UNIVERSITY OF LEICESTER

SCHOOL OF CMS

Big Data CW Report – Group 09

Module

CO7093 Big Data and Predictive Analytics

Assignment

Coursework

Submitted By

Rajalakshmi Venkatesh Kumar, Shree (SRVK1)
Vaishnavi Late (VL99)
Palanikumar, Sidheswar (SP922)

Contents

Table of Figures

1. Introduction

In course contains the developing a classification model to predict the risk of Covid-19 patients being admitted to the ICU based on their symptoms, medical history, and other relevant features. The dataset provided contains detailed patient information, with a focus on understanding the various factors that influence ICU admission. Our objective is to build a predictive model and to explore the data through cleaning, visualization, and analysis techniques that will enhance the model's performance.

1.1 Dataset Overview:

The dataset used in this coursework consist of approximately 200031 patient records related to COVID-19. It includes 21 features with demographic information (such as age, sex and patient_type), reported conditions like diabetes, hypertension, asthma and obesity and a binary target variable indicating ICU admission status (ICU: 2 for no ICU admission, 1 for ICU admission). Initial exploratory checks revealed several data quality issues, including the presence of missing values and inconsistent entries—with the use of '?' for unknowns.

2. Part 1: Building up a basic predictive model

2.1 Data Cleaning and Transformation :

2.1.1 Handling missing values :

For Data Cleaning and Transformation, we performed a multi stage cleaning process using pandas. There are lots of inconsistencies and missing values in dataset. Many features contain "?" for missing values. These are first replaced with NaN, then we analysed each column's percentage of missing values. Additionally, rows with excessive missing values across multiple columns are dropped.
The columns "index" and "USMER" are removed from the dataset because it does not contribute to analysis and not contain any meaningful information.

2.1.2 Handling Outliers using IQR method :

We have identified potential outliers on the "age" column, which is critical in predicting ICU admissions.
To identify and remove outliers, we used the IQR (Interquartile Range) method which is the difference between the 75th percentile (Q3) and the 25th percentile (Q1) of a dataset: IQR=Q3−Q1. data point is considered an outlier if it lies below:
(Lower Bound = Q1 − 1.5 × IQR) or above (Upper Bound = Q3 + 1.5 × IQR)
We applied this method to all numeric columns (e.g., age, classification_final, etc.). For each column, we computed the IQR and removed all records that fell outside the computed lower and upper bounds. In the dataset, age values below 0 or above 120 were clearly erroneous and removing such outliers resulted in a cleaner dataset with improved model performance.

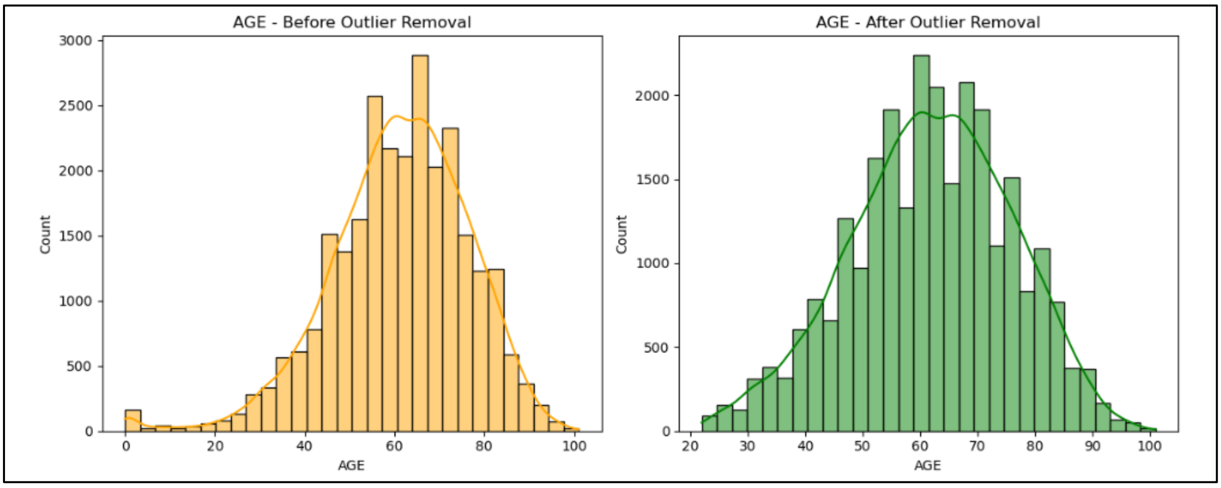


Figure 1: Outlier Removal

2.2 Data Visualization:

In this stage of the project, we carried out a detailed exploratory data analysis (EDA) using various visual techniques to understand the dataset's structure and identify key patterns. We examining the distribution of the target variable (ICU and CLASSIFICATION_FINAL) to detect any imbalance, which helped to decide whether to apply data balancing methods. ICU admissions are analysed across different age groups to highlight potential high-risk. To assess prediction accuracy, we compared model outputs with existing classification labels, such as CLASSIFICATION_FINAL. Correlation heatmaps and scatter matrix plots are used to explore relationships between variables. Additionally, we used histograms, boxplots, and feature importance graphs to study feature distributions, uncover outliers, and identify significant predictors. These findings played an essential role in guiding our data cleaning, feature selection, and modelling approach.

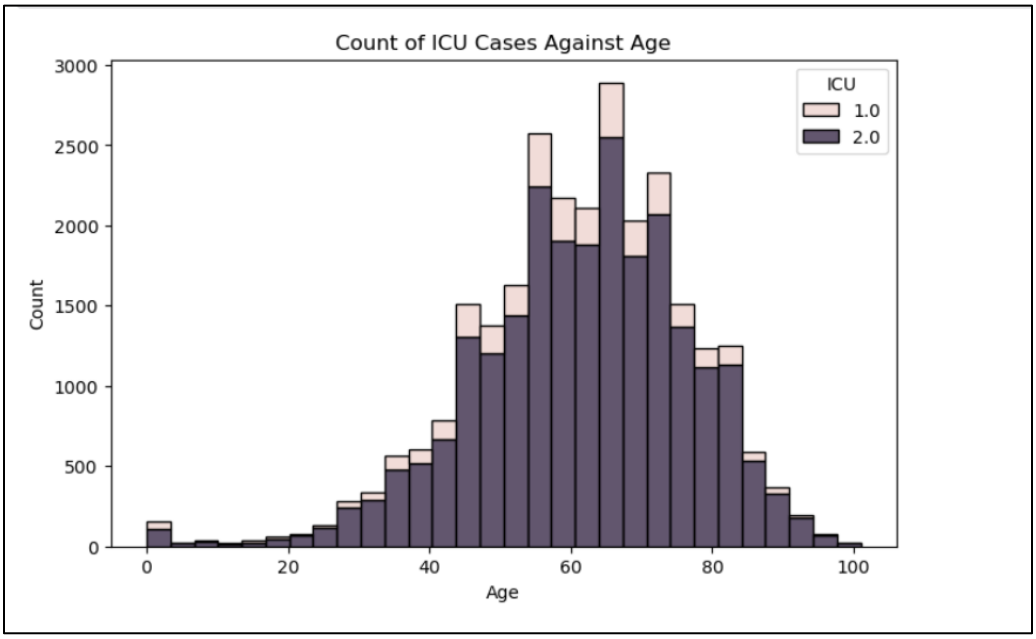


Figure 2: Count of ICU Cases against age

This histogram visualizes the distribution of ICU cases across different age groups.
ICU = 1.0: Patients who were admitted to the ICU (shown in a lighter color).
ICU = 2.0: Patients who were not admitted to the ICU (shown in a darker color).

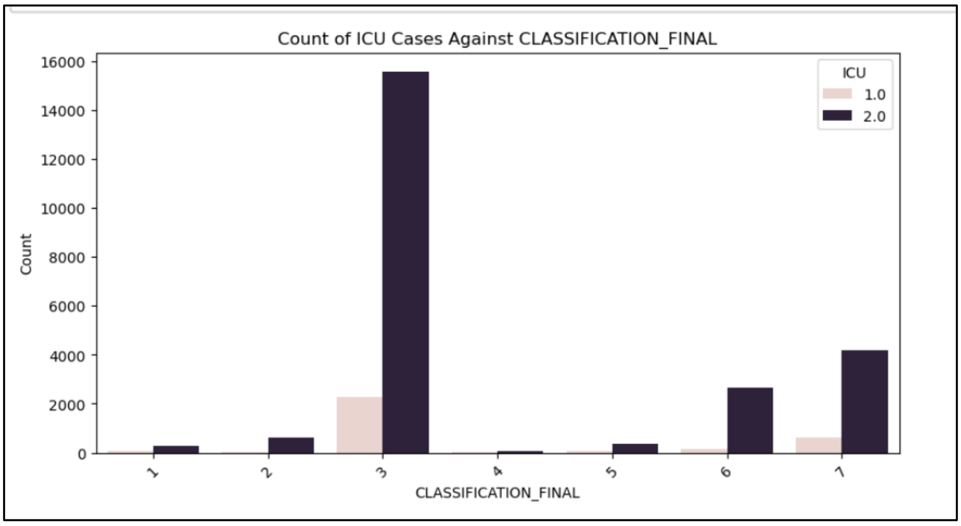


Figure 3: Count of ICU cases against CLASSIFICATION_FINAL

When visualized against ICU outcomes, we observed that patients with higher severity classification were far more likely to require ICU care.

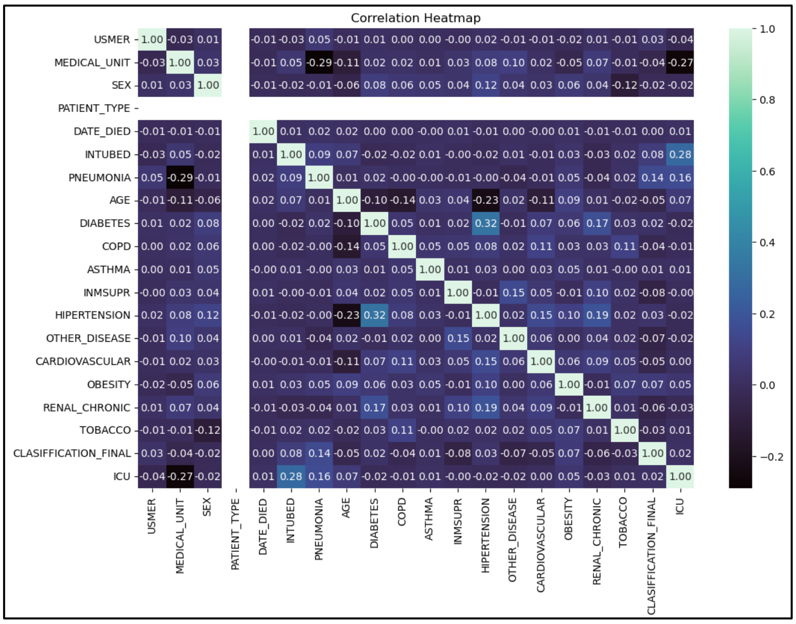


Figure 4: Correlation Heatmap

All features were binary; hence, correlation values are generally low. The strongest correlation (0.32) was between DIABETES and HIPERTENSION, indicating a mild co-occurrence. All other feature pairs had correlations close to zero, suggesting little to no linear association. By examining the scatter matrix, correlation matrix, and heatmap, we can clearly identify pairs of features with high correlation (|correlation| ≥ 0.3). Below figure indicates the highlighted high correlation heatmap.

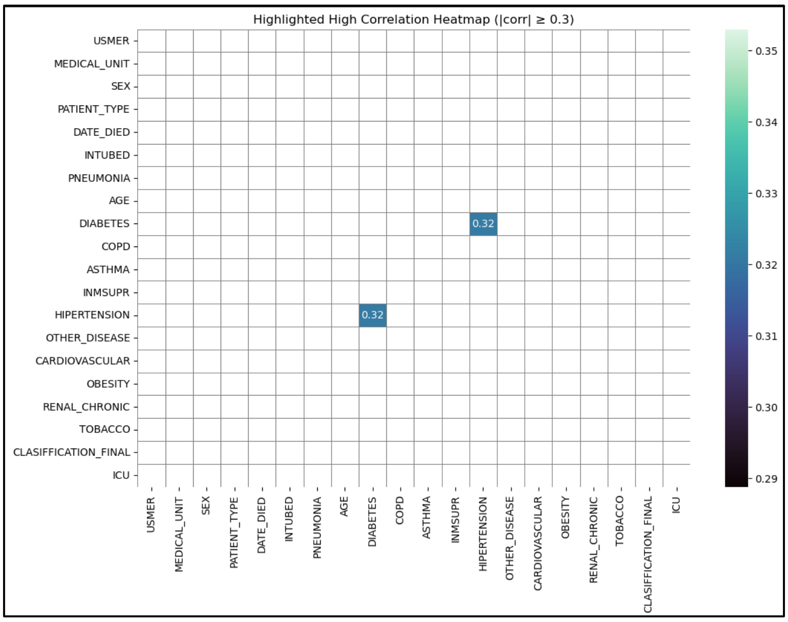


Figure 5: Highlighted high correlation heatmap

The visualizations were created to demonstrate a deeper understanding of the problem and the dataset, focusing on identifying key factors affecting the ICU admission and death probability of patients. Box plots were used to show the distribution of the AGE feature with respect to ICU admission and the final classification of patients. Scatter plots were generated to explore the relationship between the CLASIFFICATION_FINAL (patient severity classification) and AGE, with the hue representing ICU admission status. The countplots show the effect of comorbidities like Obesity and Diabetes on survival, clearly indicating higher death rates for patients with these conditions. Overall, these additional visualizations provided clear insights into how critical features like AGE, comorbidities, and patient classification influence the severity of cases, the need for ICU care, and the overall impact on the healthcare system.

2.3 Model Building:

2.3.1 Select the predictors that would have impact in predicting ICU :

Based on the exploratory data analysis and the generated visualizations, the following predictors were identified as having a significant impact in predicting ICU admission:

I. AGE :
- The boxplots and scatter plots showed that older patients are more likely to be admitted to ICU.
- As age increases, the risk of severe complications also increases, making it a crucial factor in ICU prediction.

II. CLASIFFICATION_FINAL :
- Scatter plots indicated that the patients with higher severity levels are more frequently admitted to ICU.

The most important predictors impacting ICU admission are *AGE*, *CLASIFFICATION_FINAL* and *Comorbidities* like Diabetes, Obesity, COPD, Cardiovascular disease, Chronic Renal, and Asthma. These predictors directly relate to a patient's risk level and ICU admissions.

2.3.2 Evaluate first linear model :

For predicting ICU, we have selected Age and Classification_Final predictors based on exploratory data analysis. As ICU prediction is a binary classification problem (ICU = Yes or No), we used the

Logistic Regression to build the first linear model. We selected the features as,
X = ['AGE', 'CLASIFFICATION_FINAL'] and Y = ICU Status ({2.0: 1, 1.0: 0}). The dataset is split into 80% training and 20% testing sets. Cross-validation procedure is used to enhance the model performance and stability across different data splits. The model achieved a **training accuracy of 88.1%** and a **test accuracy of 87.9%**, which shows a consistent performance without signs of overfitting. However, the classification report reveals a **significant class imbalance**, with the model heavily biased toward predicting class 1. It achieved a **perfect recall of 1.00** for class 1 but **completely failed to identify class 0**, with a recall and precision of 0.00. As a result, the F1-score: 0.47, and the **ROC-AUC score is only 0.561** which is less.

2.3.3 Evaluate the performance of model using different performance metrics :

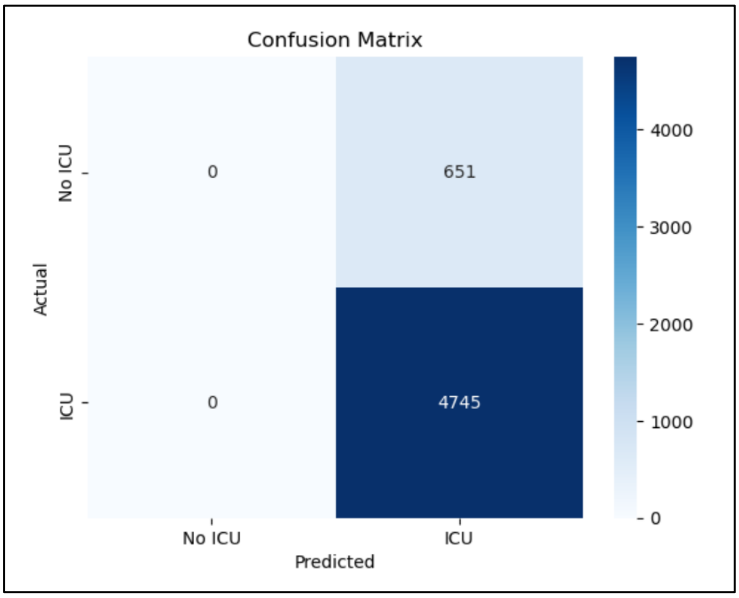


Figure 6: Confusion Matrix

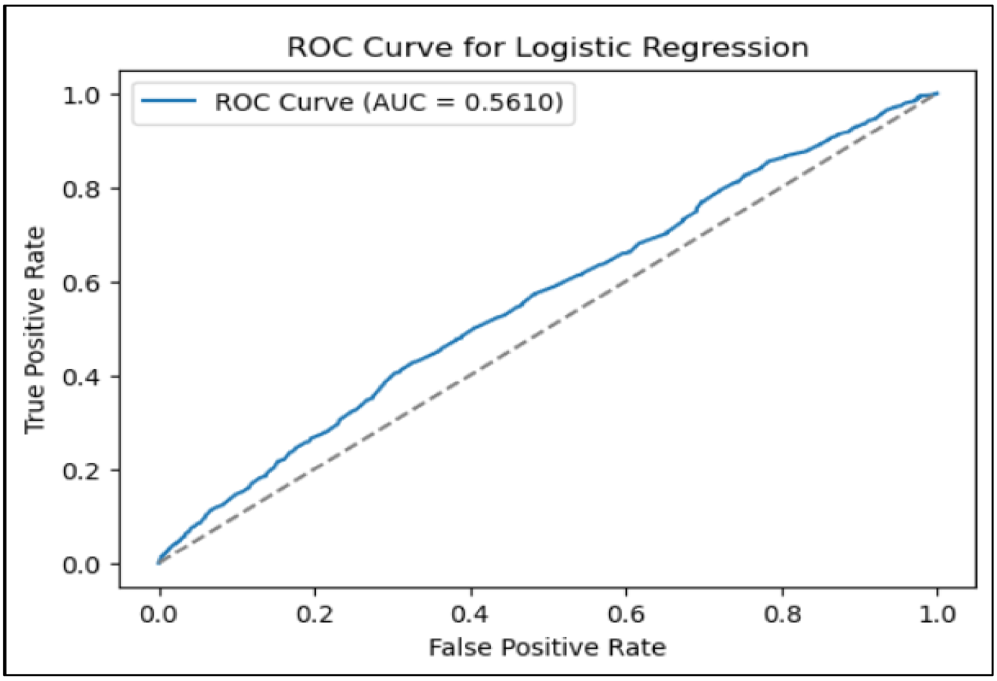


Figure 7: ROC Curve for Logistic Regression

The confusion matrix and ROC curve further confirm the model's bias toward predicting ICU cases, with **no true negatives** captured. The **ROC-AUC score of 0.561** indicates poor class separation. Although cross-validation showed **consistent accuracy (~88%)**, the model lacks the ability to correctly classify the minority class, suggesting the need for balancing techniques or model adjustments.

2.3.4 Linear model with balanced data evaluation :

To address the class imbalance in the dataset, an **under sampling technique** was applied to equalize the number of ICU and non-ICU cases. After retraining the logistic regression model on the balanced data, the classification performance improved significantly for the minority class. The **recall for ICU cases increased from 0.21 to 0.70**, and the model achieved a better **balance between precision and recall** for both classes. Although the **ROC-AUC score remained similar** before (0.8156) and after balancing (0.8135), the balanced model demonstrated **more equitable predictive power**, particularly by avoiding strong bias toward the majority class. These results confirm that balancing the dataset led to a more robust and fairer model, particularly in addressing the underrepresented ICU class.

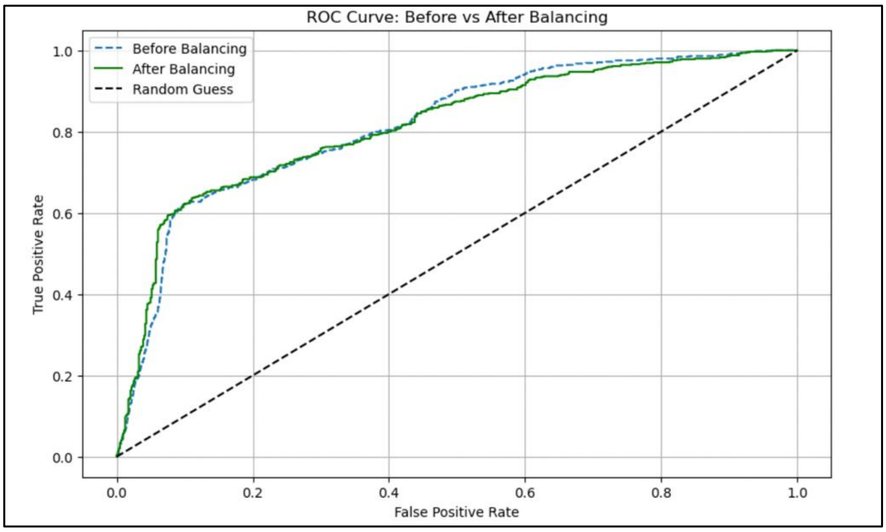


Figure 8: ROC Curve: Before vs After Balancing

3. Part 2 : Improved Model

3.1 Handling Missing Values with Imputation and feature engineering

Missing values in the dataset were handled using PySpark's Imputer, which replaces nulls with the **mean of each column**. This approach preserves the full dataset size and ensures that important statistical relationships are maintained.This strategy complies with the coursework requirement to use **maximum data points** instead of discarding incomplete records. A set of clinically relevant features (e.g., age, comorbidities, treatment indicators) was selected and transformed into a **single feature vector** using VectorAssembler. This step is required to feed structured input ito Spark ML algorithms efficiently.

Algorithms used:

- **Random Forest** was selected for its **robustness, scalability, and strong empirical performance** on medical classification tasks. A RandomForestClassifier was trained with 50 trees and a max depth of 10. This model was chosen for its ability to handle non-linear data and capture complex interactions among features.
- A **LogisticRegression** model served as a linear baseline. Its straightforward training and interpretability make it suitable for benchmarking more advanced algorithms.Though it offered decent results, it was outperformed by the Random Forest model in all metrics.
- Both models benefit from being trained on a **balanced dataset**, improving their ability to detect rare but critical cases.

3.2 Evaluation metrics and results

Both models were evaluated using:

- **Accuracy** – overall correctness
- **F1 Score** – balance of precision and recall
- **ROC AUC** – ability to distinguish between classes

Model	**Accuracy**	**F1 Score**	**ROC AUC**
Random Forest	**0.772**	**0.771**	**0.855**
Logistic Regression	0.725	0.725	0.811

Random Forest clearly outperformed Logistic Regression in all metrics, validating the effectiveness of ensemble methods for this clinical classification task.

3.3 Risk-Based Patient Segmentation Using K-Means Clustering and PCA

K-Means clustering was applied to the scaled dataset using k=2, with features first standardized using StandardScaler. This method groups patients into risk clusters based on clinical attributes.

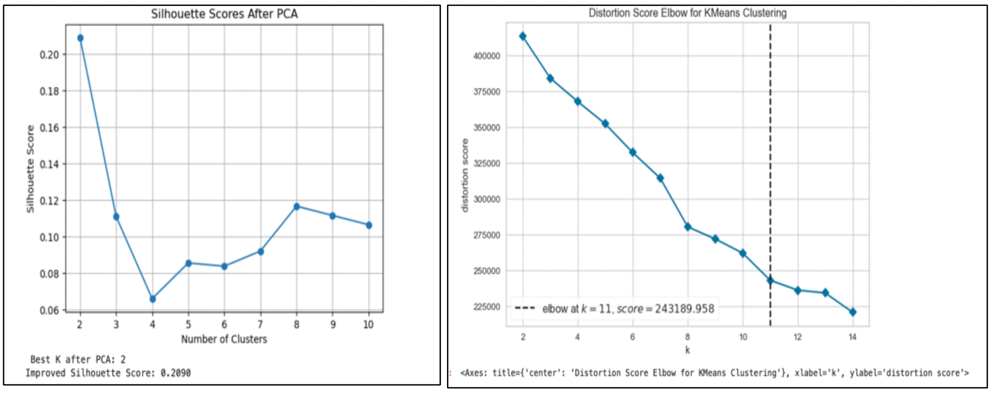
Silhouette Score was used to evaluate how well-separated the clusters are.

Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving 95% variance. This step enhances clustering performance and allows for effective 2D visualization of high-dimensional data.

The clustering quality improved post-PCA, increasing the silhouette score from **0.2292 to 0.2953**.

The best number of clusters (k) was explored using:

- **Silhouette Scores** over k=2–10
- **Elbow Method** with Yellowbrick



```
Best K after PCA: 2
Improved Silhouette Score: 0.2090
```

```
<Axes: title={'center': 'Distortion Score Elbow for KMeans Clustering'}, xlabel='k', ylabel='distortion score'>
```

Figure 9: Silhouette Scores for PCA

This multi-metric approach justifies the final choice of k and shows that clusters are well-formed based on internal metrics.

Clusters were visualized in the PCA-reduced 2D space, allowing easy interpretation of cluster boundaries and overlap.

3.4 Cluster visualization and Interpretation

The code visualizes the K-Means clustering results after PCA dimensionality reduction. The pca_features column is split into two components (x and y), representing each patient's location in a 2D space. A scatter plot is generated using Matplotlib, where each point is colored based on its cluster assignment (cluster_after). This helps evaluate how well-separated the clusters are after reducing the high-dimensional data. The visualization confirms the effectiveness of PCA in enhancing cluster interpretability.

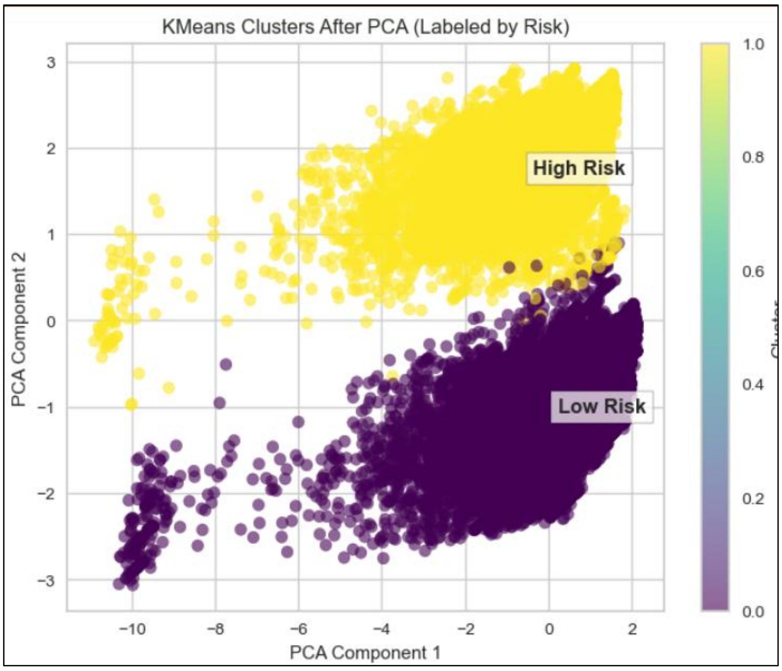


Figure 10: KMeans Clusters

Cluster 1 (High Risk Cluster)

- **Gender:** Exclusively female patients (SEX = 1.0).
- **Pregnancy:** High pregnancy rate (PREGNANT ≈ 1.98), further confirming the female identity of this cluster.
- **Age:** Slightly younger patients (AGE ≈ 52.2 vs. 52.6).
- **Clinical Classification:** Higher average value in CLASIFFICATION_FINAL (≈ 4.63), suggesting more severe cases or higher clinical attention**.**
- **Respiratory Conditions:**

  Slightly higher prevalence of PNEUMONIA and INTUBED cases, which might point to more severe respiratory symptoms.
- **Comorbidities: Lower average values in key risk-related conditions such as:**

  DIABETES, OBESITY, HIPERTENSION, and RENAL_CHRONIC, suggesting better underlying health conditions.

Cluster 0 (Low Risk Cluster)

- **Gender:** Predominantly male patients (SEX ≈ 1.99).
- **Pregnancy:** Not pregnant (PREGNANT = 1.0), as expected for a male-dominant cluster.
- **Age:** Slightly older (AGE ≈ 52.6).
- **Comorbidities:** Slightly higher prevalence of risk conditions like:

  DIABETES, OBESITY, HIPERTENSION, and RENAL_CHRONIC.
- **Clinical Classification:** Slightly lower average in CLASIFFICATION_FINAL (≈ 4.37), suggesting less severe or earlier-stage cases.
- **Medical Facility Usage:** Slightly higher averages in USMER and MEDICAL_UNIT, which might reflect different care-seeking behavior or referral patterns**.**