**Covid-19 analysis**

(Word count\*: 2,145)

*\* Excluding diagrams, code, references and titles.*

By

Oluwatimileyin Oladipo (Student ID: 2023383)

Leopoldo Rojo Romero (Student ID: 2023355)

Higher Diploma in Science in Data Analytics for Business / AI

Machine Learning

Dr. Muhammad Iqbal

CCT College

Dublin, Ireland

**Table of contents**

[**Introduction** 2](#_Toc151928354)

[**Problem Definition** 2](#_Toc151928355)

[**Objectives** 2](#_Toc151928356)

[**Data Source** 2](#_Toc151928357)

[**Implementation** 3](#_Toc151928358)

[**Early data analysis** 3](#_Toc151928359)

[**Modelling** 4](#_Toc151928360)

[**Conclusion** 9](#_Toc151928361)

[**Individual reflective reports** 9](#_Toc151928362)

[**References:** 18](#_Toc151928363)

# **Introduction**

Coronavirus 2019 (COVID-19) is an infectious disease caused by coronavirus SARS-CoV2 (World Health Organization, 2020). Individuals infected with covid-19 showed signs of mild to moderate respiratory illness, dyspnoea, and ground-glass opacity demonstrated by imaging tests and fever, and in extreme cases, infected individuals died because of covid-19 (Huang et al., 2020).

Various researchers and virologist reported on the determining factors of the survival of patients with covid-19. Risk of death and severity of symptoms were reported higher in patients with comorbidities such as pulmonary, immune system pathology, diabetes, hypertension and smoking (Richardson et al., 2020). (Galvão et al., and Justino Fernández, 2020).

# **Problem Definition**

Our motivation for choosing this project is to understand through analysis the impact of comorbidities and other features on the survival or death of a sample of covid-19 patients. Following our analysis, we intend to build a robust Machine Learning model that helps to predict a patient’s survival.

# **Objectives**

* Perform a descriptive analysis of covid-19 dataset using statistical techniques and plots, as well as displaying relationship between comorbidities and survival of a covid patient.
* Perform data cleaning by engineering features and dropping features not relevant to this project scope.
* Use Random Forest and Gaussian Naive Bayes models that predict if a patient will die or survive covid-19 given patient's current symptoms. Furthermore, perform cross validation and checks to see the performance of the models.

# **Data Source**

Our data was obtained from <https://www.kaggle.com/> in CSV format. Ethical considerations were discussed and as shown in *fig.1* data was publicly available for everyone for research purpose.

A screenshot of a computer

Description automatically generated

*Figure 1: Licencing screenshot for dataset.*

# **Implementation**

In its raw form, data had 21 columns and 1048575 rows, and after cleaning, there were 388878 rows and 16 columns in our data.

# **Early data analysis**

Data shows that most of the columns were encoded, other columns we manually encoded as guided by data dictionary in preparation for model building. [ICU & INTUBATED] columns were dropped because they had over 80% null-values and avoid altering data trueness through imputation.

Figure 2: Countplot for age categoryA bar graph with different colored bars

Description automatically generated

Age column was grouped into 4 age categories, children age (0-15), adult (15-30), old (31-60) and senior (61-120) are category 0,1,2,3,4 respectively. Count plots *fig:2* show 2, have the highest count followed by 3, 1 and 0 in that order.

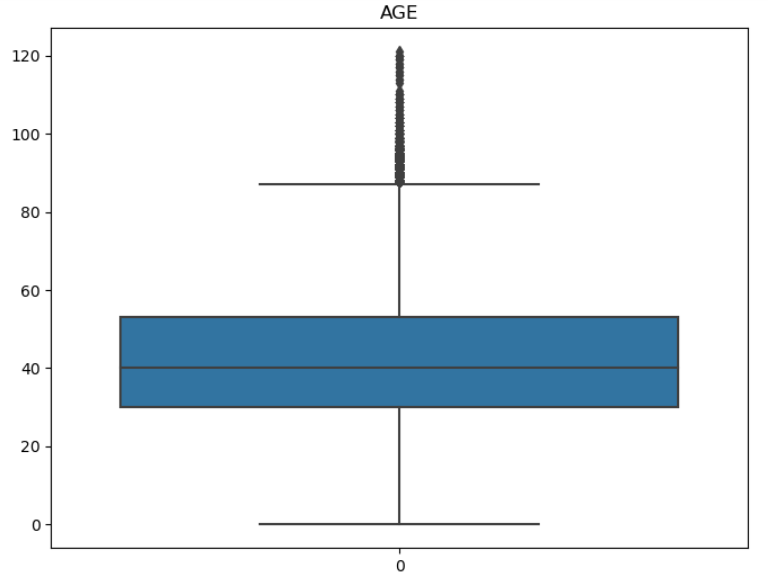


Figure 3: Boxplot for age

Boxplot*(fig 3)* shows that the majority of the patients are aged between 0 and approximately 95 years old, with some outliers above 100. Domain knowledge suggests, covid-19 has severe effects on old patients compared to younger patients, hence outliers are significant. 0 years old could be babies that are less than a year old. *Fig:4* is a correlation plot of death and survival patients in each age category. Overall, more people survived covid-19 in our sample data (*fig 5)*.

A graph with blue and orange bars

Description automatically generatedA graph with a number of squares

Description automatically generated with medium confidenceFigure 4: Age cat vs Death Figure 5: Proportion of died to survived.

# **Modelling**

**Why Gaussian Naive Bayes and Random Forest (GNB and RF)?**

We chose the died column as our target to build a model that predicts if a covid-19 patient survived covid-19 or not given independent variables. We chose these models because both are used for classification tasks. RF builds multiple decision trees and merges their predictions; therefore, it is more accurate and robust to outliers and noisy data. GNB is a probabilistic supervised learning model, that uses conditional probability to classify target.

**Results:**

**Random Forest:**

**A screenshot of a computer code

Description automatically generated**

Figure 6: Random Forest fitting

RF was fitted on at (10,20,30,40) % testing size on n\_estimators (50,100,150) and max\_depth of 3. **Max\_depth** represents depth of each individual decision tree and **n\_estimators** represent the number of decision trees that will be built in the random forest, both are hyperparameters that helps us build a more robust model.

As shown in *Fig 7*, when **n\_estimators**= 150 and **Max\_depth** =3 **precision** of RF was (63,62,62,64) at testing size of (10,20,30,40) % respectively and **accuracy** of 89 across 4 testing trail.

A close-up of a number

Description automatically generated*Figure 7: Evaluation of n\_estimator of 150 and max\_depth of 3*

*Fig 8* shows when **n\_estimators**= 100 and **Max\_depth** =3 **precision** of RF was (63,62,62,64) at testing size of (10,20,30,40) % respectively and **accuracy** of 89 across 4 testing trail, obviously there was no improvement to the precision and accuracy.

A close-up of a number

Description automatically generated

Figure 8: Evaluation of n\_estimator of 100 and max\_depth of 3

*Fig 9* shows when **n\_estimators**= 50 and **Max\_depth** =3 **precision** of RF was (73,71,69,73) and **accuracy** of (87,88,88,87) at testing size of (10,20,30,40) % respectively. Our model predicted more accurately and precisely when reduced the n\_estimators to 50.

**RF Confusion matrix:**

Fig 10 shows model predicted 8336 false positives (FP) than true positives, the FP are patients that did not die from covid-19 and model predicted them dead. However, model predicted more true-negative TN than false negative (FN) meaning model predicted to greater extent correctly patient that survived.

A graph of a number

Description automatically generated with medium confidence

Figure 9: RF confusion matrix before balancing

**Gaussian Naive Bayes:**

A screenshot of a computer code

Description automatically generated

Figure 10: GNB implementation

GNB was fitted on at (10,20,30,40) % testing size. As shown in *Fig 10*, when at testing size of (10,20,30,40) % **precision** of GNB was (47,46,46,47) respectively and **accuracy** of 85 across 4 testing trails.

**GNB Confusion matrix:**

**A screenshot of a graph

Description automatically generated**

Figure 11: GNB confusion matrix before balancing

*Fig 11* shows model predicted 9017 false positives (FP) and 7928 true positives, the FP are patients that did not die from covid, and model predicted them dead. However, model predicted 58047 true negatives (TN) than false negative (FN) meaning model predicted to greater extent correctly patient that survived.

We believe this was due to imbalance in our data, there were more patients that survived than died hence the model was bias toward patients that survived. To improve our model performance, we tried scaling, adjusting our target encoding and PCA but none of these approaches affected our model performance. We did minmax scaling because our data was skewed, and PCA to reduce dimensionality of our data, hence improving our model performance. Finally, we balanced our data because there were more patients that survived than died, hence our model was bias.

**After balancing:**

**Random Forest:**

A screenshot of a computer program

Description automatically generated

Figure 12: Model performance after balancing

After balancing, our RF model performance was significantly improved, our precision and accuracy move up to 83% and 88% respectively. The overall model performance (F-1 score) shows 89% in fig 13. Result are shown in fig 12 for testing size (10,20,30,40) % and n\_estimator of 50 and max\_depth of 3.

A screenshot of a computer

Description automatically generated

Figure 13: F1-Score after balancing

**GNB after balancing:**

After balancing the GNB accuracy went down and the precision when up as shown in fig14, its overall performance however was 89% as shown in fig15.

Figure 14: GNB evaluation after balancingA screenshot of a computer code

Description automatically generated

A close-up of a message

Description automatically generated

Figure 15: GNB F-score after balancing

**Confusion Matrix:**

A screenshot of a computer

Description automatically generatedA close-up of a chart

Description automatically generated

Figure 16: RF and GNB confusion matrix after balancing

As shown in fig 16, RF predicted better after balancing than GNB, the are more TP and TN than FP and FN in the RF prediction. In GNB, model predicted more TP than FP, while it predicted more FN than TN.

# **Conclusion**

Haven carefully evaluated our model results, we can observe that a RF model performed better for this sample data. This model is consistently predicting if a person died or survived Covid-19 with an Accuracy of 86% - 88% with a proportion of actual positives correctly identified (Recall) of 81% (survived) and 95% (died).

The results were obtained using:

1. Metrics from Sklearn to evaluate the Accuracy and Precision.

2. Classification report from Sklearn.metrics to evaluate Precision, Recall, F1-score and Accuracy.

3. Cross validation score from Sklearn.model\_selection to evaluate the Accuracy dividing the data into multiple subsets to obtain a reliable metric and compare it with our previous ones.

# **Individual reflective reports**

**Oluwatimileyin Oladipo (Student ID: 2023383, word count: 506)**

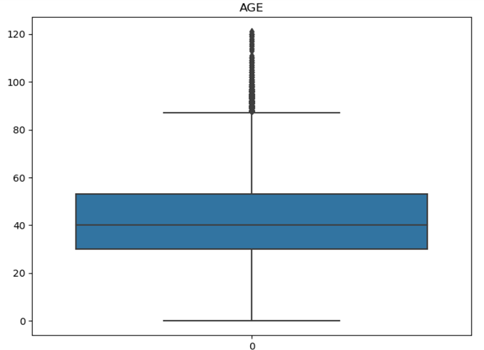
As a pair we decided to work on covid-19 case. During the deliberation, I looked for the dataset we used for our project and after we both reviewed the data, we agreed to use it.

**Data Preparation (DP):**

Data preparation also known as data preprocessing in a crucial step to a reliable data analysis and modelling. This process involves early data analysis (EDA), data cleaning, transformation, handling imbalanced data, feature engineering, handling imbalanced data, visualisation, and feature selection. DP takes over 80% of modelling process because if we put ill-prepared data into our machine learning model, we would get bad results, hence having through DP is imperative to getting a reliable result from our model. During our project, I prepared our data for analysis.

**EDA:**

EDA helps us to understand the dataset by summarizing its characteristics and often visualising them with plots. During this step I was able to familiarise with the dataset, what columns we have in our data, and how we can use them for our analysis. Furthermore, with EDA I was able to visualise distribution patterns in numerical variable for better understanding *fig 17.*

*Figure 17: Boxplot for age*

**Data Cleaning (DC):**

DC is the process of making our data clean for modelling, a task I handled during our project. DC involves but not limited to handling missing values, engineering features, transforming data. Initially 'isnull()' assessment showed no null-values, further exploration revealed missing-values coded as integers (95-98). To handle this, I converted them to NAN in other to pinpoint their locations and quantity. [ICU, INTUBATED] columns had over 80% NAN, to prevent introducing bias through imputation method I dropped them and used ‘dropna()’ I dropped other rows with NAN values.

**Feature engineering:**

The Date\_died variable contained patient that survived and died in ‘999-999-99’ and ‘dd/mm/yy’ format respectively, In other to extract our target variable from this column, I wrote a function that iterates through the variable and return 1 if patient died and 2 if patient survived as shown in *fig.18* I then assigned this call function to a new column called ‘died’, hence extracting our target variable. A screenshot of a computer code

Description automatically generated

Figure 18: Function to create target variable.

**Manipulating data to improve model performance:**

Our Initial machine learning setup had poor precision. To Optimise model performance, my approach involves dropping classification finals, scaling independent variables *fig.19,* and balancing the dataset *fig.20.*

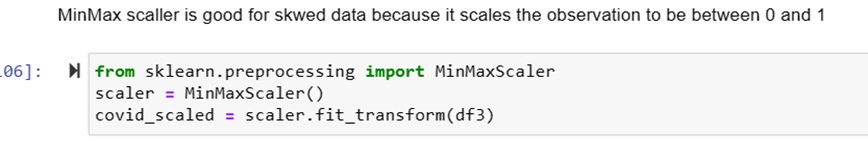


Figure 19: Scaling the independent variables.

I scaled both random forest and Gaussian Naive Bayes models, averaging accuracies at 0.89% and precisions of 0.55. Despite adjusting the target value and eliminating noise, improvements in

A white background with black text

Description automatically generated

Figure 20: Balancing x and y variables.

accuracy and precision were limited. Utilizing Principal Component Analysis (PCA) *fig.21* aims to reduce dimensionality, enhancing model performance by curbing noise in dataset. Additionally, scaling the dependent variable ensures uniformity, a pivotal feature engineering step for improved model accuracy.A blue line graph with numbers

Description automatically generated

Figure 21: PCA plotting.

**Conclusion:**

Data preprocessing is an important process in any data analysis and modelling process. “Garbage in Garbage out” meaning the quality of result is dependent on the quality of data passed into machine learning algorithm. My main contribution during our project was to prepare our data. After preparation during modelling process, I also contributed to manipulating our data, to optimise model performance. After completion of our coding, I drafted our report in collaboration with my group mate.

**Leopoldo Rojo Romero (Student ID: 2023355, word count: 501)**

**Contributions**

Following the start of the Exploratory Data Analysis (EDA), we realized that we have got missing values in our data, thus I performed the following code to find the percentage of missing values in each feature.

Tabla

Descripción generada automáticamente

Figure 22: Missing values in each feature (%).

After we notice that two of our features had more than 80% of missing values, we decided to remove them due to the did not hold enough data for our analysis objectives.

Succeeding our data preparation, we chose Random Forest and Gaussian Naive Bayes machine learning models for this dataset due that both are used for classification problems, first using decision trees and latter using probabilities.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Figure 23: Random Forest (RF) model.

Texto

Descripción generada automáticamente con confianza media

Figure 24: Results of training-testing RF model.

Gráfico

Descripción generada automáticamente con confianza baja

Figure 25: Classification report of the RF model.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Figure 26: Cross validation of the RF model.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Figure 27: Gaussian Naive Bayes (GNB) model.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Figure 28: Results of GNB model.

Imagen que contiene Gráfico

Descripción generada automáticamente

Figure 29: Classification report of GNB model.

Gráfico, Gráfico en cascada

Descripción generada automáticamente

Figure 30: Cross validation of the GNB model.

After testing our models, we noticed that we required to perform further data preparation in order to get better results, therefore we proceed to implement the following:

* Remove an additional feature that was not adding any value
* Scale our data since it was slightly skewed
* Reduce dimensionality using PCA
* Balance our data due that our machine learning models were bias to patients that survived

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Figure : RF model after further data preparation.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Figure 32: Results of RF model.

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Figure 33: Classification report of the RF model.

Imagen que contiene Texto

Descripción generada automáticamente

Figure 34: Cross validation of the RF model, after further data preparation.

Subsequently, we got to a machine learning model that was consistently predicting if a Covid-19 patient survived or died with and accuracy between 86% - 88% with a proportion of actual positives correctly identified of 81% for patients who survived and 95% for patients who did not survive.

**Team dynamics**

For this assignment and as a team, we agreed to work remotely using Whatsapp as our communication tool. Throughout these weeks, we shared ideas, information and we have been in constant communication regarding our assignment, whilst completely understanding each other’s availability from our everyday lives, thus, we have balanced the workload of the project, nevertheless, we were involved in every phase.

Furthermore, we concurred to finish our team task first since we both understand the importance of our time, we believe that carrying out our assignment with this methodology allowed us to have a better time management.

**Learning journey**

From the beginning of this assignment, we used our classes notes for guidance which helped us in the process of each phase of the project; we constantly realized that we required to implement new techniques, for instance, hyperparameter optimization to our machine learning models.

We know that the knowledge gained during this assignment will be of great support throughout the two semesters of our Higher Diploma.

**Conclusions**

Overall, as a team we think that our topic of Covid-19 led us to revise and reinforce everything that we have learned at CCT College, owing to the fact that we applied multiple techniques while performing our analysis to accomplish the objectives of this assignment, while been in fact, an interesting and meaningful topic from our point of view due to the impact that had in the world.

In my personal experience, working with my teammate is helping me to continue developing my soft skills such as, time management, problem solving, listening, critical thinking, communication, and teamwork; this assignment definitely took us beyond what we have learned in classes and pushed us into new topics in regard to Python coding.

# **References:**

*Galvão, M.H.R., Roncalli, A.G., Galvão, M.H.R. and Roncalli, A.G. (2020). Fatores associados a maior risco de ocorrência de óbito por COVID-19: análise de sobrevivência com base em casos confirmados. Revista Brasileira de Epidemiologia, [online] 23.* [*doi:https://doi.org/10.1590/1980-549720200106.*](.)

*Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M. and Xiao, Y. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet, [online] 395(10223), pp.497–506.* [*doi:https://doi.org/10.1016/s0140-6736(20)30183-5.*](.)

*Jamal, M., Shah, M., Almarzooqi, S.H., Aber, H., Khawaja, S., El Abed, R., Alkhatib, Z. and Samaranayake, L.P. (2020). Overview of transnational recommendations for COVID‐19 transmission control in dental care settings. Oral Diseases.* [*doi:https://doi.org/10.1111/odi.13431.*](.)

*Justino Fernández (2020). Catálogo de las exposiciones de arte en 1951. Anales del Instituto de Investigaciones Estéticas.* [*doi:https://doi.org/10.22201/iie.18703062e.1952.sup1.2454.*](.)

*kaggle.com. (n.d.). Covid-19 Survival Prediction (EDA and MODEL). [online] Available at:* [*https://www.kaggle.com/code/anishwarbehera/covid-19-survival-prediction-eda-and-model?cellIds=24&kernelSessionId=135588147*](https://www.kaggle.com/code/anishwarbehera/covid-19-survival-prediction-eda-and-model?cellIds=24&kernelSessionId=135588147%20) *[Accessed 15 Nov. 2023].*

*Lauer, S.A., Grantz, K.H., Bi, Q., Jones, F.K., Zheng, Q., Meredith, H.R., Azman, A.S., Reich, N.G. and Lessler, J. (2020). The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Annals of Internal Medicine, [online] 172(9), pp.577–582.* [*doi:https://doi.org/10.7326/M20-0504.*](.)

*Richardson, S., Hirsch, J.S., Narasimhan, M., Crawford, J.M., McGinn, T., Davidson, K.W., Barnaby, D.P., Becker, L.B., Chelico, J.D., Cohen, S.L., Cookingham, J., Coppa, K., Diefenbach, M.A., Dominello, A.J., Duer-Hefele, J., Falzon, L., Gitlin, J., Hajizadeh, N., Harvin, T.G. and Hirschwerk, D.A. (2020). Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. JAMA, 323(20).* [*doi:https://doi.org/10.1001/jama.2020.6775.*](.)

*World Health Organization (2020). Coronavirus disease (COVID-19). [online] World Health Organization. Available at:* [*https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19*](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19%20) *[Accessed 21 Nov. 2023].*

*World Health Organization (2021). The true death toll of COVID-19: estimating global excess mortality. [online] World Health Organization. Available at:* [*https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality.*](https://www.who.int/data/stories/the-true-death-toll-of-covid-19-estimating-global-excess-mortality.)