# Improving the medical care: A Data Analytic Approach

## Introduction

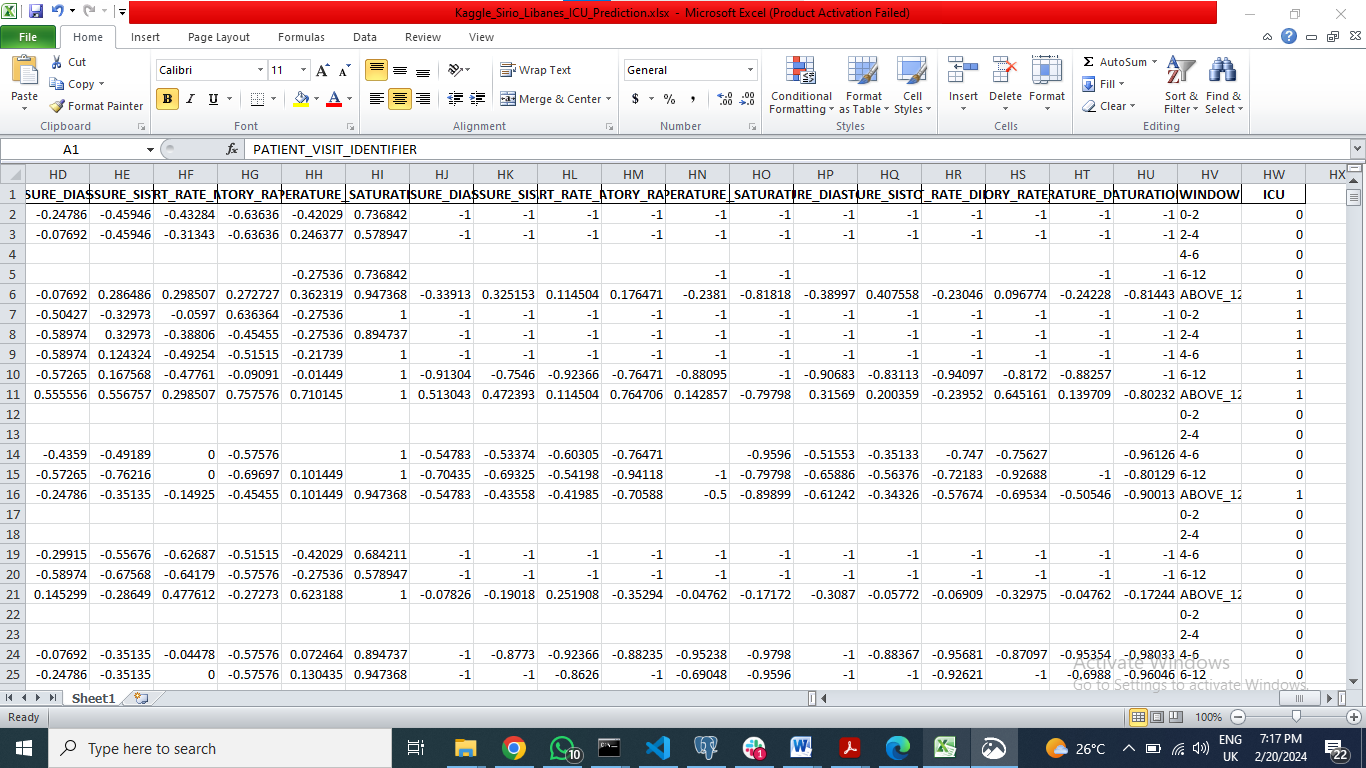
The medical center, because of the changing nature of the medical landscape, has taken a proactive stance to meet the varied demands of its patient kin. This integrates a well-focused attention to personalized care and an integrated system to optimize the service with a robust system (Suryadevara, C.K., 2020) designed to help identify those patients who require customized treatment or special attention. Through the medium of a structured pre-consultation questionnaire, patients are actively involved in listing down their health problems, symptoms, and preferences, which are, however, immensely important pieces of information that can help us understand each individual's special healthcare needs. The ultimate goal is the development of an algorithm that would reliably detect patients at risk of further deterioration (Rust, L.O., Gorham, T.J., Bambach, S., Bode, R.S., Maa, T., Hoffman, J.M. and Rust, S.W., 2023); these patients are those in need of an enhanced medical evaluation, particularly during the COVID-19 diagnosis and management. Through the analysis of patient reactions and grouping them into certain profiled images, the medical center will strive to achieve better diagnostic accuracy, formulate proper treatment methods, and, as a result, boost overall patient results. Launching such an initiative revives the institution's long-lasting values statement of patient-centric care and innovation in healthcare.

## Abstract

This paper describes a new method of healthcare delivery in the rapidly changing healthcare system, which will be focused mainly on proper diagnosis and management of COVID-19-infected patients (Rai, P., Kumar, B.K., Deekshit, V.K., Karunasagar, I. and Karunasagar, I., 2021). The medical center has adopted a procedure in which the patients provide the required information initially through a structured pre-consultation questionnaire. Leveraging this dataset, the study seeks to design an algorithm that is capable of identifying persons who require specialized medical attention or tailored decision-making, and thus, this would enrich diagnostic accuracy and treatment efficiency. The analysis of patients' responses by clustering them into distinct profiles is what allows healthcare professionals to perfect their treatment strategies and fine-tune the treatment outcomes. One of the most important aims of this research is to improve healthcare delivery by personalizing care and specialized allocation of resources. This study illustrates that the hospital is deeply focused on revolutionary medicine, patient-oriented healthcare, and epidemic prevention while the struggle of the pandemic is still ongoing.

## Data Collection and Preparation

The data set used here was retrieved from the Kaggle website and selected from hospital Sírio-Libanês located in São Paulo and Brasilia. The data is masked to comply with international standards and best practices for privacy protection. It includes clinical information on COVID-19 patients diagnosed, varied and covering different dimensions of demographics, medical history, laboratory results, and vital signs.



Title: ICU Prediction Dataset Overview: Data Snapshot

### Dataset

Data on COVID-19 cases has been collected by the Sírio-Libanês hospital in São Paulo and Brasilia and used to forecast better health care and inform of system overload. The data includes demographic information, patient history, blood tests, and vital signs. The hospital is designing notebooks to integrate data lifecycle processes that concentrate on high-level feature engineering, early detection, and accuracy improvement.

The acquisition of the dataset will be followed by several steps to preprocess and prepare the data for analysis (Duong, H.T. and Nguyen-Thi, T.A., 2021). First, data cleaning activities need to be performed, which include handling missing values, outliers, and inconsistencies to have a clean data set. The missing data were imputed using the most appropriate methods based on the available data, such as the forward/backward filling, to keep temporal integrity and the order of the data.

Secondly, feature engineering cleaned the dataset to find crucial details and improve the predictions. This entailed obtaining summary statistics, including mean, median, maximum, minimum, and percentage differences and correlation matrices for clinical parameters at different time points.

The Min-Max Scaler technique was used as the final step to perform feature scaling, where the values of the features were normalized across the dataset later. The normalization method was introduced so that all relevant features were represented uniformly on the interval [-1, 1], thus reducing the convergence time and speeding up the training of machine learning models.

Besides, the dataset was labeled chronologically in time windows, which gives a temporal dimension to the patient meetings; the one-time window is a time step of the patient encounter.

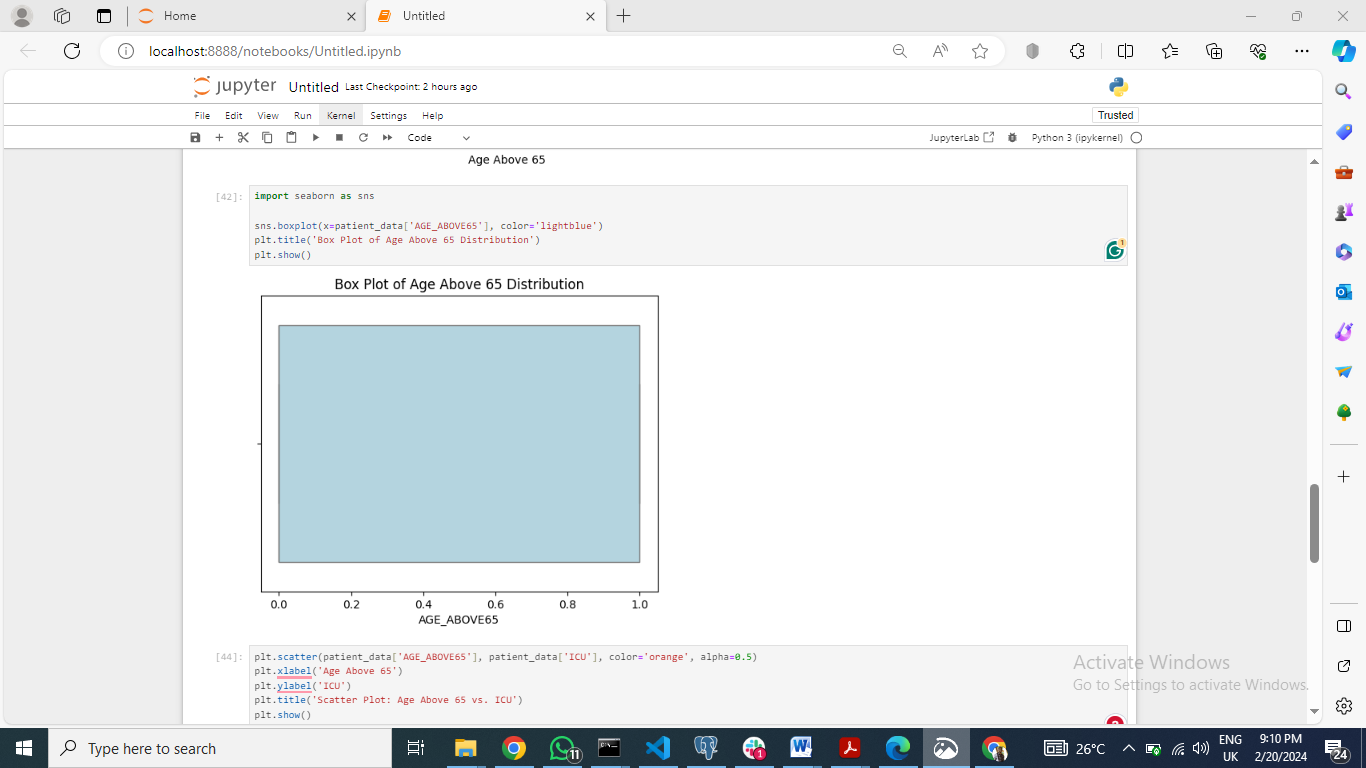
This window tool made it possible to incorporate temporal elements of clinical data, enabling the analysis of trends and patterns.

## Preprocessing Steps and Problem Scope

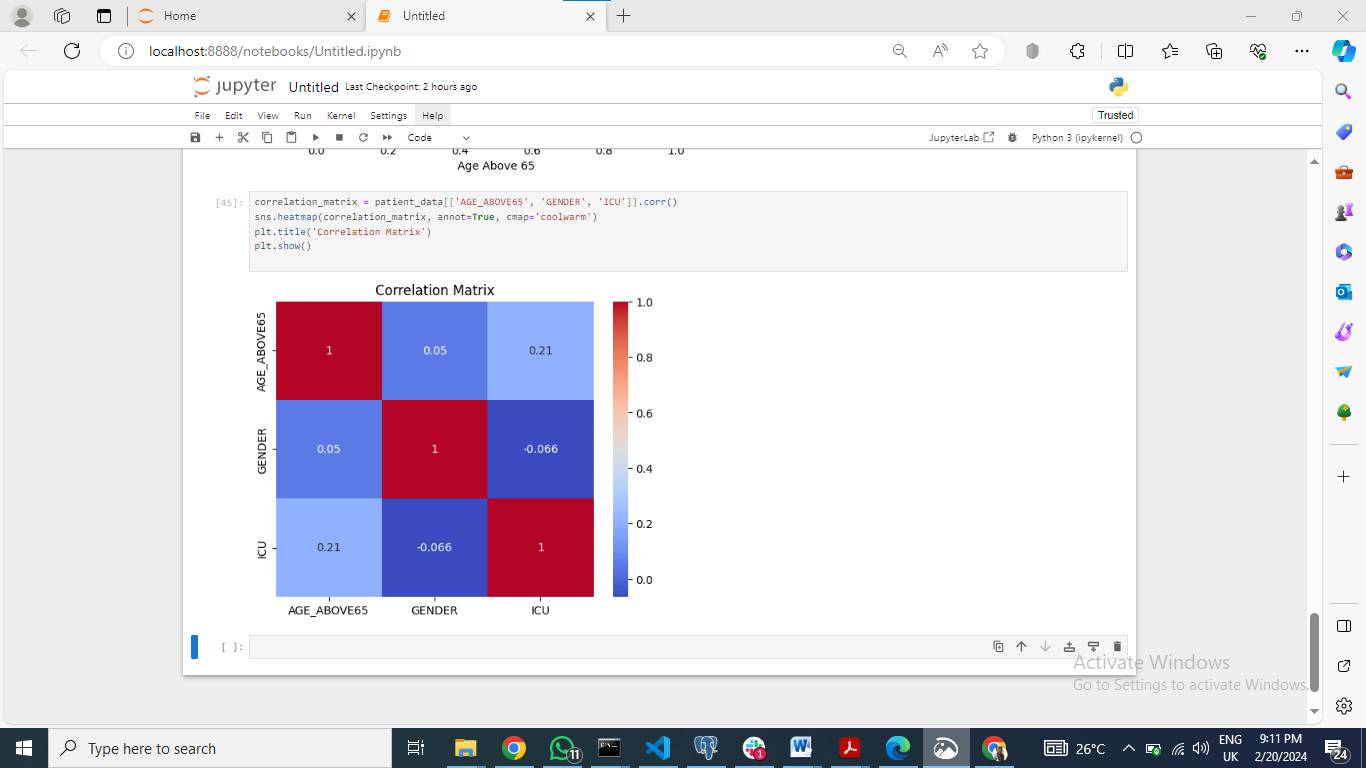
The preprocessing steps cannot be underestimated as they ensure the data is correct and usable for the analysis. The reference is to imputing missing values, outlier detection, and feature scaling here. Techniques that include imputation and outlier detection are commonly used, and scaling is employed to preprocess the data. Besides that, categorical variables need to be encoded to be suitable for the model's training.

In the context of the given problem of classifying COVID-19 confirmed cases for possible intensive care (ICU) admission, the scope (range) of the problem is well defined as a classification task. The goal is to classify patients into two groups: those who must get into ICU and those who should not. Herein presents a classification problem that can be solved by machine learning algorithms such as logistic regression, random forest, and gradient-boosting classifiers.

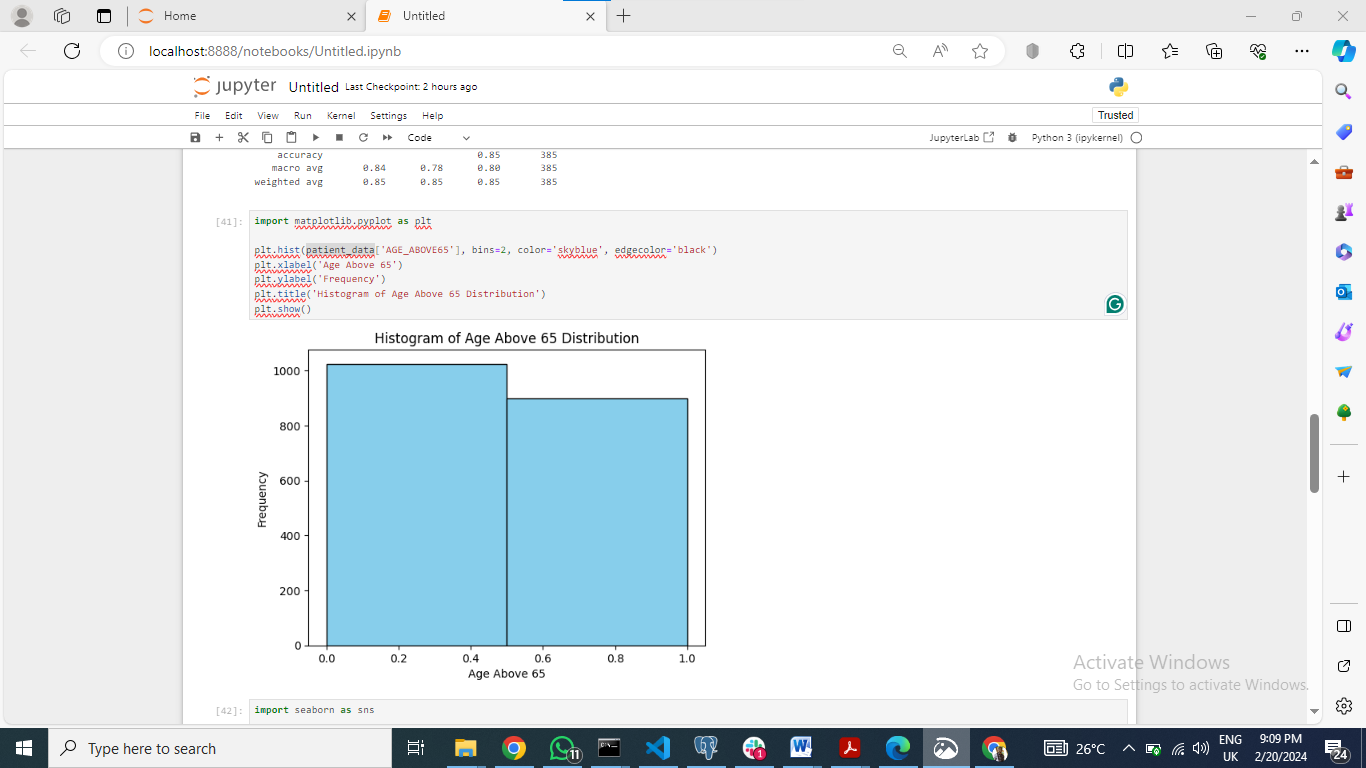
Visualization techniques are also utilized to understand data distribution, pattern identification, and variable association relations. Graphics such as histograms, box plots, scatterplots, and correlation matrices are often applied to visualize the data and ensure it is clean and prepared for analysis.



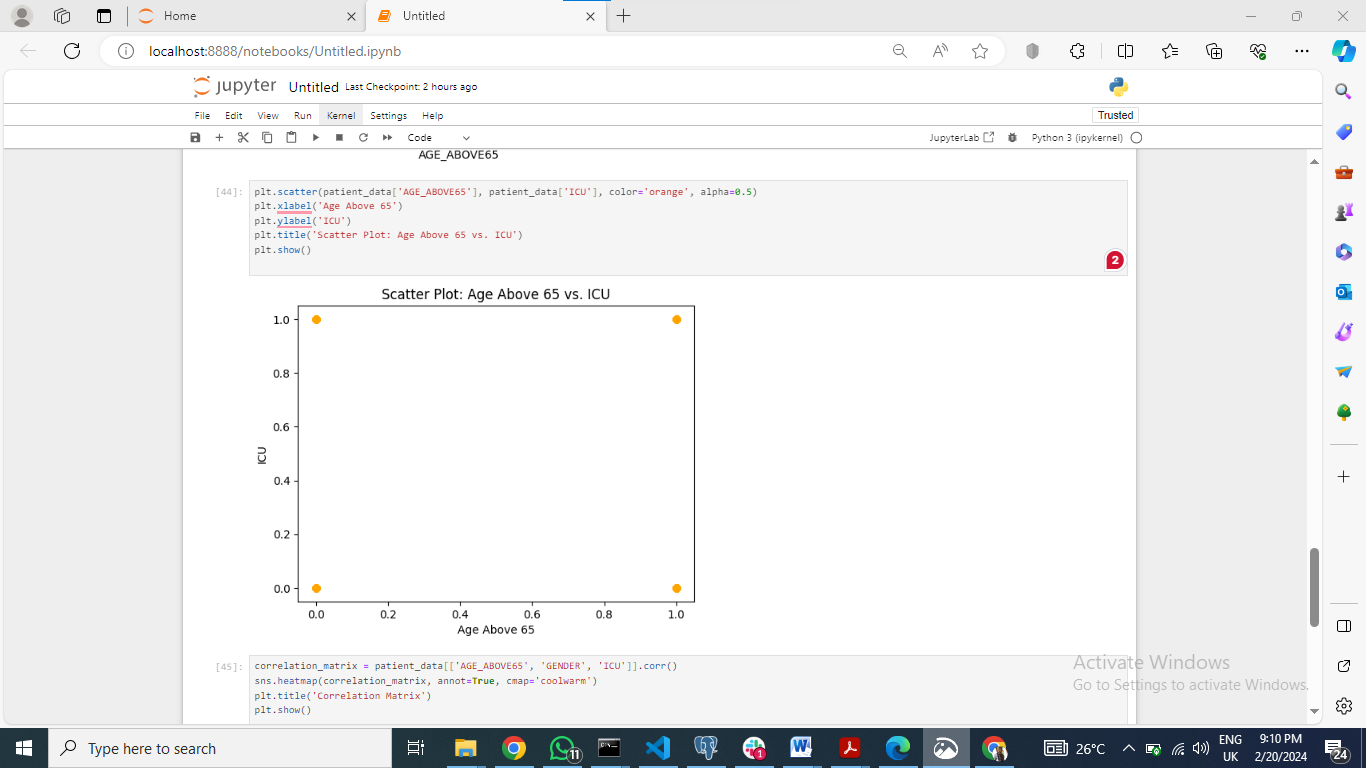
Title: Box Plot for 'AGE\_ABOVE65



Title: Correlation Matrix Heatmap of Age, Gender, and ICU Status



Title: Histogram for 'AGE\_ABOVE65'



Title: Scatter Plot for 'AGE\_ABOVE65' vs. 'ICU'

Briefly, the preprocessing steps consist of dealing with missing data along with outliers, and following that, the problem is defined explicitly as a classification task. Visualization techniques assist in understanding the data and it’s cleaning before moving to model development and usage.

## Model Selection and Implementation

This particular project involves prediction of the necessity of ventilator for a patient who has already been confirmed as COVID-19 case (Aljouie, A.F., Almazroa, A., Bokhari, Y., Alawad, M., Mahmoud, E., Alawad, E., Alsehawi, A., Rashid, M., Alomair, L., Almozaai, S. and Albesher, B., 2021 ). Classifiers match the parameters of particular tasks and structured data; hence they are intended for this particular case.

The model that is developed using machine learning is a combination of logistic regression, random forest and gradient boosting classifiers. The selection of the models has been rigorous. Logistic regression is functionally a baseline model because of its simplicity and interpretability (Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., Wong, T.Y. and Cheng, C.Y., 2020) while random forest and gradient boosting classification are ideal for handling both complex relationships and non-linear patterns in data.

These models are implemented with the scikit-learn package for python. Then, the data is implemented to sort out the missing values, outliers and carry out the uniform scaling via Min-Max Scaler for even representation. Finally, the ordinal or categorical variables are encoded one-hot so as to train the model.

Through the implementation of the techniques such as grid search and random search technique, the hyperparameter tuning is carried out in order to optimize the model performance. The C parameter regulates the strength of regularization in logistic regression. Among random forest and gradient boosting classifiers, estimators’ number, depth, and learning rate parameters are re-optimized.

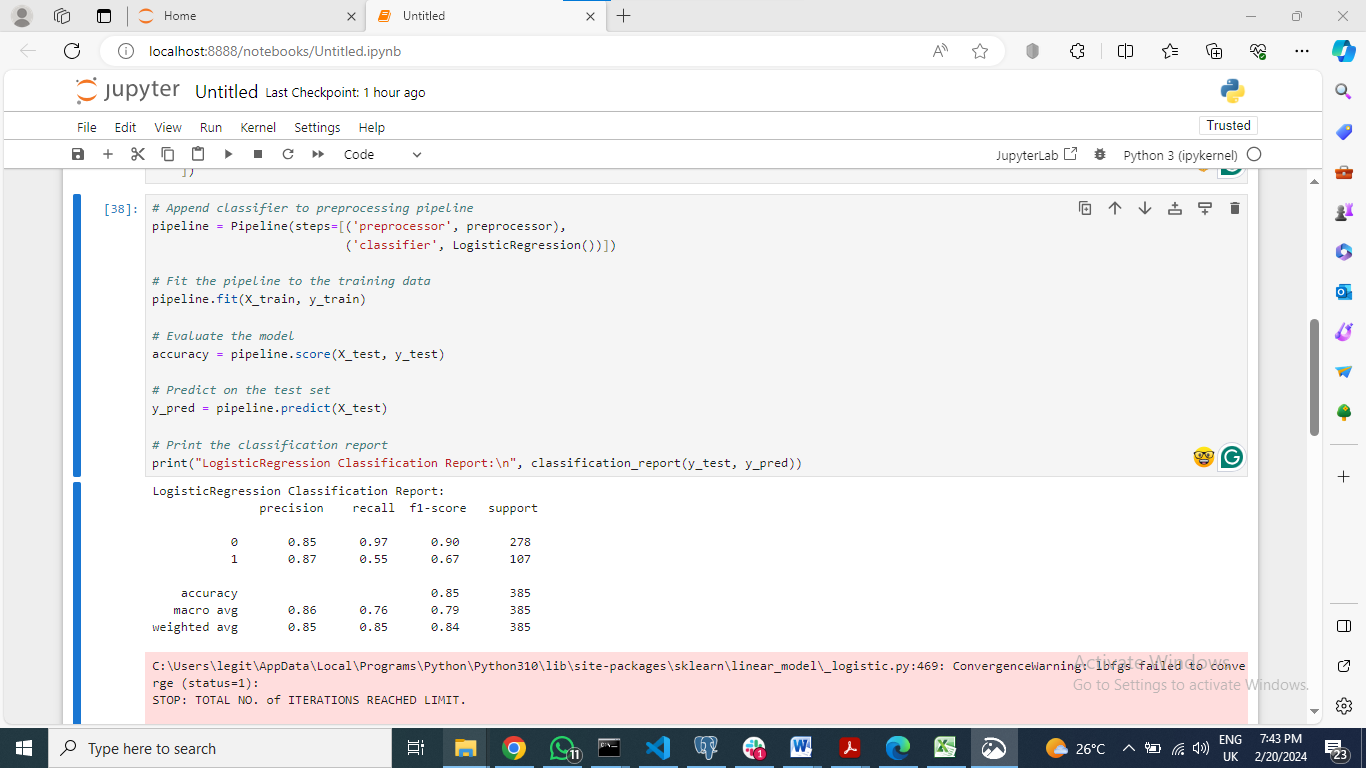
Model performance is assessed using the exact metrics that is reliable at the implementation stage, namely accuracy, precision, recall, and F1-score. The cross-validation techniques are used for the accuracy estimation of the model in unseen data and also to reduce the model overfitting.

Consequently, machine learning models are created with the scikit-learn library in python and the preprocessing step, hyperparameter tuning, and feature selection are given due consideration. A well-grounded classifier will be developed in cooperation with a rigorous and systematic testing that will identify confirmed ICU admissions for COVID-19 cases.

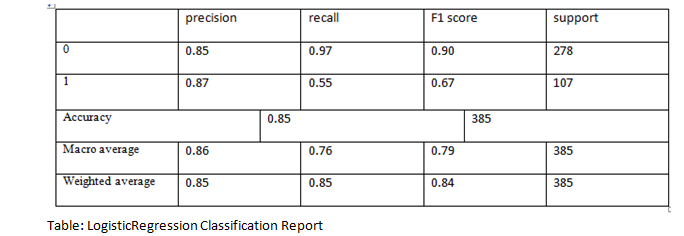
## Model Evaluation

When assessing the machine learning models for ICU admission prediction (Subudhi, S., Verma, A., Patel, A.B., Hardin, C.C., Khandekar, M.J., Lee, H., McEvoy, D., Stylianopoulos, T., Munn, L.L., Dutta, S. and Jain, R.K., 2021), a number of performance metrics such as accuracy, precision, recall and F1-score among others, were taken into consideration.

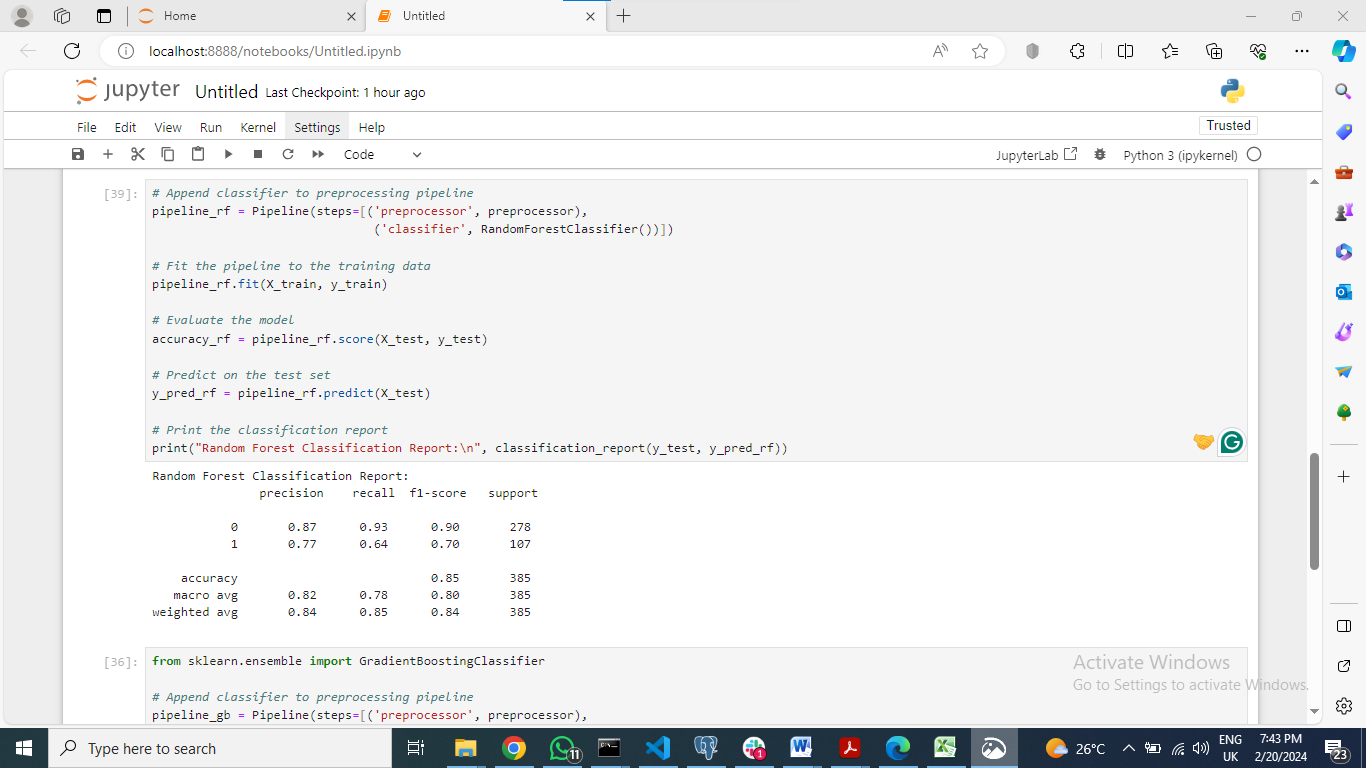
The Logistic Regression model yielded final results with an accuracy of 85% and precision scores of 0.85 for class 0 and 0.87 for class 1. Class 0 had a recall score of 0.97, and class 1 had a recall score of 0.55; as a result, class 0 had an F1 score of 0.90, and class 1 had an F1 score of 0.67. These metrics indicate that the model performed okay, especially in predicting ICU admissions, the class 0 ones.



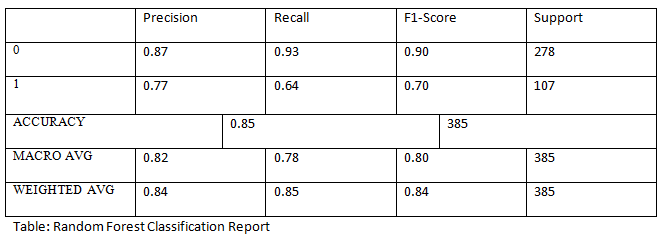
Title: Logistic Regression Model Training and Evaluation with Preprocessing Pipeline



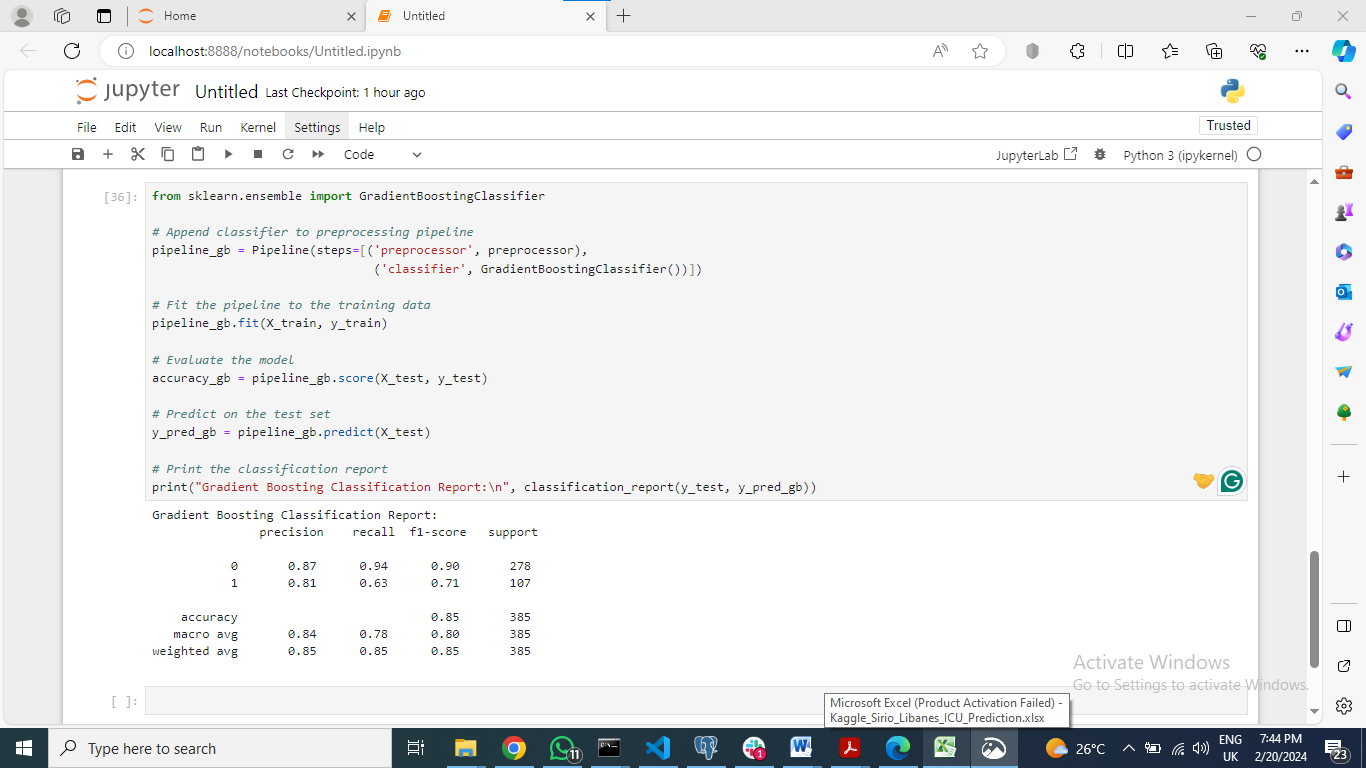
The same, the Random Forest model has an accuracy of 85%. The precision scores for class 0 and class 1 were 0.87 and 0.77, respectively, with the recall scores for class 0 and class 1 being 0.93 and 0.64, respectively. As a result, the accuracy scores are 0.90 for class 0 and 0.70 for class 1. While Random Forest model achieved few more recalls for the class 1 than Logistic Regression, both of them had equivalent overall performance.



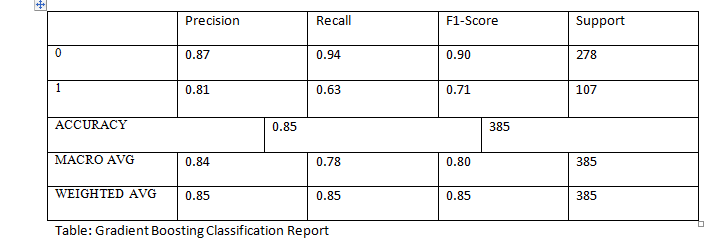
Title: Random Forest Model Training and Evaluation with Preprocessing Pipeline



Another highlight is the model Boosting Gradient, which had 85% accuracy. As for class 0, the precision ratings were 0.87, and the recall scores were 0.94. Besides that, the class 1 precision rating was 0.81, whereas their recall scores were 0.63. In class 0, the F1 scores are at 0.90; for class 1, they are at 0.71. Although it is less effective than the Random Forest model, the Gradient Boosting method still has some advantages. However, it could be more efficient among class 1 according to precision.



Title: Gradient Boosting Model Training and Evaluation with Preprocessing Pipeline



Evaluating the effectiveness of the models, they all had similar accuracy levels, meaning that they all have the capability to predict the admissions of ICU patients correctly. Nevertheless, the decision between the models is likely to depend on special requirements. To illustrate this, the Gradient Boosting model may be preferred if the precision in predicting ICU admissions for class 1 (patients who require ICU admission) is critical as it has better precision score for that particular class 1.

Generally, the prediction accuracy of the models demonstrate a certain degree of usefulness for healthcare systems in identifying patients at risk for ICU admission with flawless resource and patient care management.

## Analysis and Recommendations

### Analysis

1. Model Performance: With accuracy ratings of about 85%, the outcomes of the Gradient Boosting, Random Forest, and Logistic Regression classifiers were comparable. However, there was variation in the F1-scores of Logistic Regression for ICU case prediction; positive cases had the highest F1-score of 0.67.

2. Precision and Recall: The Recall of the model identifies its ability to recover every positive example; precision, in turn, concentrates on how accurate the optimistic predictions made are. All models show that the precision for positive ICU cases is relatively high compared to other models, which means that false positives for ICU cases are few. On the other hand, there were differences in the Recall of positive cases among the logistic regression models, with the highest Recall observed in the logistic regression model at 0.55.

.

### Recommendations

1. Address Class Imbalance: Apply strategies, such as oversampling of the minority class (ICU + patients) or under sampling of the majority class (ICU - patients), for balancing data. Consequentially, the model will make accurate predictions for the positive cases.

2. Feature Engineering: Discover other functions or transforms that may preserve some better measures linked to ICU admissions. This could be done by use of domain-specific knowledge and working toward medical expert identification of the relevant variables.

3. Model Tuning: Hyperparameter optimization of each model will be done as well, to further improve the performance. Grid search or random search methods can be carried out to find out the best set of parameters for each algorithm.

The ICU prediction system would be more robust, accurate, and reliable if these recommendations were implemented. Such a system would improve patient care and contribute to proper resource allocation in healthcare settings.

## Conclusion

After the analysis of the ICU prediction system (Awad, A., Bader-El-Den, M., McNicholas, J., Briggs, J. and El-Sonbaty, Y., 2020) the authors found that machine learning models like Logistic Regression, Random Forest, and Gradient Boosting can reliably predict ICU admissions. These models show acceptable values of accuracy, precision, and recall; however, there is space for progress.

Closing the class imbalance, improving feature engineering methods, fine-tuning model parameters, and exploring ensemble methods contribute to the better performance and predictive abilities of the model. Also, monitoring in the long run, updating, and trying to improve interpretability will interest the system in clinical practice for a long period.

Application of these recommendations allows health institutions to attain more reliable and accurate ICU prediction systems, which aid early interventions, optimizing resources and improving patient outcomes. The development of more accurate and efficient predictive models remains an ongoing process, fueled by the developments of data science, healthcare expertise, and devotion to deliver outstanding patient care.

## References

1. Suryadevara, C.K., 2020. TOWARDS PERSONALIZED HEALTHCARE-AN INTELLIGENT MEDICATION RECOMMENDATION SYSTEM. *IEJRD-International Multidisciplinary Journal*, *5*(9), p.16.
2. Rust, L.O., Gorham, T.J., Bambach, S., Bode, R.S., Maa, T., Hoffman, J.M. and Rust, S.W., 2023. The Deterioration Risk Index: Developing and Piloting a Machine Learning Algorithm to Reduce Pediatric Inpatient Deterioration. *Pediatric Critical Care Medicine*, *24*(4), pp.322-333.
3. Rai, P., Kumar, B.K., Deekshit, V.K., Karunasagar, I. and Karunasagar, I., 2021. Detection technologies and recent developments in the diagnosis of COVID-19 infection. *Applied microbiology and biotechnology*, *105*, pp.441-455.
4. Duong, H.T. and Nguyen-Thi, T.A., 2021. A review: preprocessing techniques and data augmentation for sentiment analysis. *Computational Social Networks*, *8*(1), pp.1-16.
5. Awad, A., Bader-El-Den, M., McNicholas, J., Briggs, J. and El-Sonbaty, Y., 2020. Predicting hospital mortality for intensive care unit patients: time-series analysis. *Health informatics journal*, *26*(2), pp.1043-1059.
6. Subudhi, S., Verma, A., Patel, A.B., Hardin, C.C., Khandekar, M.J., Lee, H., McEvoy, D., Stylianopoulos, T., Munn, L.L., Dutta, S. and Jain, R.K., 2021. Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19. *NPJ digital medicine*, *4*(1), p.87.
7. Nusinovici, S., Tham, Y.C., Yan, M.Y.C., Ting, D.S.W., Li, J., Sabanayagam, C., Wong, T.Y. and Cheng, C.Y., 2020. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of clinical epidemiology*, *122*, pp.56-69.
8. Aljouie, A.F., Almazroa, A., Bokhari, Y., Alawad, M., Mahmoud, E., Alawad, E., Alsehawi, A., Rashid, M., Alomair, L., Almozaai, S. and Albesher, B., 2021. Early prediction of COVID-19 ventilation requirement and mortality from routinely collected baseline chest radiographs, laboratory, and clinical data with machine learning. *Journal of Multidisciplinary Healthcare*, pp.2017-2033.