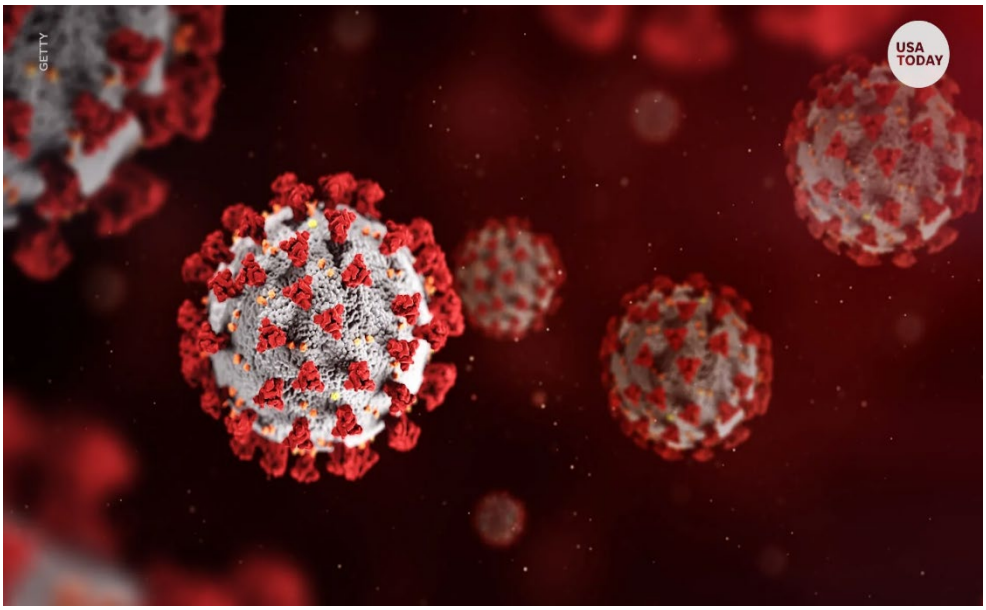
Forecasting ICU Admissions for COVID-19 Patients in Brazil: A Data-Driven Approach



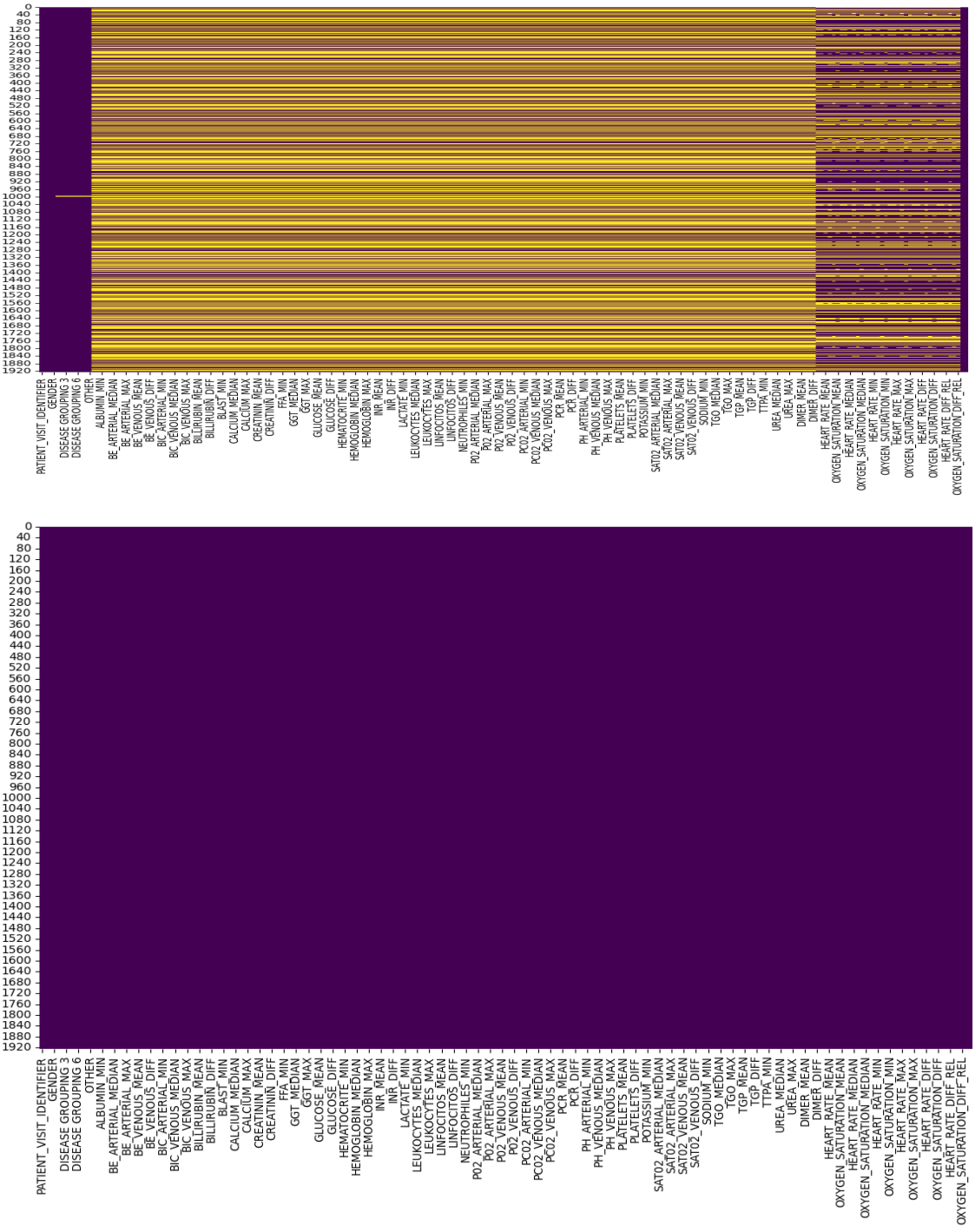
Abstract

The COVID-19 pandemic has strained healthcare systems worldwide, particularly in countries like Brazil with diverse healthcare infrastructure. Efficient ICU bed allocation has become critical. This project aims to use machine learning techniques to predict ICU admissions for COVID-19 patients in Brazil, helping optimize resource management. By analyzing patterns and risk factors associated with ICU admissions, the model can support decision-making and improve healthcare delivery. Data for this study is sourced from a Kaggle dataset provided by Sírio-Libanês Hospital (https://www.kaggle.com/datasets/S%C3%ADrio-Libanes/covid19).

Data Description, Exploration and Processing

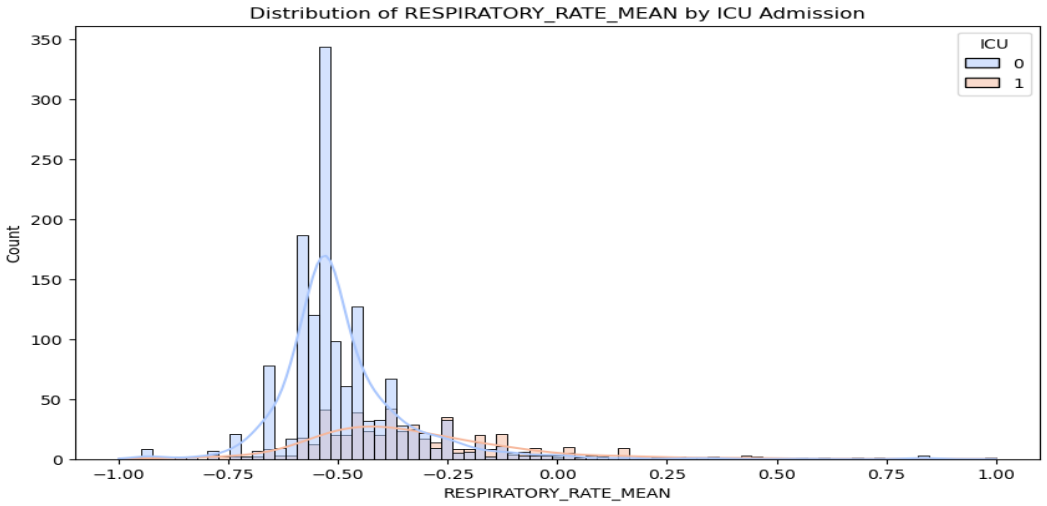
Initial exploration of the dataset showed that the data had been cleaned and scaled to fit between -1 and 1. Also, a lot of missing values were noticed, which needed to be handled carefully to keep the data useful for analysis. Forward-fill and backward-fill techniques were used to fill these gaps. Forward-fill replaces missing values with the last known value, while backward-fill uses the

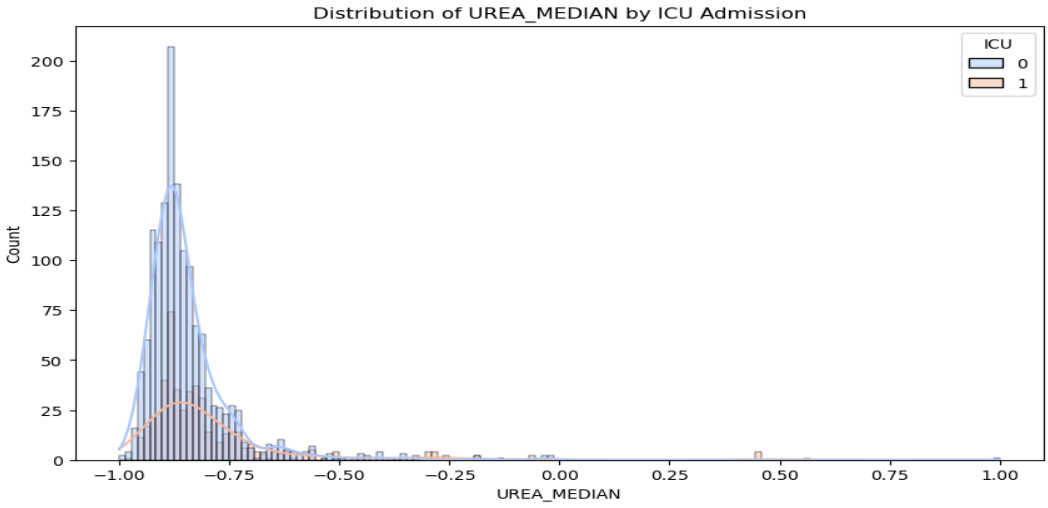
next available value to fill in the gaps. Using both methods helped maintain the flow of data and reduce the impact of missing information.

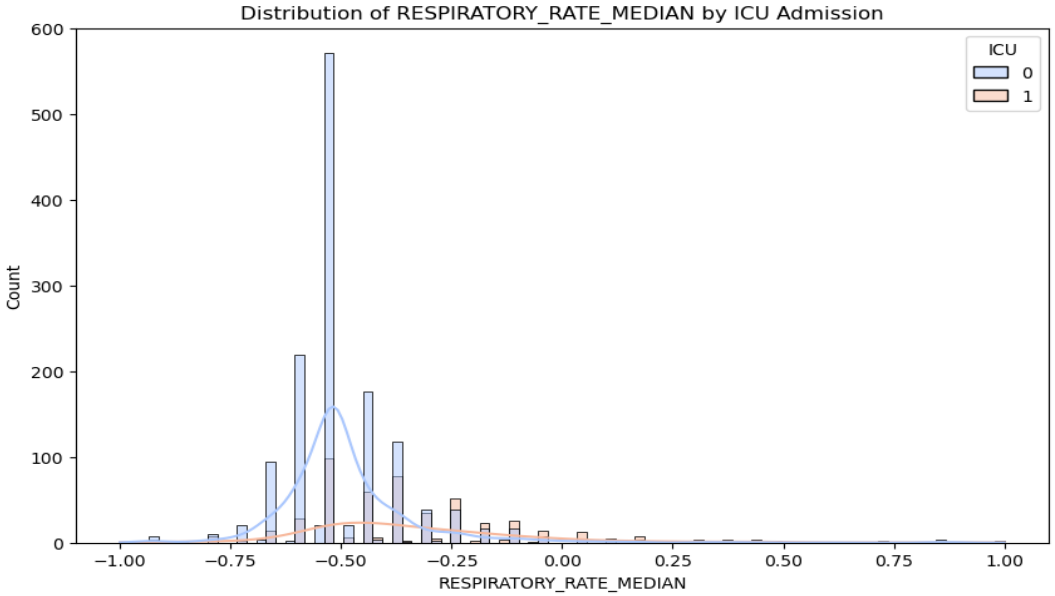


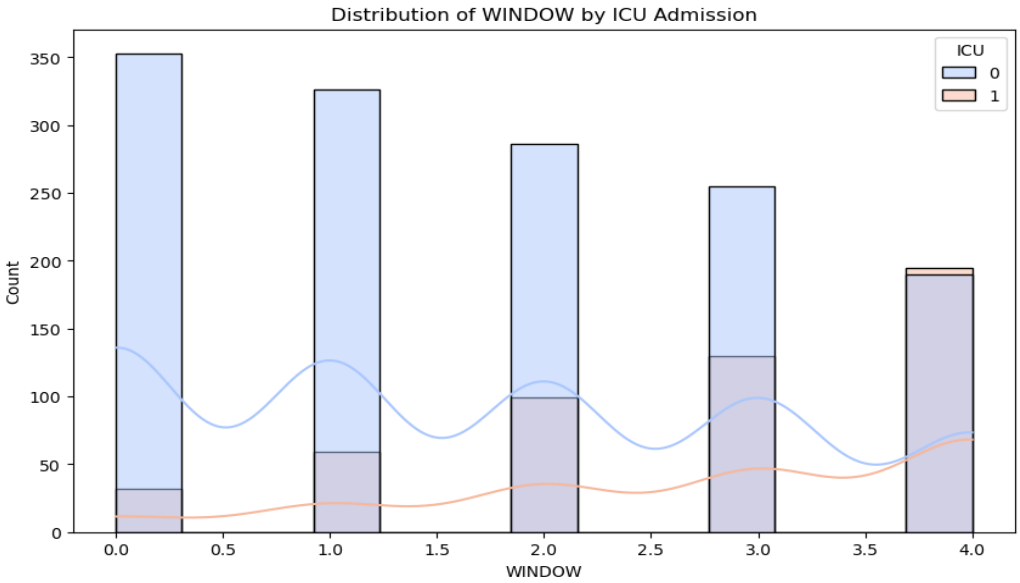
The dataset includes important patient-level details, such as clinical and demographic features. Key features include patient identifiers, age percentiles, and whether the patient was admitted to

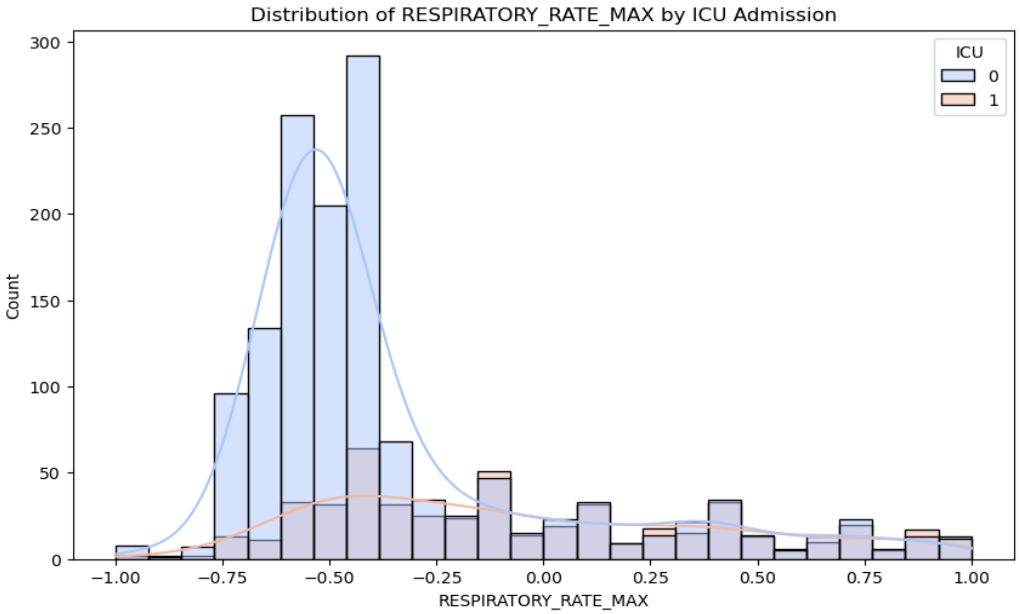
the ICU. Age percentiles show where a patient stands compared to others by age, while ICU admission status is the main outcome of interest. Proper handling of missing values was crucial to preserve the quality of the dataset, and the combination of forward-fill and backward-fill techniques helped achieve this. After handling the missing data, the dataset was further analyzed to understand how ICU admissions varied across different periods and age groups. To do this, various visualizations, such as heat maps and count plots, were used. Heatmaps helped visualize relationships between variables, while count plots showed the number of ICU admissions across different categories. These visual tools make it easier to spot trends and patterns in the data.

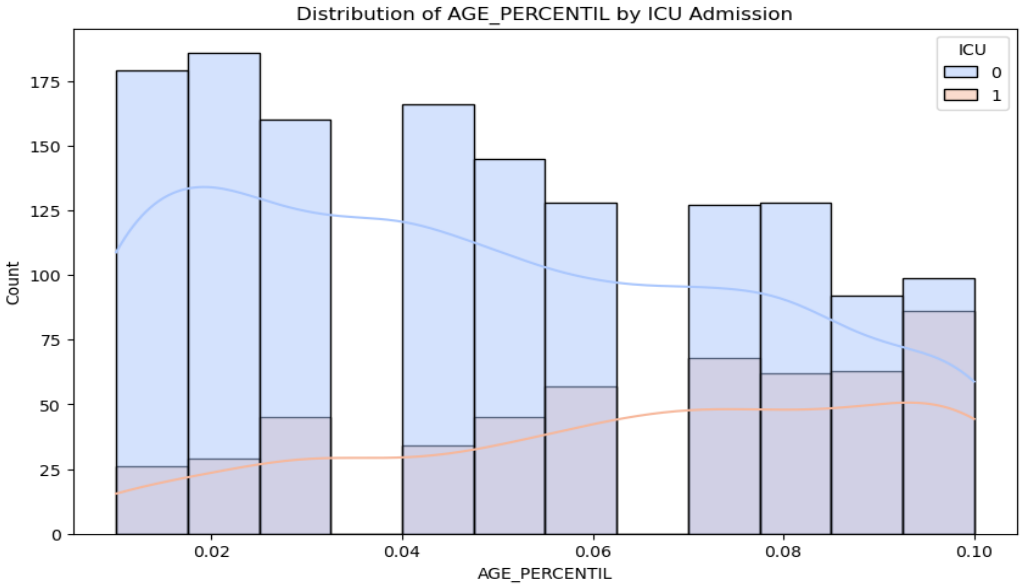





Distribution of RESPIRATORY_RATE_MEDIAN by ICU Admission
ICU
0
1
Count
RESPIRATORY_RATE_MEDIAN
600
500
400
300
200
100
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

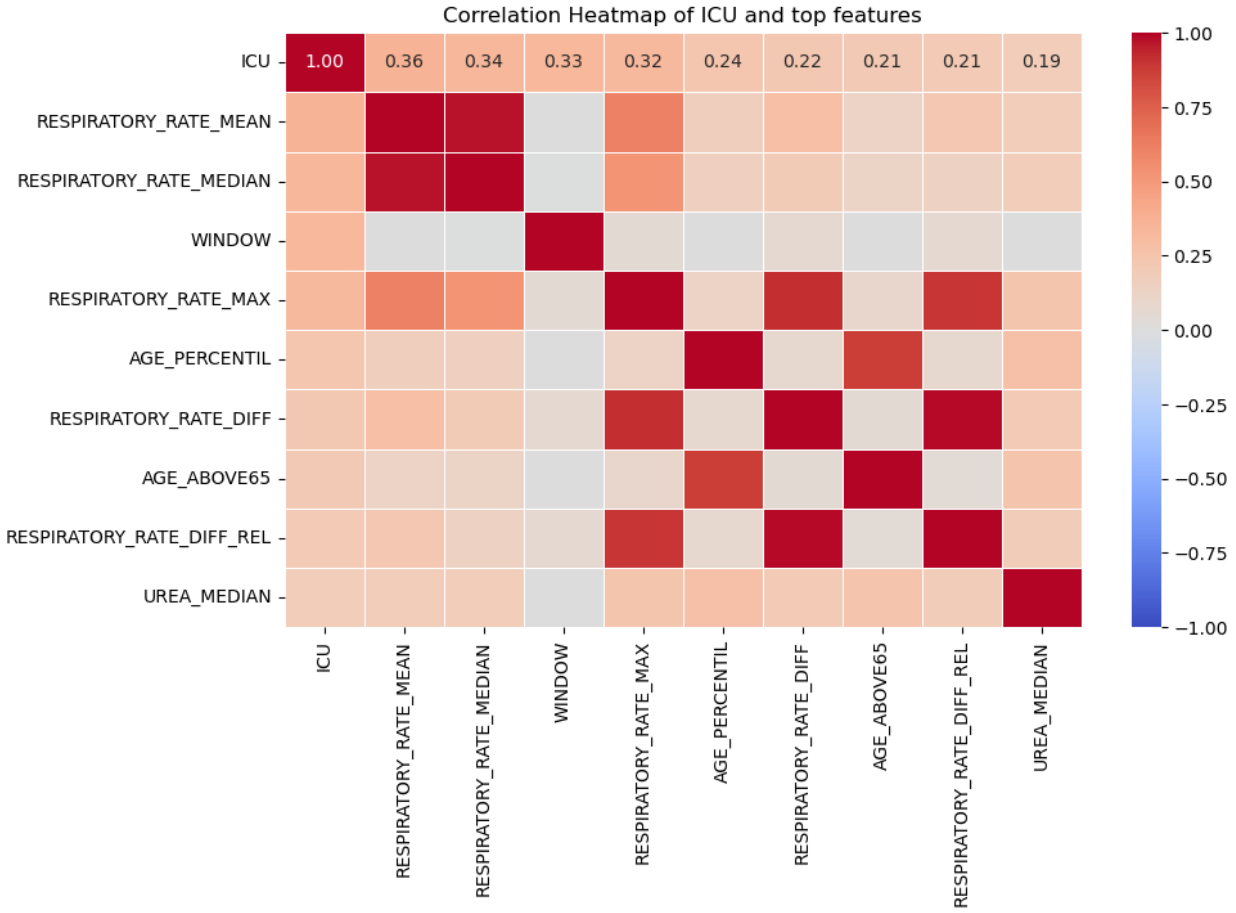

Distribution of WINDOW by ICU Admission
ICU
0
1
Count
WINDOW
350
300
250
200
150
100
50
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0





A

correlation analysis was also performed to identify which features were most related to ICU admissions. This analysis found that age percentiles and specific periods had a significant impact on ICU admissions. To prepare the data for machine learning models, these features were converted into numerical values, a necessary step to ensure the models could process them properly. These preprocessing steps and initial exploration provided a solid foundation for building predictive models.



Machine Learning Models

Three machine learning models were employed to predict ICU admissions: Logistic Regression, Random Forest Classifier, and Decision Tree Classifier. The following sections detail the model training, evaluation, and optimization processes. the window column was dropped before working on the models. This is because there would be a lot of assumptions if it had to be included in the features.

Logistic Regression

Logistic Regression was selected for its interpretability and simplicity. The model was trained using a training set and evaluated on a test set. Performance metrics, including accuracy, precision, recall, and the confusion matrix, were computed to assess the model's effectiveness.

Cross-validation was performed to ensure robustness, and Bayesian Optimization (BayesSearchCV) was used to fine-tune hyperparameters. The best parameters were identified, improving the model's performance.

```
Logistic Regression Model Accuracy:  0.8
              precision    recall  f1-score   support

           0       0.82      0.93      0.87       278
           1       0.71      0.47      0.56       107

    accuracy                           0.80       385
   macro avg       0.77      0.70      0.72       385
weighted avg       0.79      0.80      0.79       385

[[258  20]
 [ 57  50]]

Text(0.5, 1.0, 'Logistic Regression Confusion Matrix')
```

Key Observations¶

Overall **Accuracy**: 80% of predictions were correct.

- Class **0 (Non-ICU patients)**:
  - Precision: 82%
  - Recall: 93%
  - F1-Score: 0.87
- Class **1 (ICU patients)**:
  - Precision: 71%
  - Recall: 47%
  - F1-Score: 0.56
- **Confusion Matrix**:
  - **258** true negatives
  - **20** false positives

- **57** false negatives
- **50** true positives

- **Class Imbalance**: There are significantly more non-ICU cases (278) than ICU cases (107), which may affect the model's ability to predict ICU cases effectively.
- **Cross Validation**: mean accuracy 79.06% with a standard deviation of 0.0334
- **BayesSearch:** After Optimization, accuracy increased to 81%

Random Forest Classifier

The Random Forest Classifier was employed to handle complex interactions between features. With 200 estimators and a maximum depth of 3, the model was trained and evaluated.

```
Random Forest Accuracy (with class weight): 0.812987012987013
              precision    recall  f1-score   support

           0       0.89      0.85      0.87       278
           1       0.65      0.72      0.68       107

    accuracy                           0.81       385
   macro avg       0.77      0.78      0.77       385
weighted avg       0.82      0.81      0.82       385
```

Key observations

Overall accuracy: **81%**.

- **Class 0 (Non-ICU patients)**:
  - Precision: 89%
  - Recall: 85%
  - F1-Score: 0.87
- **Class 1 (ICU patients)**:
  - Precision: 65%
  - Recall: 72%
  - F1-Score: 0.68
- The **macro average** F1-score is **0.77**, indicating balanced performance across both classes.
- The **weighted average** F1-score is **0.82**, reflecting the overall effectiveness of the model.

Decision Tree Classifier

The Decision Tree Classifier was also used to understand the decision-making process and feature importance. A grid search was conducted to find the optimal hyperparameters, including max depth, min samples split, and min samples leaf. The Decision Tree model's performance was assessed and compared with the other models.

Decision tree Accuracy : 0.7532467532467533

	precision	recall	f1-score	support
0	0.82	0.85	0.83	278
1	0.56	0.50	0.53	107
accuracy			0.75	385
macro avg	0.69	0.68	0.68	385
weighted avg	0.75	0.75	0.75	385

```
Parameters: {'criterion': 'gini', 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2}
New accuracy: 0.8
```

Key Observations¶

model accuracy: 75%.

- Class 0 (Non-ICU patients):
  - Precision: 82%
  - Recall: 85%
  - F1-Score: 0.83
- Class 1 (ICU patients):
  - Precision: 56%
  - Recall: 50%
- After hyperparameter tuning, new accuracy: 80%

Results

Logistic Regression achieved the highest accuracy (81%), demonstrating strong performance in identifying non-ICU patients, but struggled with ICU cases (precision: 71%, recall: 47%).

Random Forest had an accuracy of 81% as well, showing good precision for non-ICU patients (89%) and improved recall for ICU patients (85%), though its precision for ICU cases dropped to 65%. The Decision Tree classifier has an accuracy of 75% accuracy, with a precision of 82% and a recall of 85%. After optimization, the decision tree model increased its accuracy to 80%. All models faced challenges due to class imbalance, indicating a need for techniques like SMOTE, feature engineering, and exploration of ensemble methods to enhance predictive performance, particularly for ICU patients**.**

Conclusion

This study demonstrates the potential of machine learning models to predict ICU admissions for COVID-19 patients in Brazil. Healthcare providers can better allocate ICU resources and improve patient outcomes by leveraging data and advanced modeling techniques. Future research should incorporate additional features, such as treatment history, to enhance prediction accuracy and model robustness. It should also include time stamps for the vitals, that way the window column can be better utilized. All the models did well, but they had trouble accurately predicting the minority class, this indicates an imbalance in the data that needs to be addressed. This could be addressed using techniques like SMOTE, feature engineering, and the exploration of ensemble methods to enhance predictive performance, particularly for ICU patients. On the overall performance of all the models, Random Forest outperformed the rest, and it will be the ideal model to use in future predictions.

Author: Ousman Njie