**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](#_Toc151734566)

[*Word count* 3](#_Toc151734567)

[**Data Description** 3](#_Toc151734568)

[**Data Cleaning and Preprocessing** 3](#_Toc151734569)

[**Machine Learning Models** 4](#_Toc151734570)

[***Outcomes*** 5](#_Toc151734571)

[*Random Forests* 5](#_Toc151734572)

[*Random Forests - SMOTE* 8](#_Toc151734573)

[*Principal Component Analysis - PCA* 10](#_Toc151734574)

[*Random Forests - PCA* 12](#_Toc151734575)

[*Artificial Neural Networks* 16](#_Toc151734576)

[***Assessment*** 18](#_Toc151734577)

[**Conclusion** 18](#_Toc151734578)

[**Reflective journal** 19](#_Toc151734579)

[**References** 19](#_Toc151734580)

# **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: 90

Data preparation and preprocessing: 236

Machine learning models: 188

Total: 741

Outcomes:

Assessment:

Conclusion:

Reflective journal:

Total:

# **Data Description**

The dataset used is from the Kaggle repository (Garg, 2023, p.1); the link and the licence link are presented below. 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.

Dataset: <https://www.kaggle.com/datasets/ayushggarg/covid19-vaccine-adverse-reactions>

Licence: <https://creativecommons.org/publicdomain/zero/1.0/>

# **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), I applied this model; however, I had to scale the data and handle the imbalanced class.

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.

c) Visualise your comparison of ML modelling outcomes. You may use a statistical approach to argue that one feature is more important than other features (for example, using PCA).

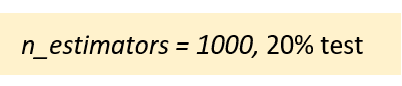
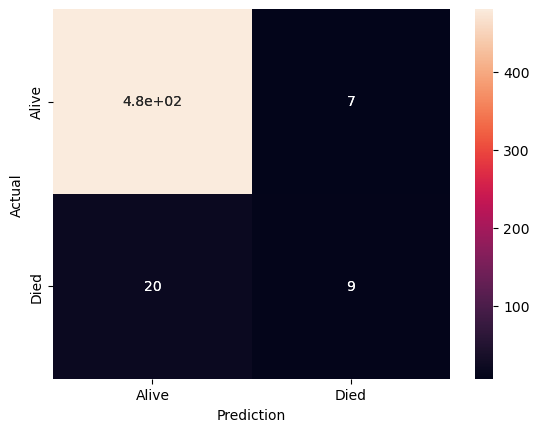
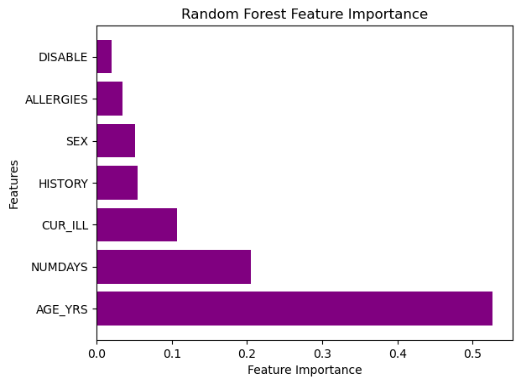
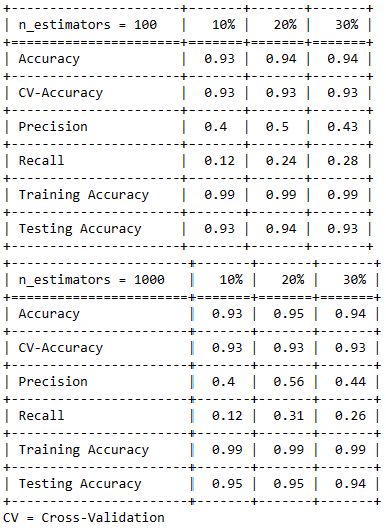
d) Cross-validation methods should be used to justify the authenticity of your ML results.

## ***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. Although they present high accuracy, inclusive for the Cross-Validation (CV) we can see that the imbalance class might be interfering in the model performance as we can see in confusion matrix (CM), precision and recall values.

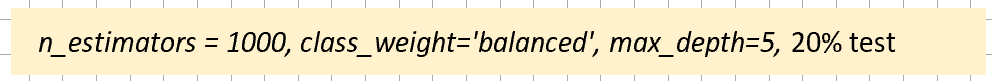
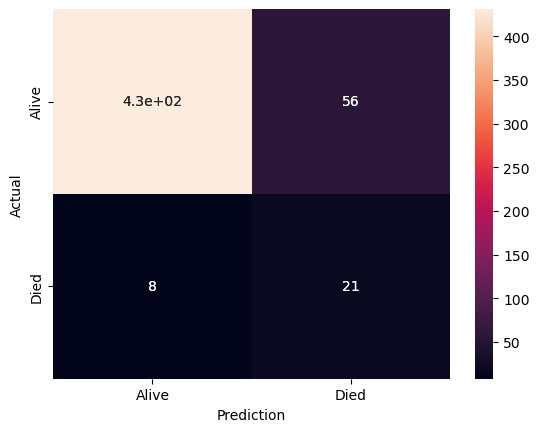
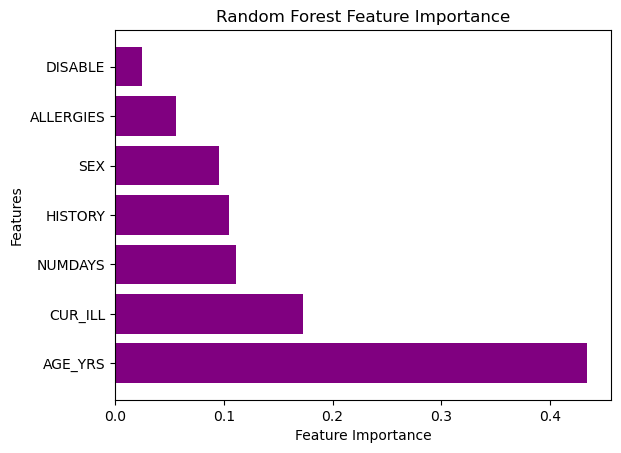
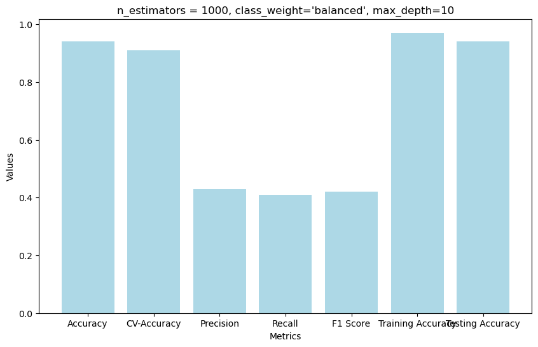
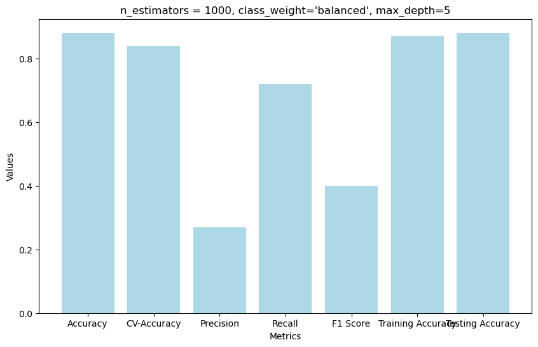
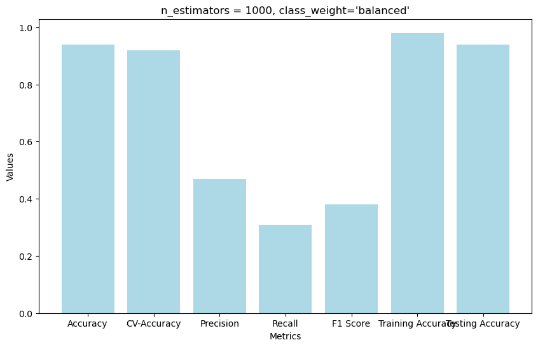
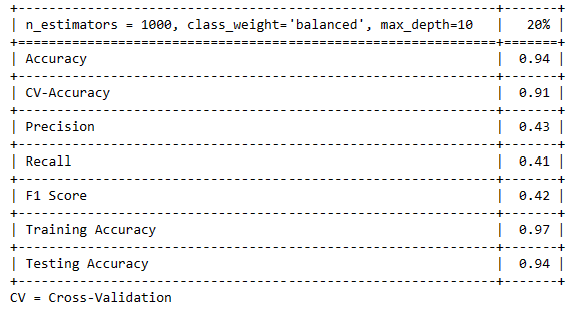
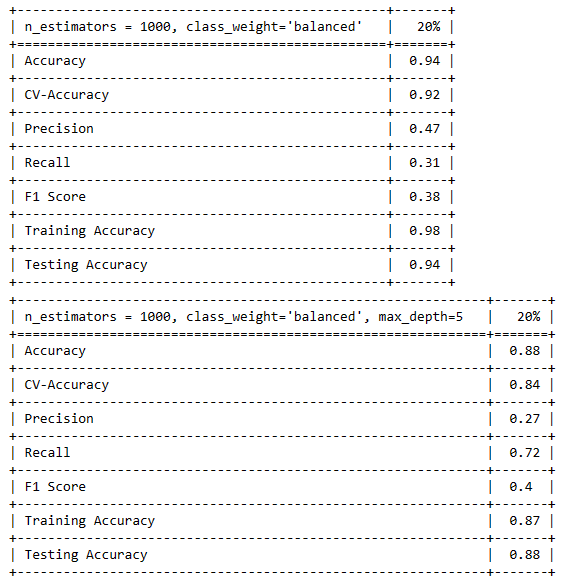
Feature importance shows that age and the number of days from the vaccination date to the onset date were feature which contribute the most to the model's predictions.



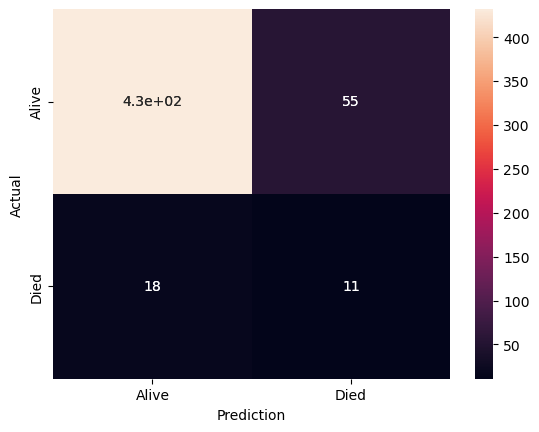
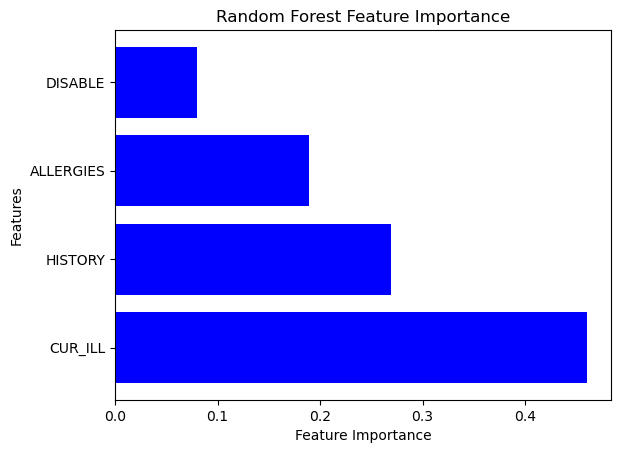
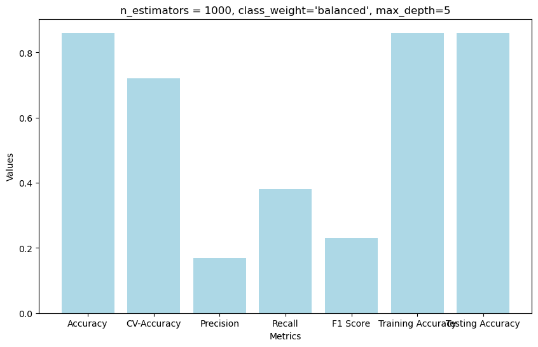
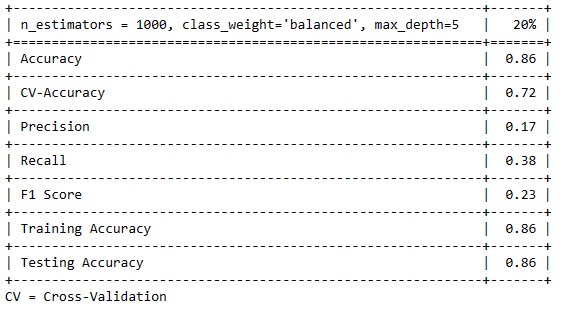
Following, are the adjustment of hyperparameters with focus on the n\_estimators = 1000 and 20% test split. The table and charts below show how the insertion of class weight balance and the number of nodes expansions (5) contributed to improve the recall, which means that the model could correctly predicted 72% of the positive instances. However, a higher precision value was also desirable because it measures the accuracy of positive predictions.

Now the feature importance chart shows that ages and occurrence of any illness at the vaccination time contributed to the models’ prediction.

Adjusting hyperparameters:



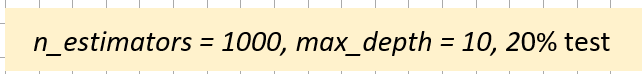
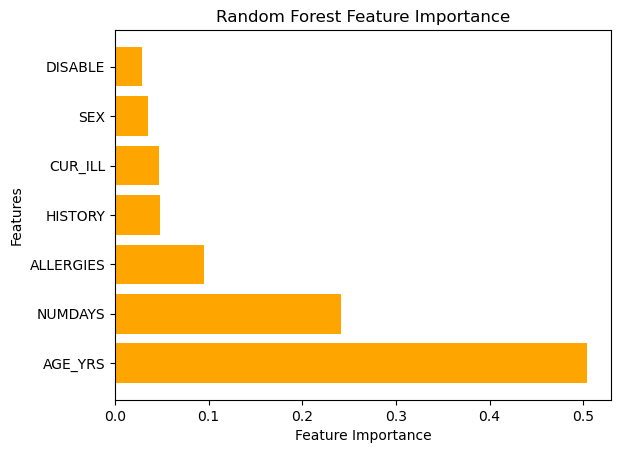
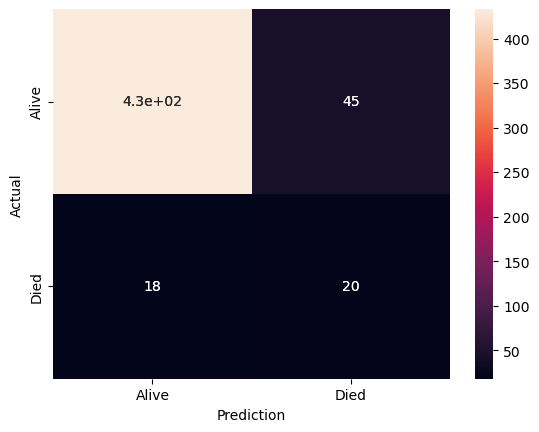
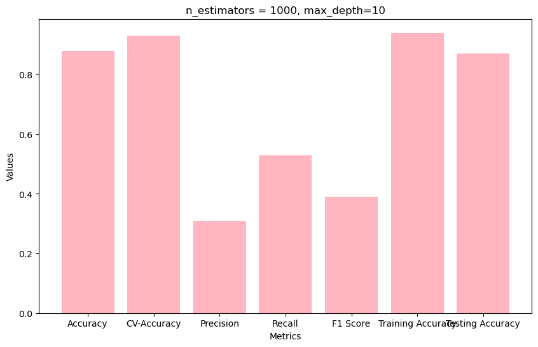
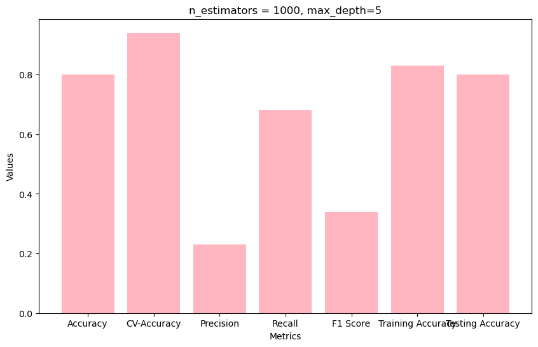
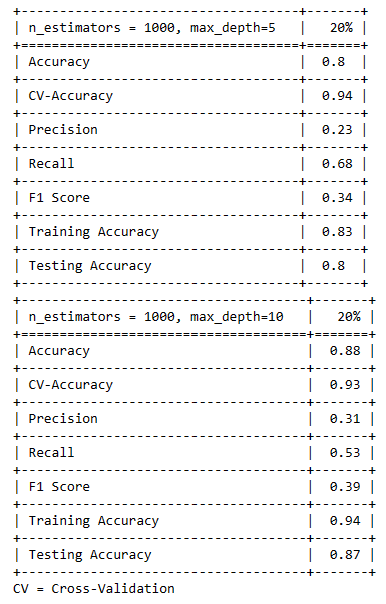
Here, I focused on features (DISABLE, CUR\_ILL, HISTORY, ALLERGIES) to identify which can contribute more to the death risk after the vaccination. I used the best hyperparameters found earlier to obtain these charts and table. However, any improvement was obtained, but I identified that the most important feature in this case was the presence of any illness at the vaccination time.



### *Random Forests – SMOTE*

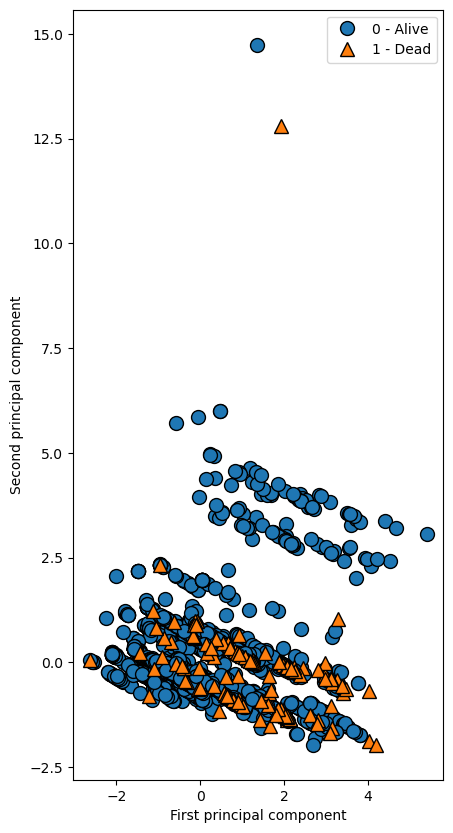
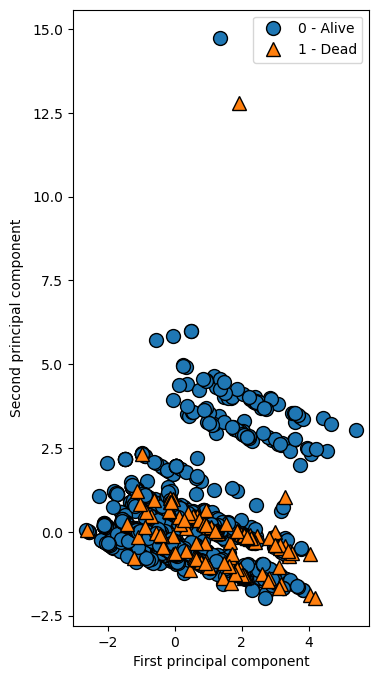
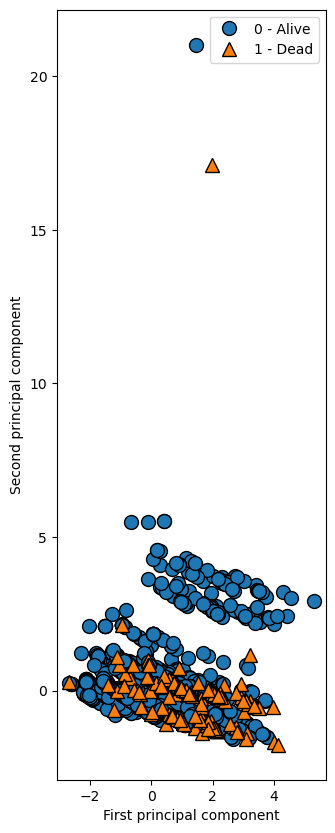
I applied the SMOTE only on the training set to keep the test set representative of the real sample. However, even after adjusting hyperparameters I continue with low precision and recall, meaning that oversampling did not improve the model’s performance for the positive class.

Adjusting hyperparameters:



### *Principal Component Analysis*

Considering the charts below we can see that the sample slipping did not result in considerable differences, as well as the sum of the first and second principal components that indicated that ~38% of the original data is retained in the principal components.

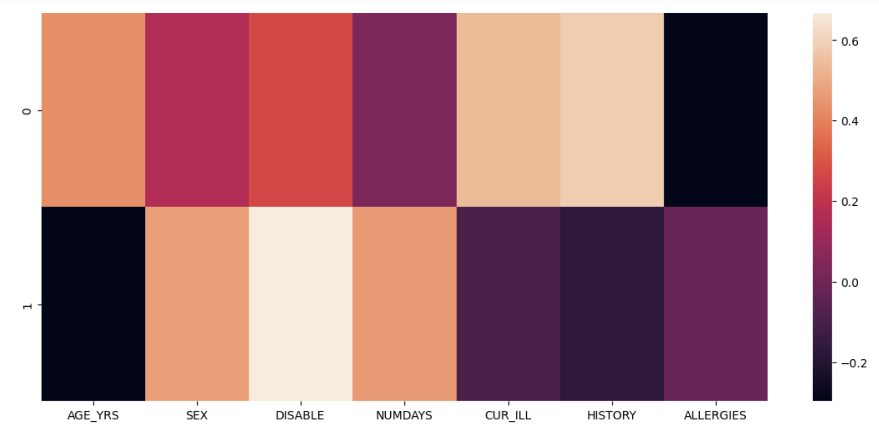


20% test

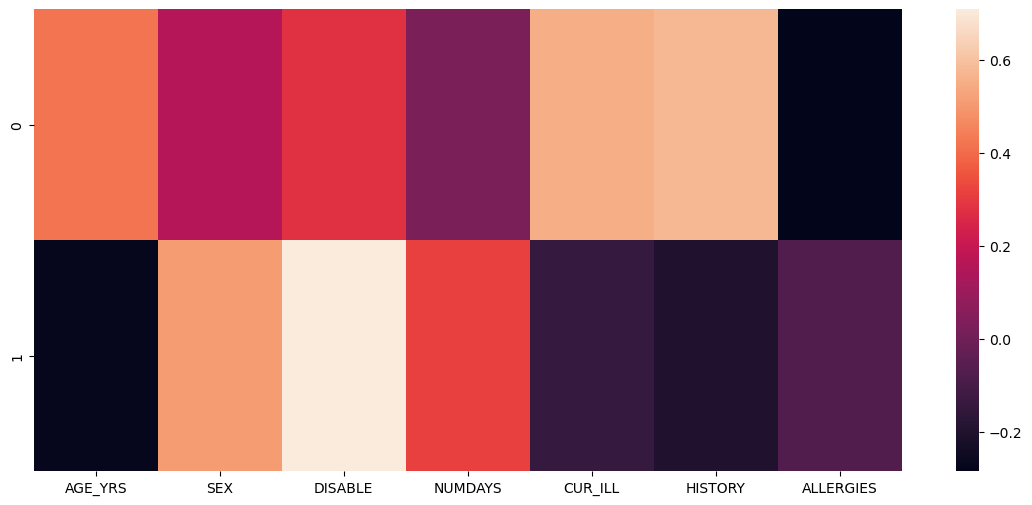
30% test

10% test

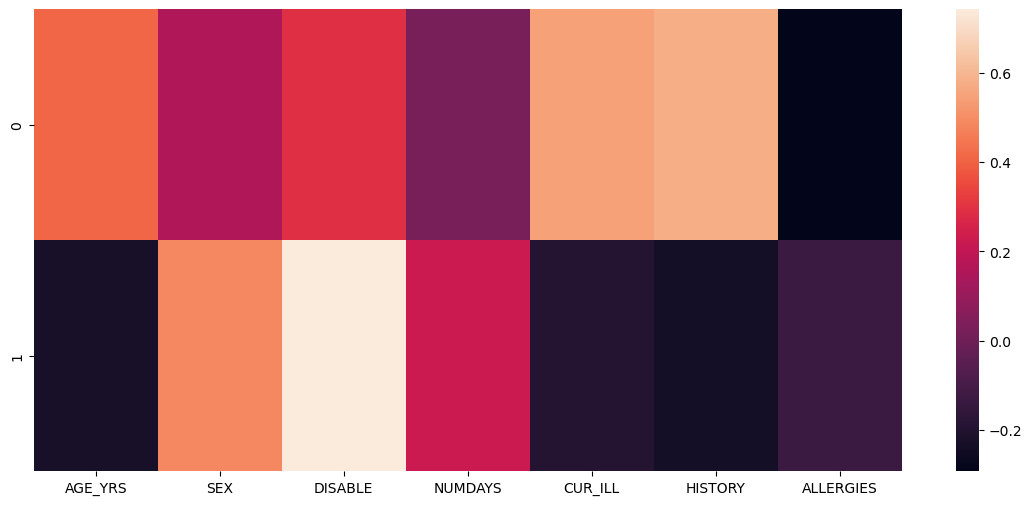
Below is presented the heatmap with the features for each class and we can see that ‘DISABLE’ feature is the feature that most contributed to the death of patients. Regarding to alive class, seamed that the absent of record for ‘HISTORY’ and ‘CUR\_ILL’ features most contributed to this target.



10% test



20% test



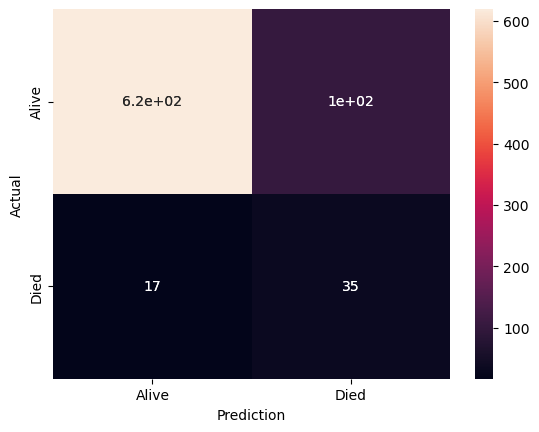
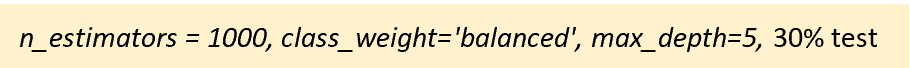
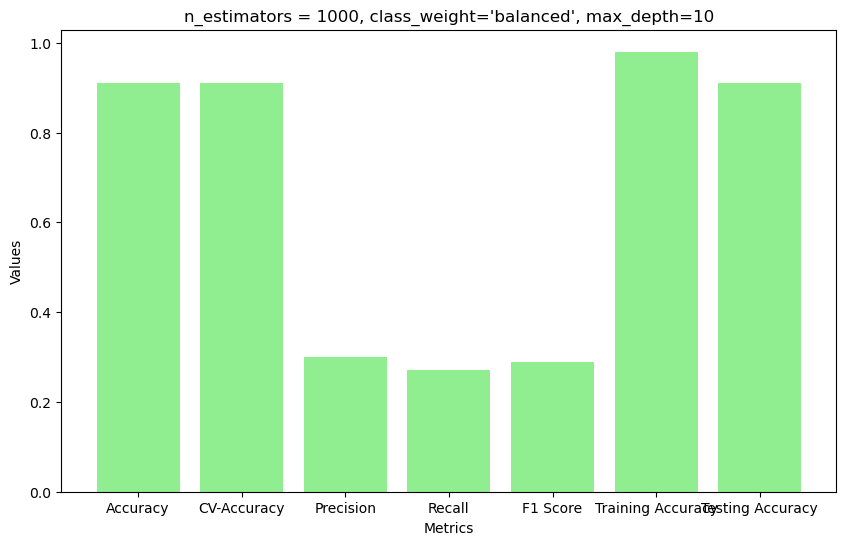
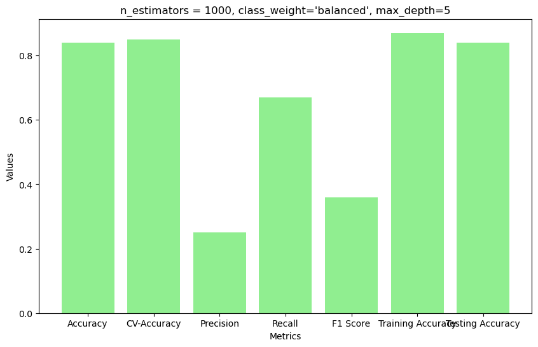
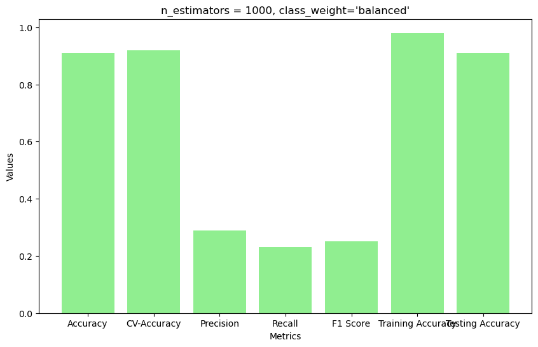
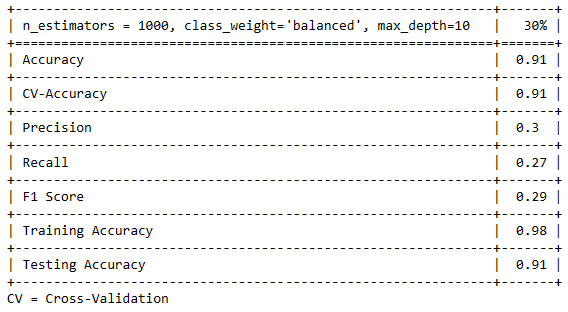
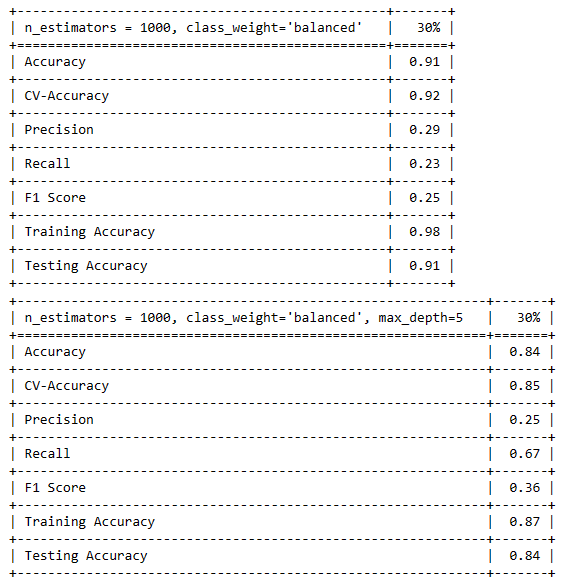
30% test

### *Random Forests - PCA*

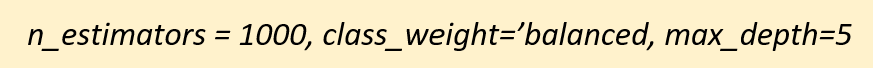
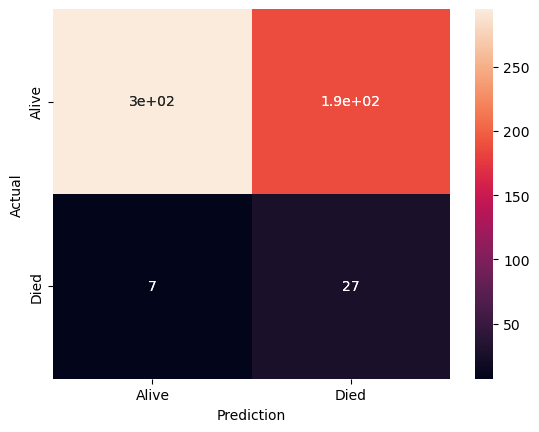
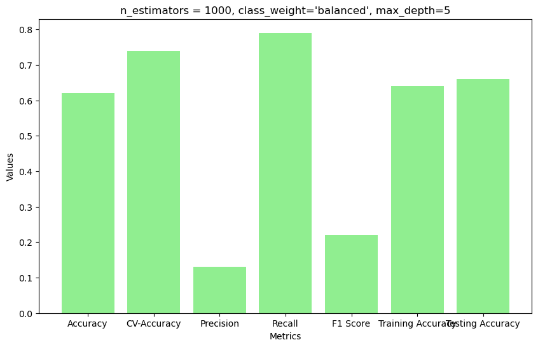
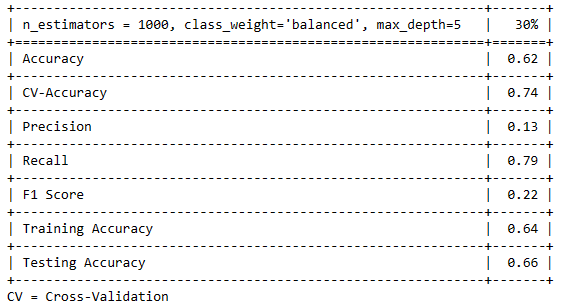
After adjusting the hyperparameters I did not see substantial improvements in the model’s prediction. The recall was slightly high, but I found 5% higher with data without SMOTE and PCA. The precision is another metric that continue low.

In the CM, presented below, we can see that almost double the result for the category false positive when compared with the original data. This result can be considered as a type I error when the model predicts died when the patient is alive.

Adjusting hyperparameters:



The results below are from RF with PCA and focus on the features (DISABLE, CUR\_ILL, HISTORY, ALLERGIES). We can see that the accuracy is not good, but the model performs better on unseen data, according to the CV accuracy as well as the most capture of the positive instances (recall). However, the precision still too low meaning low ability to avoid false positives.



Accuracy (0.62): This is the overall correctness of the model, calculated as the ratio of correctly predicted instances to the total instances. In this case, 62% of predictions were correct.

CV-Accuracy (0.74): This might refer to Cross-Validation Accuracy, which is a measure of how well the model performs on a set of data that it hasn't seen during training. A cross-validated accuracy of 74% suggests that the model performs better on unseen data compared to the overall accuracy.

Precision (0.13): Precision is the ratio of true positive predictions to the total predicted positives. It measures how many of the predicted positive instances are actually positive. A precision of 0.13 indicates that the model has a low ability to avoid false positives.

Recall (0.79): Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total actual positives. It measures how many of the actual positive instances were correctly predicted by the model. A recall of 0.79 indicates that the model is good at capturing most of the positive instances.

F1 Score (0.22): The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall. A low F1 score (0.22 in this case) suggests that there is an imbalance between precision and recall.

Training Accuracy (0.64): This is the accuracy of the model on the training dataset. It indicates how well the model has learned from the training data.

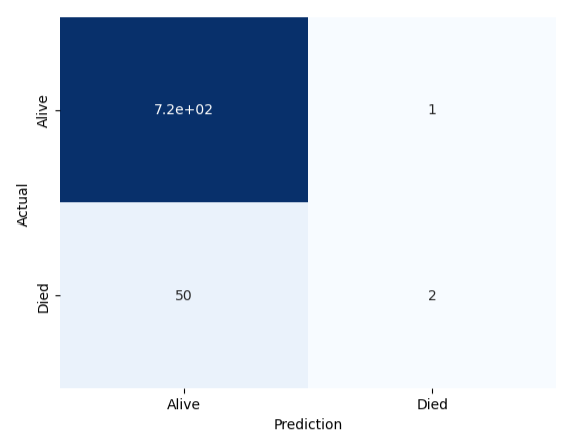
Testing Accuracy (0.66): This is the accuracy of the model on a separate testing dataset. It gives an indication of how well the model generalizes to new, unseen data.

In summary, the model seems to perform moderately well on the training and testing datasets, with better performance on cross-validated data. However, there is room for improvement in terms of precision, recall, and the overall balance reflected in the F1 score. Depending on the specific goals of the model, you might need to fine-tune it to achieve a better balance between precision and recall.

### *Artificial Neural Networks*

A close-up of a graph

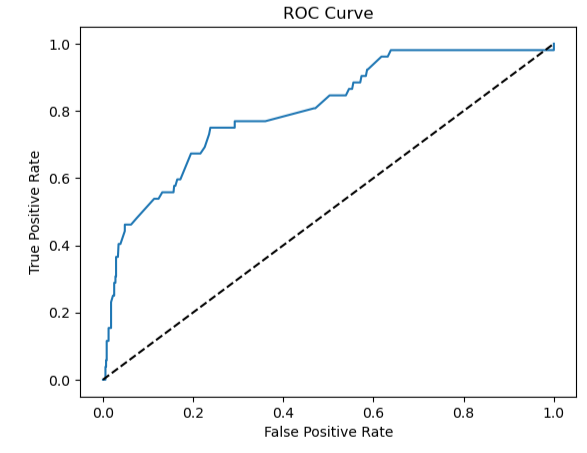
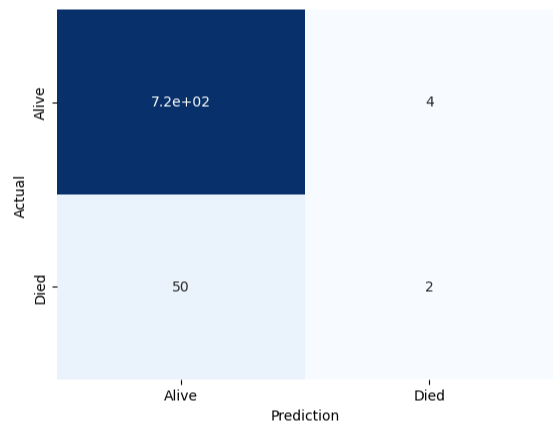
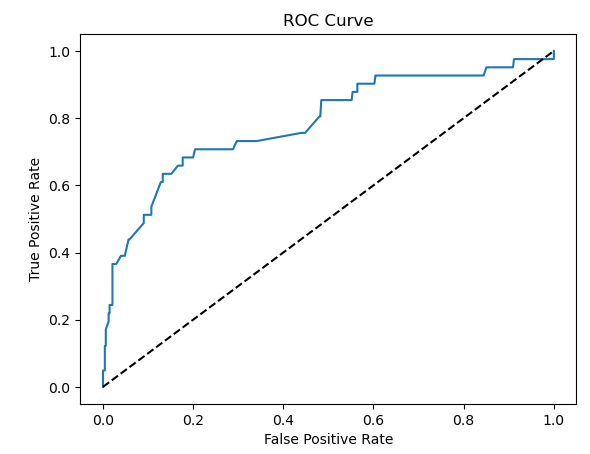
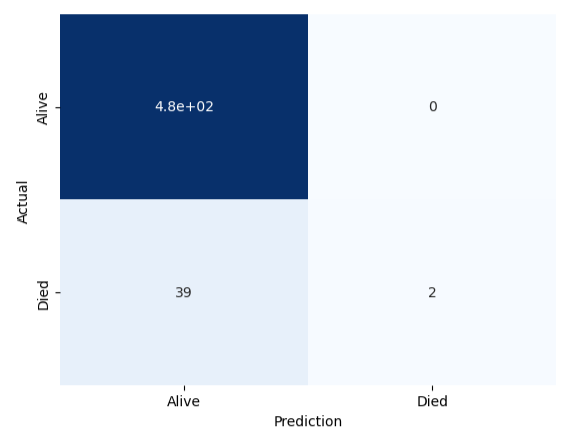
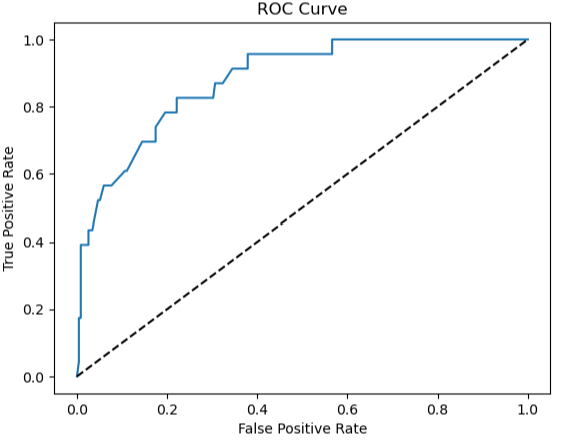
Description automatically generated



30 % test

20 % test

10 % test



Adjusting hyperparameters:

## ***Assessment***

b) Multiple machine learning approaches (at least two) using hyperparameters and a **comparison** between the chosen modelling approaches.

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**

# **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.