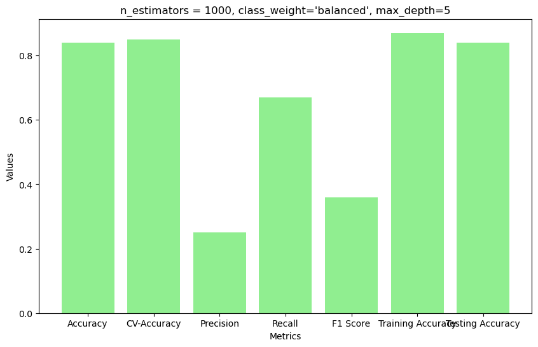
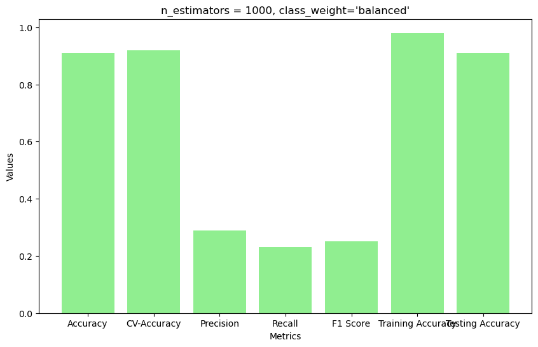
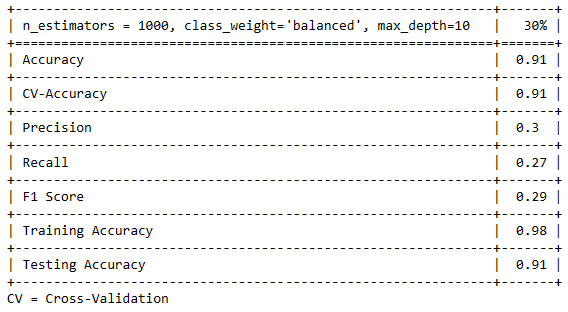
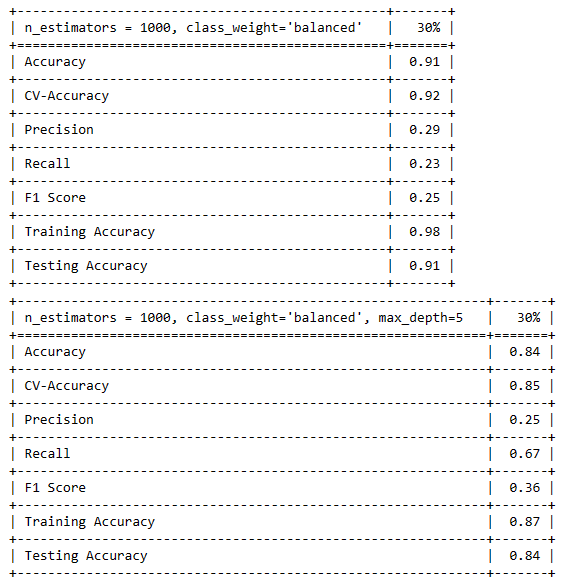
20% test



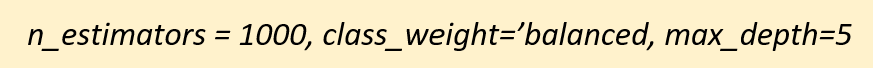
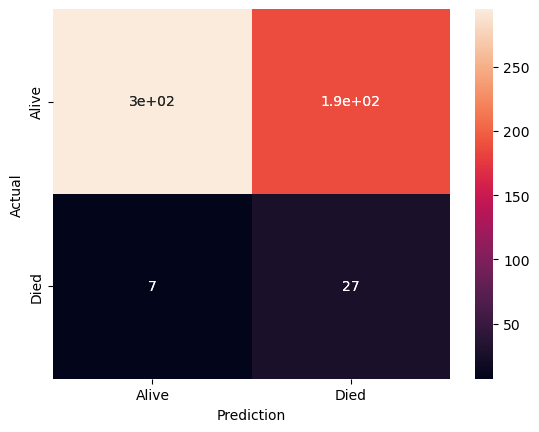
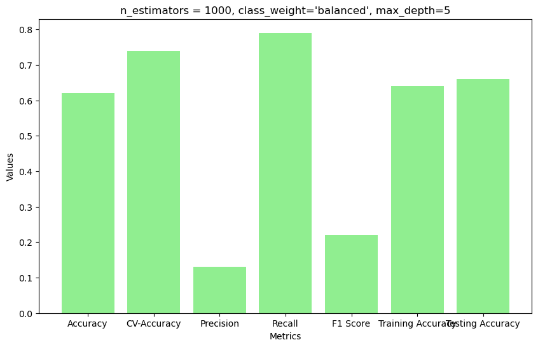
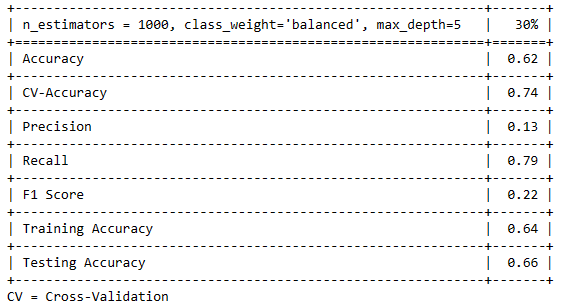
30% test

### *Random Forests - PCA*

Results of hyperparameters in RF-PCA performance.



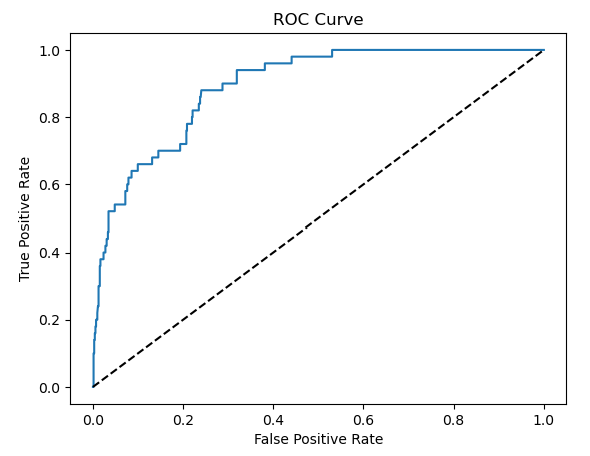
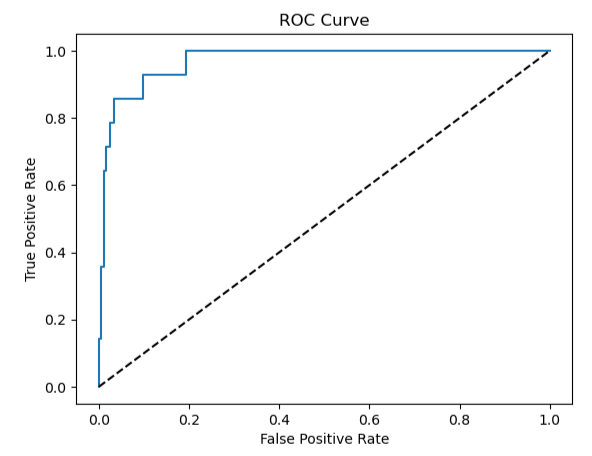
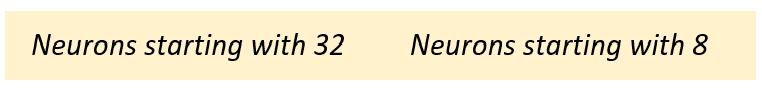
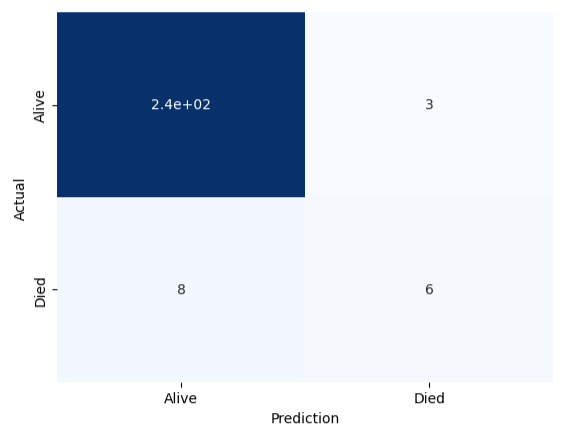
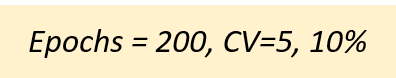
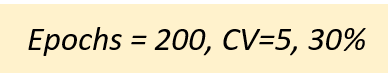
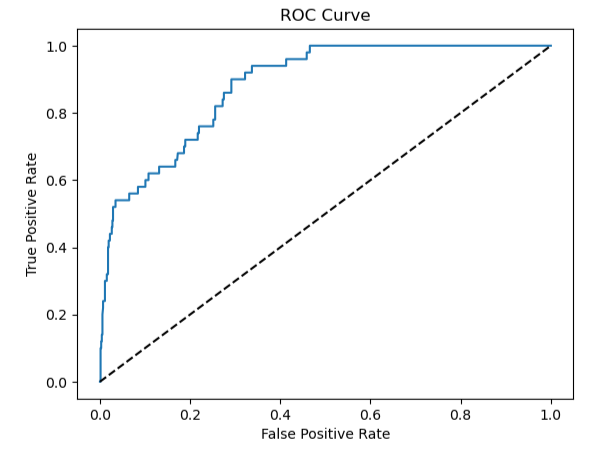
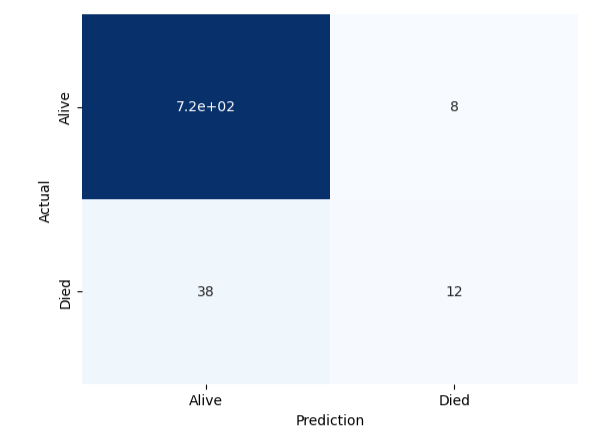
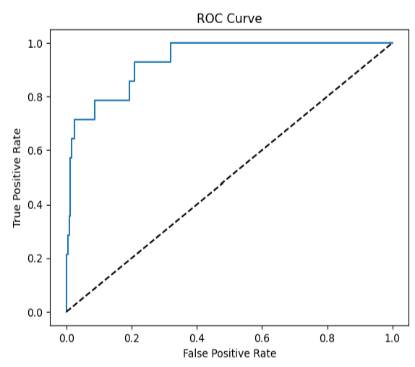
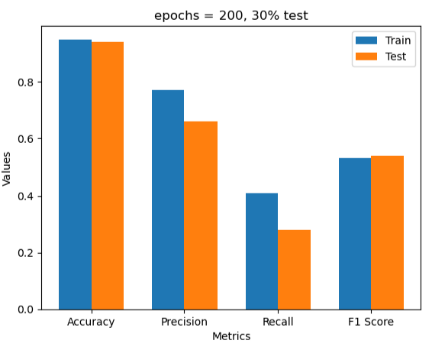
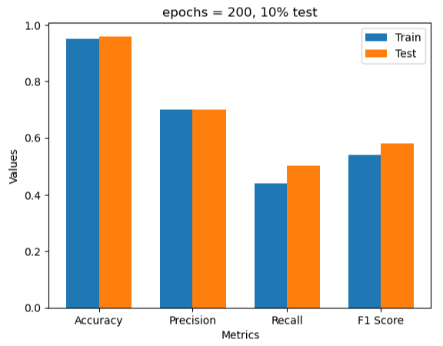
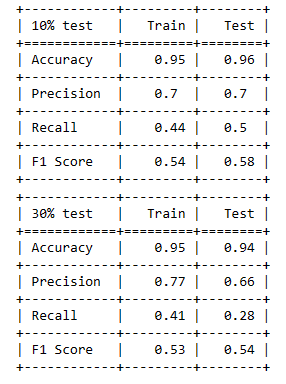
RF-PCA results with focus on four features.



### 

### *Artificial Neural Networks*

Charts and table show results after adjusting hyperparameters for model’s performance with scaled data and cross-validation. ROC curves are presented for different neurons number.



## ***Assessment***

Comparing RF-SMOTE and initial data (estimator=1000, class\_weight, max\_dept=5), SMOTE had improved the cross-validation accuracy but not the precision that were similar in both. Recall was ~20% greater in the initial data than in SMOTE, thus only the hyperparameters set contributed to improve positive predictions. Although RF is robust to overfitting, seems that it happened in SMOTE results as the testing accuracy (87%) was small than the training (94%), meaning that the models learned noises and peculiarities of the data.

Confusion-matrix for both do not showed improvement on the target but had an increase in false negative. The ‘AGE\_YRS’ feature was the most important, however when focusing on the illnesses features the first is ‘CUR\_ILL’, followed by ‘HISTORY’ and ‘ALLERGIES’. Therefore, the presence of any illness or pre-existed illness at the vaccination day may contribute to the death of vaccinated patient.

PCA results showed that ‘DISABLE’ was the feature that most contributed to the death. The hyperparameters not resulted in substantial improvements on model’s performance. Beside that the recall was 5% higher than the initial data and precision continued low, like the anterior procedures. Although the accuracy was not high the RF-PCA was not overfitting and the precision continued low.

Only with the ANN model was possible improve the precision metric, which was 1.5 higher than the initial data meaning reduction in false negative that contribute to the ‘died’ class. However, the recall presented 0.6 times reduction compared with RF-PCA. In general, the model performed well in the train and test with high accuracy but seems that is in eminence of overfitting as the results are similar. The reduction in the number of initial neurons improved the true positive but was desirable that the true negative had increased, probability it happened because of the imbalanced class.

3. Interpret and explain the results obtained, discuss overfitting / underfitting / generalisation, provide a rationale for the chosen model and use visualisations to support your findings. Comments in Python code, conclusions of the project should be specified at the end of the report. Harvard Style must be used for citations and references.

# **Conclusion**

I aim to apply two ML models to evaluate which model can achieve the most excellent accuracy, precision and recall in the prediction of people's death after COVID-19 vaccination and identify which feature can contribute more to the death risk after the vaccination, such as the presence of allergies or illnesses pre-existed.

# **Reflective journal**

A close-up of a text

Description automatically generated

# **References**

Müller, A. C. and Guido, S. (2017). Introduction to machine learning with Python: a guide for data scientists. 1st ed. United States of America. O’reilly Media.

Scikit-learn.org. (2012). 3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.22 documentation. [online] Available at: https://scikit-learn.org/stable/modules/grid\_search.html [Accessed 19 Nov. 2023].

World Health Organization (2023). *WHO COVID-19 dashboard*. [online] World Health Organization. Available at: <https://covid19.who.int/>.

Khuroo, M.S., Khuroo, M., Khuroo, M.S., Sofi, A.A. and Khuroo, N.S. (2020). COVID-19 Vaccines: A Race Against Time in the Middle of Death and Devastation! *Journal of Clinical and Experimental Hepatology*. doi:https://doi.org/10.1016/j.jceh.2020.06.003.

Bansal, H., Balamurugan Balusamy, T. Poongodi and Firoz Khan KP (2021). *Machine Learning and Analytics in Healthcare Systems*. CRC Press.

Garg, A. 2023. *COVID-19 World Vaccine Adverse Reactions*. [online] Available at: <https://www.kaggle.com/datasets/ayushggarg/covid19-vaccine-adverse-reactions>. [Accessed 15 Nov. 2023].

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer.