**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](#_Toc151910803)

[*Word count* 3](#_Toc151910804)

[**Data Description** 3](#_Toc151910805)

[**Data Cleaning and Preprocessing** 3](#_Toc151910806)

[**Machine Learning Models** 4](#_Toc151910807)

[***Outcomes*** 5](#_Toc151910808)

[*Random Forests (RF-I)* 5](#_Toc151910809)

[*RF-SMOTE* 8](#_Toc151910810)

[*RF-PCA* 9](#_Toc151910811)

[*ANN* 10](#_Toc151910812)

[***Assessment*** 11](#_Toc151910813)

[**Conclusion** 11](#_Toc151910814)

[**Reflective journal** 12](#_Toc151910815)

[**References** 12](#_Toc151910816)

# **Introduction**

Since 2020, more than 7 million people have lost their lives around the world to a lethal virus that causes a disease named COVID-19 (World Health Organization, 2023). This pandemic has substantially impacted global public health, the economy, and people's lives.

Many scientists worked to develop vaccines currently considered the fastest (Khuroo et al., 2020, p.1). As with any other vaccine, the COVID-19 vaccines have been reported to have several side effects, from mild to severe, like fever to cardiac problems and, in some cases, 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 better accuracy in the prediction of people's death after COVID-19 vaccination and identify which feature can contribute to the death risk after the vaccination, such as the presence of allergies or illnesses pre-existed.

## *Word count*

Introduction: 175

Data description: 77

Data preparation and preprocessing: 231

Machine learning models: 239

Outcomes: 123

Assessment: 137

Conclusion: 106

Total: 1088

Reflective journal: 514

# **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 ML models work efficiently with numbers, and it is possible to make standardization and normalization, which is essential for models like ANN (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. 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.

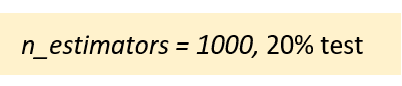
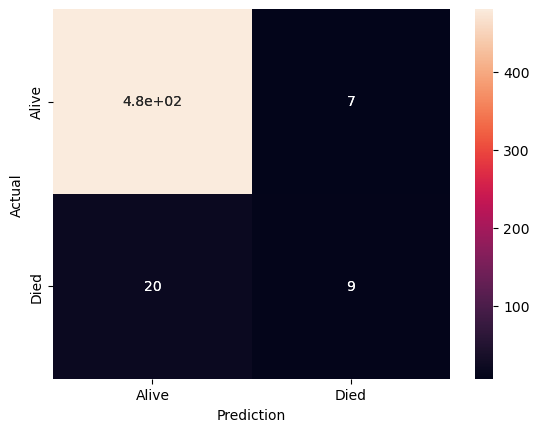
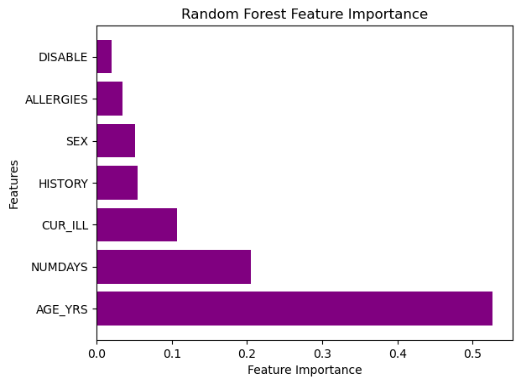
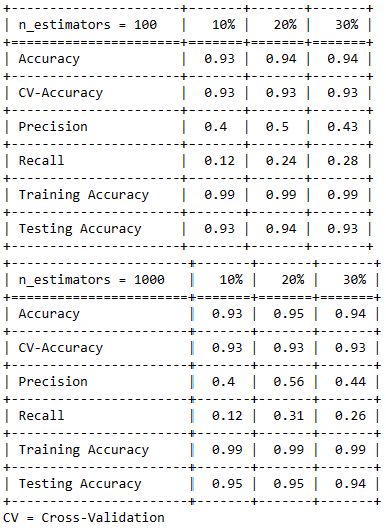
ANN is powerful for classification tasks and works best with features with the same meaning (Müller and Guido, 2017, p.117-118). This model suits my goals as I have homogeneous features and binary target, ideal for the classification method. I also scaled the data as required.

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.

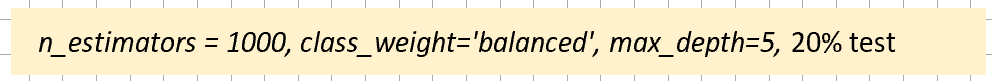
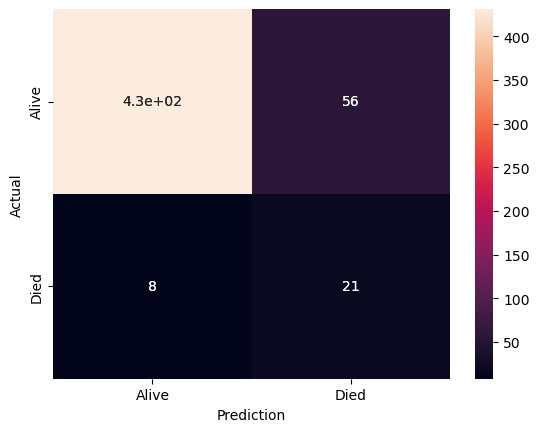
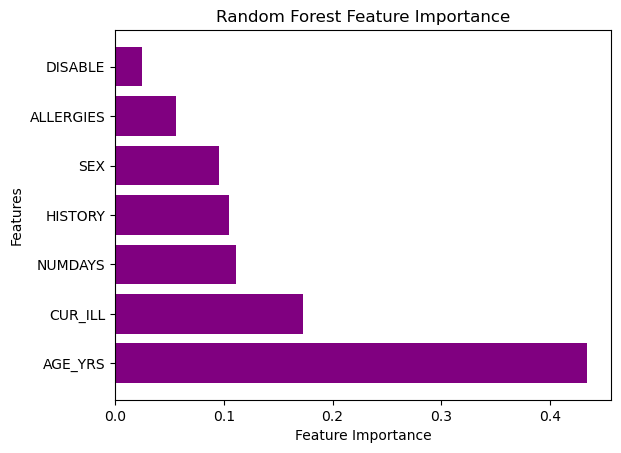
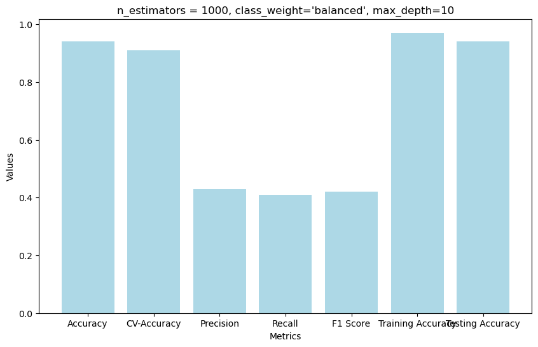
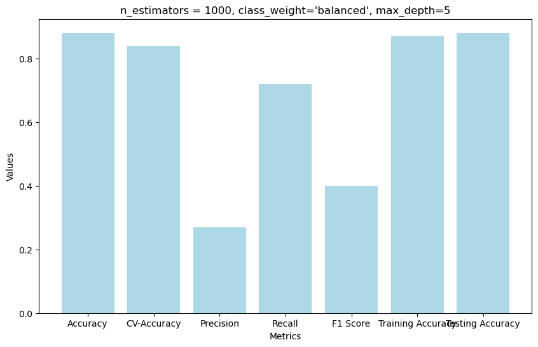
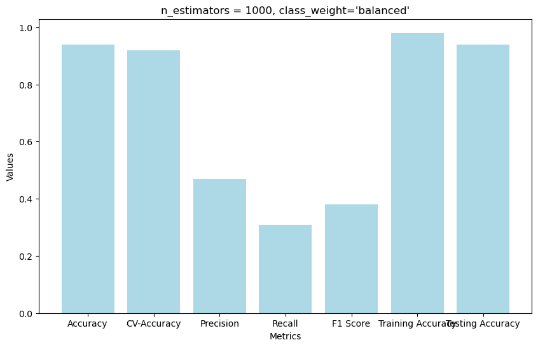
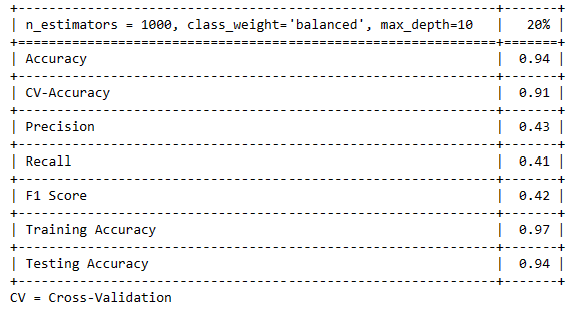
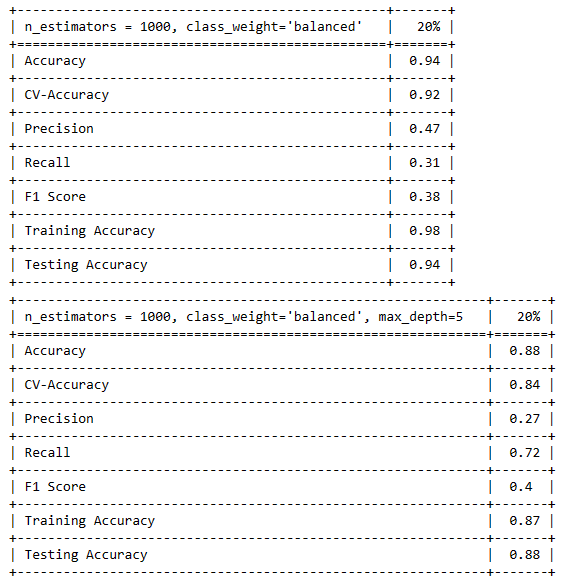
## ***Outcomes***

### *Random Forests (RF-I)*

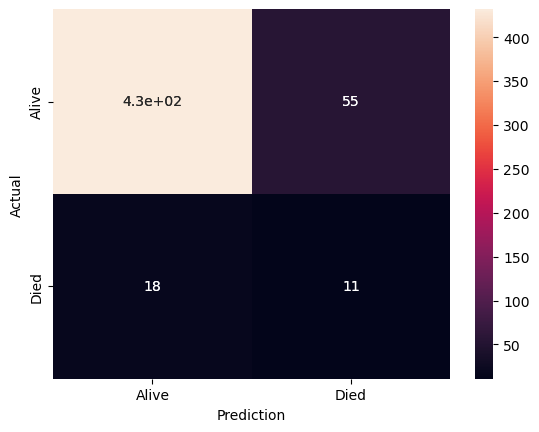
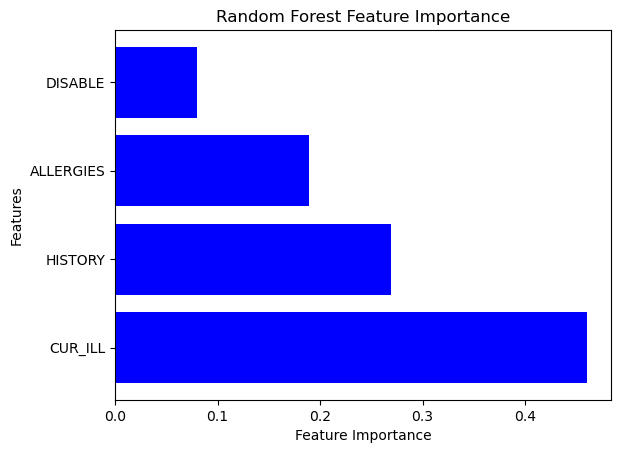
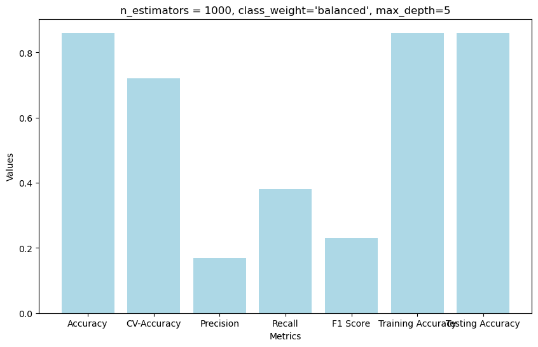
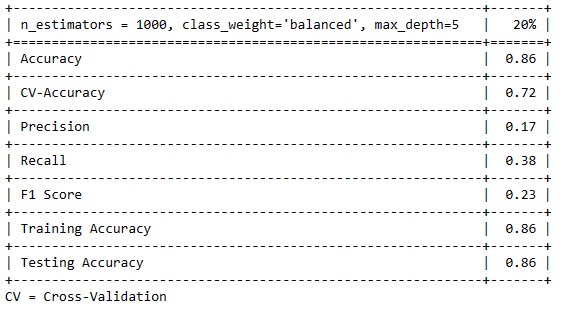
Figures below represent the results for hyperparameters adjustment in the number of trees (n\_estimators) and the sample split.



RF-I (initial data: estimator=1000, class\_weight, max\_dept=5) results of the hyperparameters adjustment, focusing on the 20% test, which presented greater performance anteriorly.

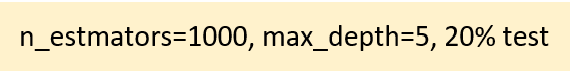
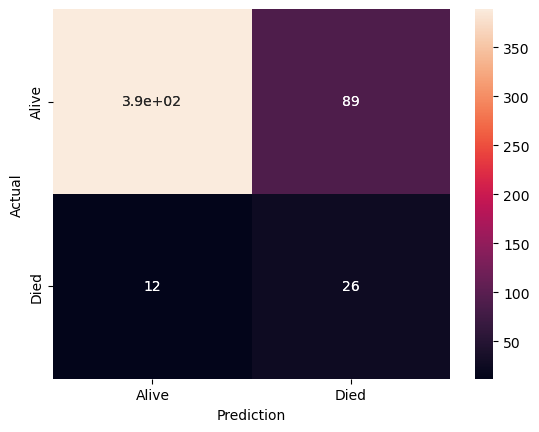
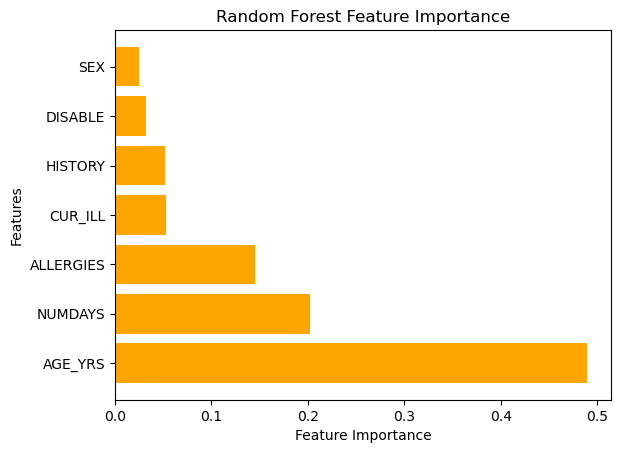
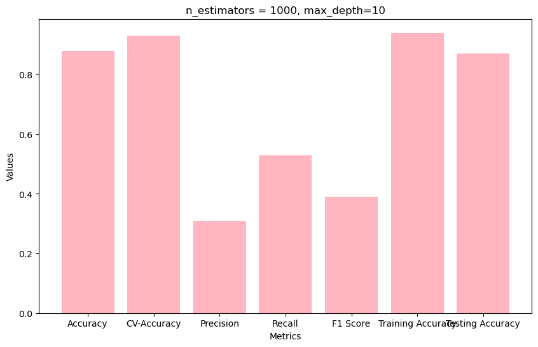
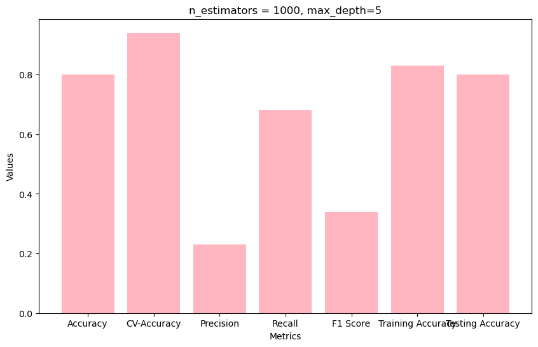
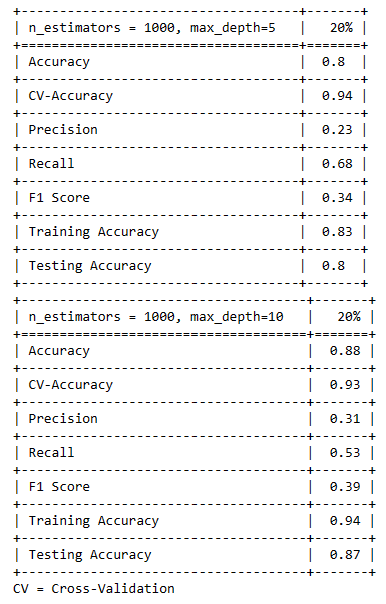


Results for four features in focus to identify which can contribute more to the risk of death. I used the best hyperparameters found earlier.



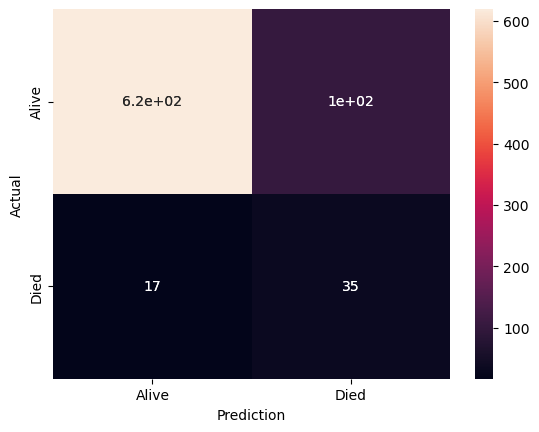
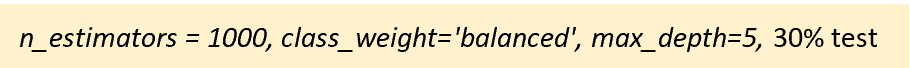
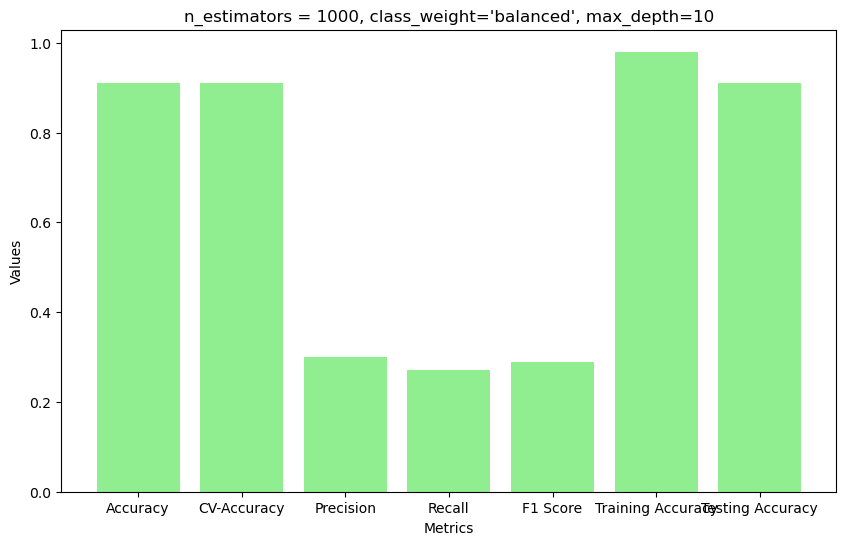
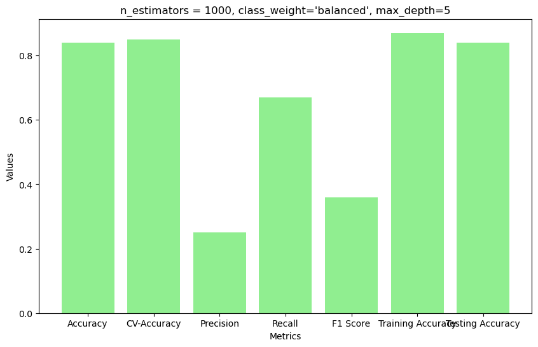
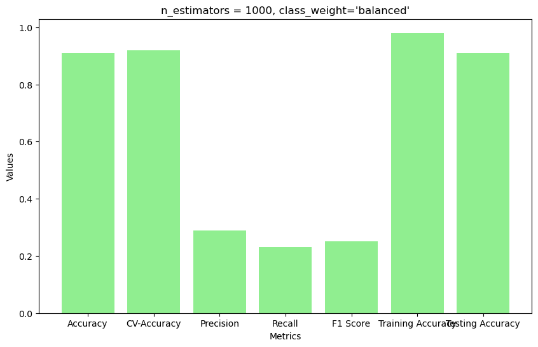
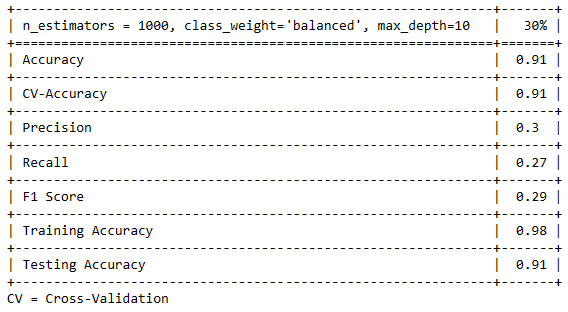
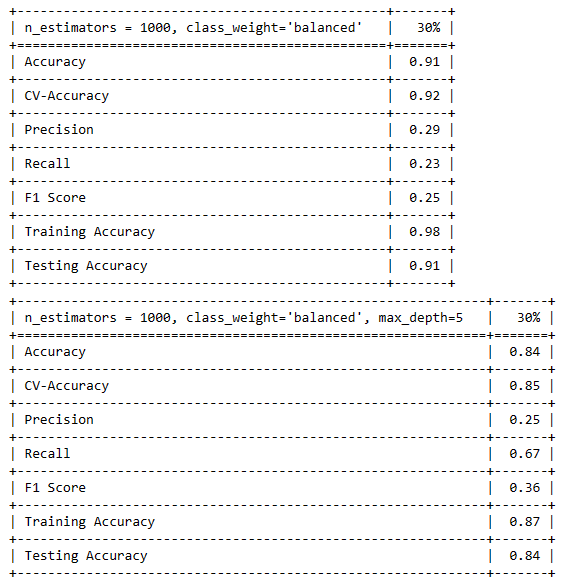
### *RF-SMOTE*

Results for SMOTE applied to the training set, and then RF performance with hyperparameter adjustments.



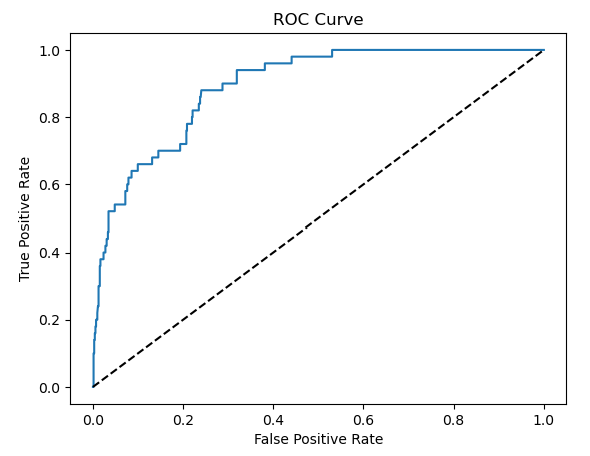
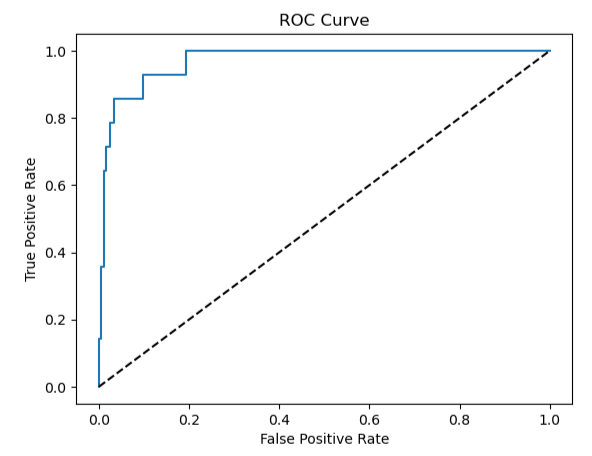
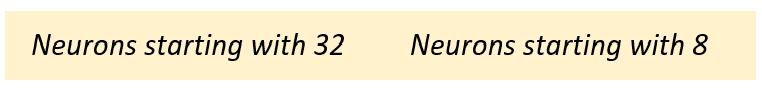
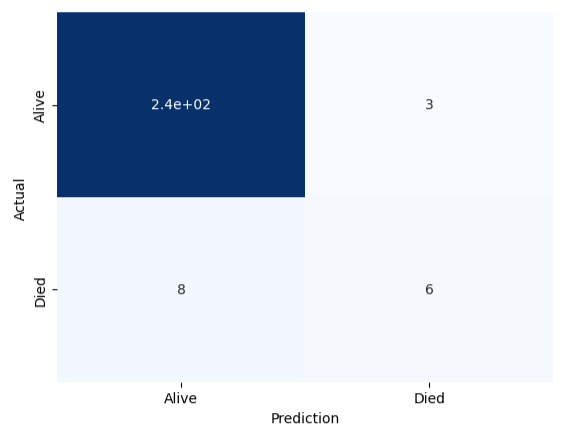
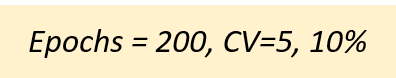
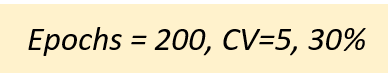
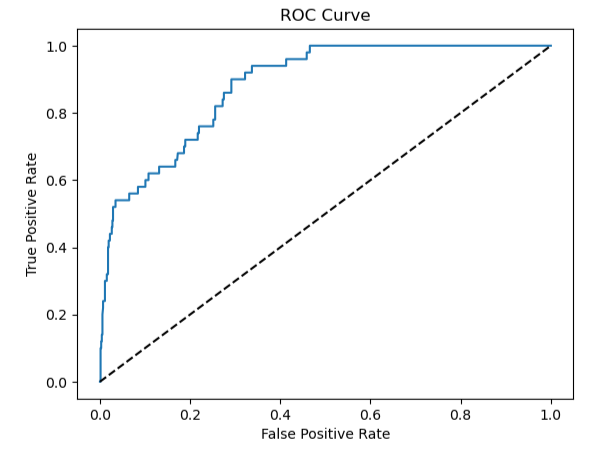
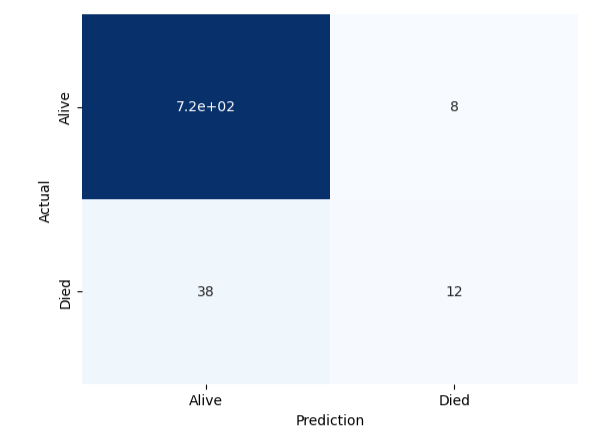
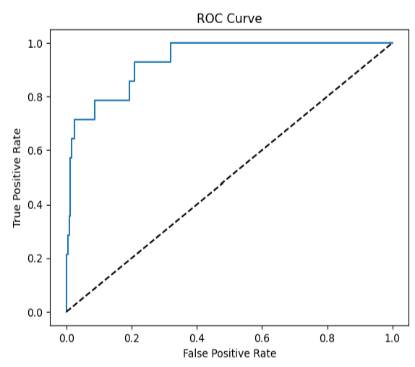
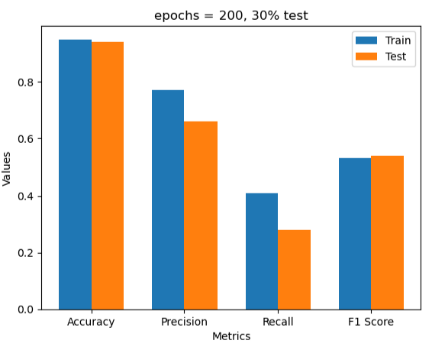
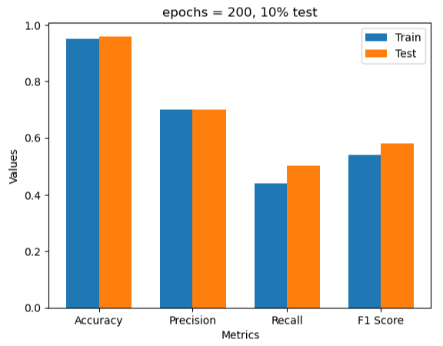
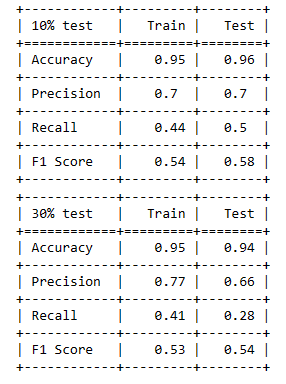
### *RF-PCA*

Results of hyperparameters in RF-PCA.



### *ANN*

Below are adjusting hyperparameter results for the model’s performance with cross-validation and ROC curves for different neuron numbers.

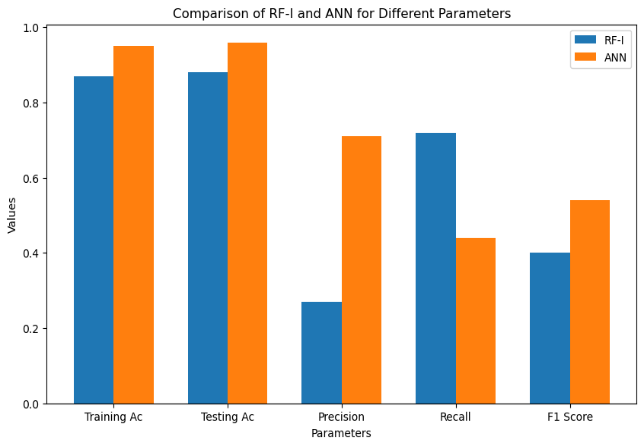
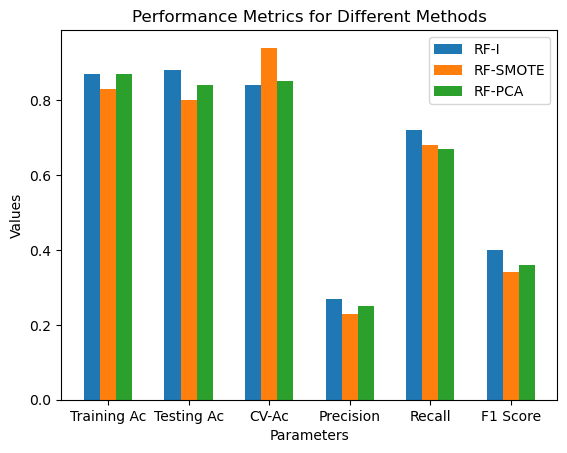


## ***Assessment***

The graph on the left compares the methods used in RF. The RF-I, which already contains class balance in its hyperparameter, was already able to improve performance compared to SMOTE and PCA methods. In addition, RF-PCA had no improvements, probably because it captured ~39% of data variance.

The recall metric was 2.4 times higher RF-I when compared with no hyperparameters adjustment, meaning that even with the adjustments in the imbalanced class, they were not enough to improve the negative class (target class) predictions, and precision remained low in all methods.

When we compare RF-I with ANN (right chart), we notice a significant increase in precision and, in contrast, a reduction in recall; however, the trade-off of these metrics (F1) is still better in the ANN than in RF-I. The precision improvement, 2.6 times higher than RF-I, shows that the model reduced the false positive predictions, which were predicted as positive when negative, indicating that the model performance is improving.



# **Conclusion**

Considering the feature importance results, the presence of any illness (‘CUR\_ILL’) or pre-existing illness (‘HISTORY’) on the vaccination day may contribute to the death of a vaccinated patient. However, more hyperparameter adjustment, other methods to handle class imbalance and a better feature selection will be necessary to improve these models’ performance.

In general, all models provided a reasonable generalization from the training to the test set, accordingly with the cross-validation accuracy, with the ANN proving to be more suitable for this dataset because of the improved precision that reduced the false positive predictions. Overfitting and underfitting did not occur due to the robustness of these models.

# **Reflective journal**

I chose to develop this project independently to avoid prejudice to my colleagues and myself because my schedule was incompatible with theirs. Therefore, I designed all the work, from the data search to charts and conclusions, as well as the choice of the code used.

I started this project searching for data about COVID-19, which was my topic of interest, and for it, I used sites like Kaggle, Google Dataset Search, and Central Statistics Office. Ultimately, I chose a dataset from Kaggle and elaborated on some questions I would like to answer with that data.

Then I started reading the class materials and the book “Introduction to Machine Learning with Python: A Guide for Data Scientists” to choose what machine learning model would be more suitable to my data and answer my questions.

I elaborated a schedule to accomplish all the tasks of this project. However, as I was developing the project, I realized that it would need much more time to do it, mainly because of the amount of text data that I only had the first contact with tools to clean it last week. Thus, it was impractical to restart the project with the time that I had. Because of this, I decided to continue with how I handle those text data, knowing that there are appropriate tools to do it, and I will use them on other occasions. It probably caused the lack of variance in the data that I found in the PCA, but I handled this problem reasonably when I used the ANN model.

Another problem that arose was the poor metrics results for precision and recall, which led me to apply the method known as Synthetic Minority Over-sampling Technique (SMOTE) to reduce the class imbalance that I had because the minority class (target) was almost 15 times smaller than the majority class. It works in the same manner as the hyperparameter ‘class\_weight’ of the Random Forests, proving the robustness of this model in my project. However, I saw an improvement in the recall metric that meant prediction improvement on the positive data.

When I worked with the ANN model, I could see an improvement in the precision metric, meaning that the model improved its ability to predict false positive that was actually negative and can contribute to better results in estimating the target class (negative class in my case). In addition, the cross-validation accuracy showed that the model was generalizing well. Thus, the model was not overfitting or underfitting.

Therefore, there is still room for improvement in the models, as always, and I would like more time to test more methods to improve them. Because in the step that I am in now is when I can see more about my data problems and raise some possible solutions, despite having already tried to deal with some initial problems, such as the imbalanced class. However, I am so grateful for doing this project and others in this course who are contributing to forging my analytical mind, and I am learning even more how to use tools to deal with.

# **References**

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

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

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].

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.

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/>.