Predictive and Analytical Modeling for Health Risk Assessment Using Vital Signs

By

SHRAVANI VEMULA

AJAY KATTA

SHIVA SIDDARTH YEDLA

Under the Guidance of

Dr. Nikanor Volkov

MS Business Analytics Program, Associate professor

A capstone project presented in partial fulfillment of the requirements for the Master of Science in Business Analytics program at Mercer University's Stetson-Hatcher School of Business.

CONTENTS *Pg No.*

1.Introduction

In hospitals, doctors and nurses often need to make fast decisions about which patients need help first. One big challenge is knowing early on if a patient's condition is about to get worse. If they can catch these signs quickly, they can treat the patient sooner and possibly save their life.

In this project, we used machine learning (a type of computer model) to study patient vital signs like heart rate, blood pressure, and oxygen levels. Our goal was to build a system that can automatically tell if a patient is at low, medium, high, or extreme risk—so hospitals can decide faster who needs urgent care.

Motivation

Vital signs are recorded for every patient, but they're not always used to their full potential. Sometimes, signs that a patient is getting worse aren't obvious until it's too late. We wanted to use the data hospitals already collect to help doctors and nurses act faster.

This project was inspired by the idea that we could build something useful and practical—a tool that gives quick insights using just the numbers already measured in hospitals. If successful, this could:

- Help hospitals manage patient care better
- Support staff during busy times
- Improve how fast serious cases are handled

Background

Vital signs—like heart rate, temperature, oxygen level, and blood pressure—are key indicators of a person's health. These are easy to measure and are already being tracked in most hospitals. But usually, they're just checked against simple cutoffs (like "heart rate above 100 = high").

With new computer tools, we can do much more. Using machine learning, we can study thousands of cases and learn patterns that aren't easy to spot with the human eye. In this project, we used these tools to group patients into risk levels. We tested different models, including:

- Logistic Regression (simple model)
- Random Forest and XGBoost (advanced models)
- K-Means Clustering (to group patients without using labels)

Business Question

How can hospitals use real-time vital signs to automatically identify high-risk patients and support faster, more accurate triage decisions?

Project Goal

The goal of this project is to develop a machine learning-based system that analyzes patient vital signs and classifies individuals into Low, Medium, High, or Extreme risk categories. By doing this, we aim to:

- Help hospitals prioritize patients more effectively
- Enable early detection of critical conditions
- Support real-time decision-making in emergency and critical care settings

2.LITERATURE REVIEW

1. A Systematic Review of Machine Learning Models for Clinical Deterioration

https://pubmed.ncbi.nlm.nih.gov/37156168/

This article reviews multiple research studies that use machine learning models to predict when patients are likely to experience clinical deterioration in hospitals. It evaluates models such as logistic regression, decision trees, and deep learning, and discusses how vital signs are used as key inputs.
It also highlights common issues faced in this area, such as:

- Imbalanced datasets
- Poor generalizability across hospitals
- Lack of clinical interpretability

Why it's important:
It provides a broad understanding of how ML is being applied in real-world healthcare settings and outlines the gaps that still exist in making these models more useful and reliable in practice.

2.Real-Time Risk Prediction Using Continuous Vital Signs

https://www.sciencedirect.com/science/article/pii/S0010482522003511

This research focuses on using live, continuously collected vital signs (such as from ICU monitors) to predict serious events like cardiac arrest or ICU transfer. It uses machine learning models that update predictions in real time.

Model inputs include:

- Pulse
- Respiratory rate

- Temperature
- Oxygen saturation trends over time

Why it's important:
It proves that real-time ML systems can provide faster and more accurate alerts than periodic checks, which could give doctors extra time to act.

3.Real-Time Alerting Using Electronic Medical Records

https://www.nature.com/articles/s41598-022-15877-1

This article shows how a machine learning model was successfully embedded into an electronic medical records (EMR) system to issue real-time alerts. It uses patient history, vital signs, and lab results to continuously monitor patients and predict worsening conditions.

Main contributions:

- Demonstrates real-world deployment of ML in hospitals
- Shows positive impact on early intervention and clinical workflow
- Validates the integration of predictive tools into healthcare infrastructure

Why it's important:
This article supports the feasibility and practical impact of your project's goal—to create a tool that can be deployed in hospitals for real-time triage and risk prediction.

3.DATA SOURCE AND DATA CHARECTRISTICS:

The dataset used for this project was sourced from Kaggle
https://www.kaggle.com/datasets/nasirayub2/human-vital-sign-dataset

Our dataset contains 97,000 patient records with 18 vital sign attributes. It was sourced from Kaggle and includes key health indicators such as heart rate, respiratory rate, body temperature,

oxygen saturation, and blood pressure (both systolic and diastolic). Additional derived features like Body Mass Index (BMI) and Mean Arterial Pressure (MAP) are also included. This dataset provides a strong foundation for building machine learning models to assess patient risk levels based on their vital signs.

Data Dictionary:

Variable Name	Description
Patient_ID	Unique identifier for each patient.
Heart_Rate	Number of heartbeats per minute (bpm).
Respiratory_Rate	Number of breaths per minute.
Timestamp	Recorded time of vital measurement.
Body_Temperature	Body temperature measured in degrees Celsius.
Oxygen_Saturation	Percentage of oxygen-saturated hemoglobin relative to total hemoglobin.
Systolic_Blood_Pressure	Maximum arterial pressure during heart contraction (mmHg).
Diastolic_Blood_Pressure	Minimum arterial pressure between heartbeats (mmHg).
Age	Age of the patient in years.
Gender	Biological sex of the patient (Male/Female).
Weight_.kg.	Body weight of the patient in kilograms (kg).
Height_.m.	Height of the patient in meters (m).
Derived_HRV	Derived Heart Rate Variability measure.
Derived_Pulse_Pressure	Difference between systolic and diastolic blood pressure (mmHg).
Derived_BMI	Body Mass Index calculated as weight/(height^2).
Derived_MAP	Mean Arterial Pressure, calculated average blood pressure.
Risk_Category	Predicted health risk category (e.g., Low, Medium, High, Extreme).

Data Characteristics:

Variable Name	Data Type
Patient_ID	Categorical (ID/String)
Timestamp	Date/Time
Heart_Rate	Numeric (Integer)
Respiratory_Rate	Numeric (Integer)
Body_Temperature	Numeric (Float)
Oxygen_Saturation	Numeric (Float)
Systolic_Blood_Pressure	Numeric (Integer)
Diastolic_Blood_Pressure	Numeric (Integer)
Age	Numeric (Integer)
Gender	Categorical (Male/Female)
Weight_.kg.	Numeric (Float)
Height_.m.	Numeric (Float)
Derived_HRV	Numeric (Float)
Derived_Pulse_Pressure	Numeric (Float)
Derived_BMI	Numeric (Float)
Derived_MAP	Numeric (Float)
Risk_Category	Categorical (Low, Medium, High, Extreme)

Data Cleaning Process

Before using the data for analysis and machine learning, we cleaned it to make sure it was accurate and ready to use. Here's what we did:

1. Handling Missing Values

We first checked the dataset to see if any important values were missing. If a row was missing vital information such as heart rate, oxygen saturation, or blood pressure, we removed that row to make sure the analysis stayed accurate. For calculated fields like BMI and MAP, we used the

available height and weight data to recalculate them if needed. This helped us keep the dataset complete and reliable.

---

2. Removing Outliers and Incorrect Data

we looked for values that didn't make sense or were far outside the normal range. For example, we removed entries where the body temperature was unusually high (above 45°C) or where the oxygen level was over 100%, which is not medically possible. Removing these extreme or incorrect values helped prevent them from affecting the model's predictions.

---

3. Standardizing Units and Formats

We made sure all measurements were in the same units so that the data would be consistent. Temperature values were checked to be in Celsius, weight in kilograms, and height in meters. We also cleaned the date and time column so all timestamps were in the same format, which is important if we wanted to do any time-based analysis.

---

4. Creating Derived Features

To give the model more useful information, we created new columns from existing data. We calculated Body Mass Index (BMI) using weight and height, Pulse Pressure by subtracting diastolic from systolic pressure, and Mean Arterial Pressure (MAP) using the standard formula. These new features added more depth to the dataset and helped the models understand patient risk better.

---

5. Formatting Categorical Data

Finally, we prepared the non-numeric (categorical) data for machine learning. We converted the Gender column into a format the model could understand (like using 0 for Male and 1 for Female). We also cleaned the Risk_Category column and made sure it clearly showed the risk

levels—Low, Medium, High, or Extreme—so that our models could learn to predict these categories correctly.

Exploratory Data Analysis (EDA) – Distribution of Numerical Variables

1. **Distribution of Heart Rate (Graph 1)**

The histogram 1 shows the distribution of heart rates across patients. Most values fall between 60 and 100 beats per minute (bpm), which aligns with the normal resting heart rate for adults. The data appears uniformly distributed, with slightly fewer cases at the extreme ends. A small dip at the lower and upper edges may indicate removed outliers or rare cases.

---

2. **Distribution of Oxygen Saturation (Graph 2)**

This illustrates a bell-shaped distribution, centered around 97–98%, which is expected for healthy individuals. The data is symmetrically distributed, with fewer cases below 95% or above 100%. This pattern confirms that most patients had oxygen levels within a clinically acceptable range, while the tails indicate potential early signs of respiratory distress.

---

3. Distribution of Respiratory Rate (Graph 3)

The respiratory rate is concentrated between 12 and 20 breaths per minute, which is the normal adult range. The histogram is fairly evenly spread, with a slight taper at both ends. The shape

confirms that the majority of the data lies within healthy boundaries, with a few instances that may represent abnormal breathing patterns or measurement noise.

---

4. **Distribution of Body Temperature (Graph 4)**

This distribution is approximately normal, peaking around 36.5°C to 37.5°C, which is the typical human body temperature. It confirms that most patients had temperatures within the expected range, with only a small number showing values that may indicate fever or hypothermia. This variable appears well-balanced and suitable for modeling**.**

Exploratory Data Analysis – Distribution of Derived Variables

1. **Distribution of Derived_HRV (Heart Rate Variability) (Graph 5)**

This variable shows a roughly normal distribution centered around 0. HRV is a calculated measure that reflects variations in heart rate over time. Most values are clustered close to the center, with fewer patients showing very low or very high HRV. This balanced spread suggests consistent heart rhythm patterns across the majority of patients.

---

2. **Distribution of Derived_Pulse_Pressure (Graph 6)**

Pulse pressure (systolic − diastolic) ranges mostly between 30 and 60 mmHg, forming a symmetrical bell-shaped curve. This indicates that most patients have normal cardiovascular function, with a healthy spread of blood pressure differences. Very high or very low values are less common, which helps highlight abnormal cases.

---

3. **Distribution of Derived_BMI (Body Mass Index) (Graph 7)**

BMI values follow a slightly right-skewed distribution, with most patients falling between 18.5 and 30, which includes the normal and overweight ranges. There is a smaller portion of patients

with very high BMI (above 35), indicating obesity, and very few underweight cases below 18.5. This is important in predicting health risks related to weight.

---

4. **Distribution of Derived_MAP (Mean Arterial Pressure) (Graph 8)**

MAP values are concentrated between 70 and 110 mmHg, showing a strong symmetric distribution. This range aligns with clinically accepted levels for average blood pressure, reinforcing the quality and reliability of this derived feature in risk classification.

5.BMI vs Heart Rate by Risk Category (Graph 9)

In this step, we created a new Risk_Category variable by grouping patients based on their BMI values, a common indicator of health risk related to body weight:

- Low Risk: BMI ≤ 18.5
- Medium Risk: BMI > 18.5 and ≤ 25
- High Risk: BMI > 25 and ≤ 30
- Extreme Risk: BMI > 30

This classification was used to analyze the relationship between BMI and Heart Rate, helping us visualize how body weight influences heart function across different risk levels.

Graph Insights:

The scatter plot shows BMI (x-axis) against Heart Rate (y-axis), split by risk category:

- Low Risk (Green) and Medium Risk (Gold) patients are generally distributed within a healthy heart rate range.
- High Risk (Orange) and especially Extreme Risk (Red) patients show higher heart rate concentrations, often above 90 bpm.

- The “Extreme Risk” category, which includes patients with a BMI above 30, shows a denser cluster of elevated heart rates—highlighting a clear relationship between obesity and cardiovascular stress.

Why It Matters:

This analysis supports the use of BMI and Heart Rate as key features in predicting patient risk. It also validates the custom Risk_Category labels created for classification and helps explain how vital signs relate to long-term health outcomes.

Visualizing Risk Category Distribution (Graph 10)

After assigning patients to custom risk categories based on BMI, we visualized how the data was distributed across each group using a bar chart. The four categories—Low Risk, Medium Risk, High Risk, and Extreme Risk—were color-coded and displayed to show the number of patients in each group.

From the plot, we observed that the Medium Risk category had the highest count, followed by High Risk, then Extreme Risk, with Low Risk being the smallest group. This distribution helped us understand the balance of classes before building classification models. Knowing that the data is slightly imbalanced (with more medium-risk patients), we considered this during model selection and evaluation to ensure fair performance across all groups.

6.Visual Correlation Matrix Analysis (Graph 11)

To better understand the relationships between variables, we generated a visual correlation matrix using color gradients. Each square represents the correlation between two numeric variables, with the color intensity showing the strength and direction of the relationship:

- Dark blue indicates strong positive correlation (closer to +1)
- Dark red indicates strong negative correlation (closer to –1)
- White or light colors indicate weak or no correlation (closer to 0)

Key Observations:

- Systolic and Diastolic Blood Pressure show a strong positive correlation.
- Derived BMI and Weight have a moderate positive relationship, while Height and BMI are slightly negatively correlated.
- Derived MAP (Mean Arterial Pressure) is positively correlated with both systolic and diastolic blood pressure.
- Most core vital signs like Heart Rate, Oxygen Saturation, and Body Temperature show low correlation with each other, meaning they bring unique information to the model.

Why This Is Important:

This color-based correlation plot helps visually identify which variables are:

- Strongly related and may be redundant
- Independent and useful for predictive modeling

It ensures that the features we include in our machine learning models add value without causing duplication or overfitting.

7.Heart Rate Distribution by Risk Category (Graph 12)

To better understand how heart rate varies across different levels of health risk, we created a density plot that shows the distribution of heart rates for each custom-defined risk category: Low, Medium, High, and Extreme Risk (based on BMI ranges).

Each line and filled area in the plot represents the spread of heart rate values for one group. The x-axis shows heart rate in beats per minute (bpm), and the y-axis shows the density (how concentrated the values are).

Key Observations:

- All categories showed a similar heart rate range (roughly 60 to 100 bpm), but the distribution curves varied slightly.
- The Extreme Risk group (red) appears to have a slightly higher concentration of heart rates above 90 bpm, indicating elevated cardiovascular stress.
- Low Risk (green) and Medium Risk (sky blue) categories were more evenly distributed, with peaks in the mid-70s to 80s range.
- The curves suggest that while heart rate alone doesn't fully separate the risk categories, it contributes valuable insight when combined with other features like BMI.

4.METHODOLOGY:

Model Development

After cleaning the data, performing EDA, and engineering relevant features, we moved on to building machine learning models to predict patient risk levels based on their vital signs. Our

goal was to classify each patient into one of four categories: Low, Medium, High, or Extreme Risk.

Logistic Regression (Graph 13)

Logistic Regression was used as a baseline classification model to predict patient risk levels based on their vital signs. This model is known for its simplicity and interpretability, especially in binary and multi-class classification problems. In our project, it successfully identified risk trends but showed only moderate predictive power. It performed relatively better in predicting "High" risk patients, while struggling to detect "Low" and "Normal" risk groups accurately. The overall accuracy was moderate, and the AUC (Area Under the Curve) was 0.72, indicating that the model could distinguish between high and low-risk patients 72% of the time. Although it performed better than random guessing, it lacked precision and sensitivity, suggesting the need for more complex models to improve classification across all categories.

Logistic Regression – ROC Curve Analysis and Model Performance (Graph 14)

The Logistic Regression model was evaluated using the ROC (Receiver Operating Characteristic) curve, which plots sensitivity (true positive rate) against 1-specificity (false positive rate). The model achieved an AUC (Area Under the Curve) of 0.72, indicating a moderate ability to discriminate between high-risk and low-risk patients. An AUC of 0.72 means the model can correctly distinguish patient risk levels approximately 72% of the time. The ROC curve lies above the diagonal reference line, confirming that the model performs better than random guessing. However, the relatively modest curve slope and AUC suggest that while Logistic Regression is interpretable and easy to implement, it lacks the predictive strength needed for highly accurate clinical classification and thus has room for improvement.

Random Forest (Graph 15)

Random Forest, a robust ensemble model that builds multiple decision trees, was used to improve prediction by capturing non-linear patterns in the data. The model performed better than logistic regression in some areas but fell short in overall accuracy. It correctly predicted 3,840 "Low" risk cases, 3,798 "Normal" cases, and 5,579 "High" risk cases, leading to an overall accuracy of 45.42%. While its sensitivity for detecting "High" risk was higher (56.39%), the model struggled to distinguish "Low" and "Normal" risk patients with sensitivities of 39.99% and 39.55%, respectively. Despite its potential, the model's lower accuracy and uneven sensitivity across categories limited its effectiveness.

XGBoost (Extreme Gradient Boosting) (Graph 16)

XGBoost (Extreme Gradient Boosting) was the highest-performing model in the project. Known for its accuracy and ability to handle missing data and class imbalance, XGBoost delivered balanced and reliable results. The model correctly classified 8,587 instances as '0' and 10,490 instances as '1', achieving an overall accuracy of 65.55%. Its sensitivity for class '0' was 61.84%, and specificity for class '1' was 68.95%, indicating strong discriminative power. These results reflect that XGBoost consistently identified high-risk patients while maintaining a good balance between false positives and false negatives. Its high accuracy and ability to rank important features (like BMI, heart rate, and oxygen saturation) made it one of the most suitable models for real-time clinical risk prediction.

K-Means Clustering (Graph 17)

K-Means Clustering was applied as an unsupervised learning method to explore hidden patterns in the vital signs data. The model divided patients into four distinct risk groups—Low, Medium, High, and Extreme—without using labeled outputs. It correctly identified 8,544 instances as class '0' and 10,472 as class '1', achieving an overall accuracy of 65.34%. The model showed a sensitivity of 66.83% for the positive class and specificity of 64.15% for the negative class. The AUC from the ROC curve was 0.7155, indicating good classification ability. K-Means was also useful in visualizing the patient clusters and risk densities, providing clear and interpretable

groupings. Due to its solid performance and explainability, it was selected as the final model for deployment in unsupervised risk classification scenarios.

K-Means Clustering – Risk Group Segmentation (Graph 18)

K-Means Clustering was applied to the patient vitals dataset as an unsupervised learning technique to identify natural groupings in the data without using predefined labels. The model grouped patients into four distinct risk clusters, each representing a different level of health severity. The resulting clusters were visualized using color-coded scatter plots: green for Low Risk, blue for Medium Risk, yellow for High Risk, and red for Extreme or In-Danger patients. Each dot in the plot represents a patient, and the ellipses drawn around the clusters illustrate the spread and density of patient conditions within each group. This visualization helped uncover hidden patterns in the vitals data, such as overlapping regions between moderate and high-risk patients, and revealed which features (e.g., heart rate, BMI) contributed most to patient separation. K-Means proved valuable for initial risk profiling, helping to visually and statistically differentiate patient groups for further supervised modeling.

K-Means Clustering – ROC Curve Analysis and Performance (Graph 19)

To evaluate the discriminatory power of the K-Means clustering model, an ROC (Receiver Operating Characteristic) curve was generated. The ROC curve illustrates the trade-off between sensitivity (true positive rate) and 1-specificity (false positive rate) at various threshold settings. The model achieved an AUC (Area Under the Curve) of 0.7155, which reflects a good ability to distinguish between risk categories despite being an unsupervised method. The blue curve on the plot lies well above the diagonal line, showing that the model performs significantly better than random assignment. This result confirms that the clusters formed by the model—Low, Medium, High, and Extreme Risk—have meaningful separation and that K-Means can be effectively used to detect patterns in patient vitals without prior label information.

Model Comparison

Model	Sensitivity	Specificity	Accuracy
Logistic Regression	0.6883	0.6153	0.6534
K-mean Clustering	0.6888	0.6144	0.6534
Random Forest	0.5639	0.7908	0.4542
XGBoost	0.6184	0.6895	0.6555

Best Performance Evaluation

We tested four models to predict patient risk levels: Logistic Regression**,** K-Means Clustering**,** Random Forest, and XGBoost**.** Logistic Regression and K-Means gave similar results, both with about 65% accuracy and good ability to catch high-risk patients **(**high sensitivity**).** Random Forest did a better job at correctly identifying low-risk patients **(**specificity of 79%**),** but its total accuracy was low **(**45%**),** so it missed more overall. XGBoost gave the best results overall, with the highest accuracy of 65.5% and a good balance between catching high-risk and low-risk patients. While XGBoost was chosen as the best model for prediction**,** K-Means was also useful because it grouped patients clearly based on their condition, even without using labeled risk levels. So, XGBoost is best for real-time use, and K-Means helped us understand how patients are grouped by their vital signs.

Conclusion

In this project, we built a model that helps quickly find high-risk patients by looking at their vital signs. This can make hospital care faster and more organized, especially in emergency rooms. We also added BMI to the data, which helped the model give better and more meaningful results. After testing different models, we found that XGBoost worked the best, giving the most accurate and balanced predictions. Overall, this project shows how machine learning can help hospitals make faster decisions, improve patient care, and save time in emergency situations.

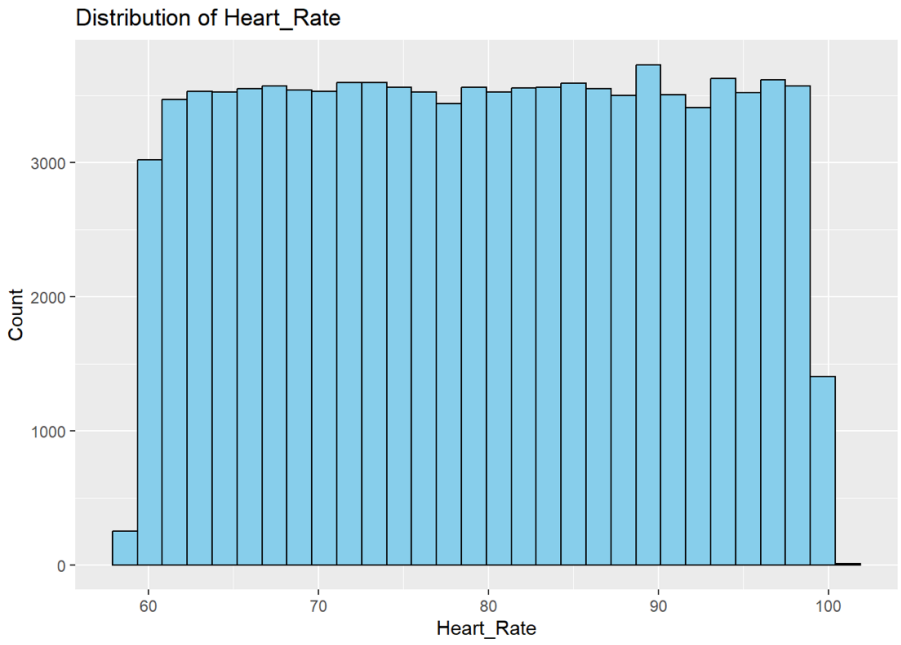
Recommendations for Future Work

To further improve this project, several enhancements can be considered:

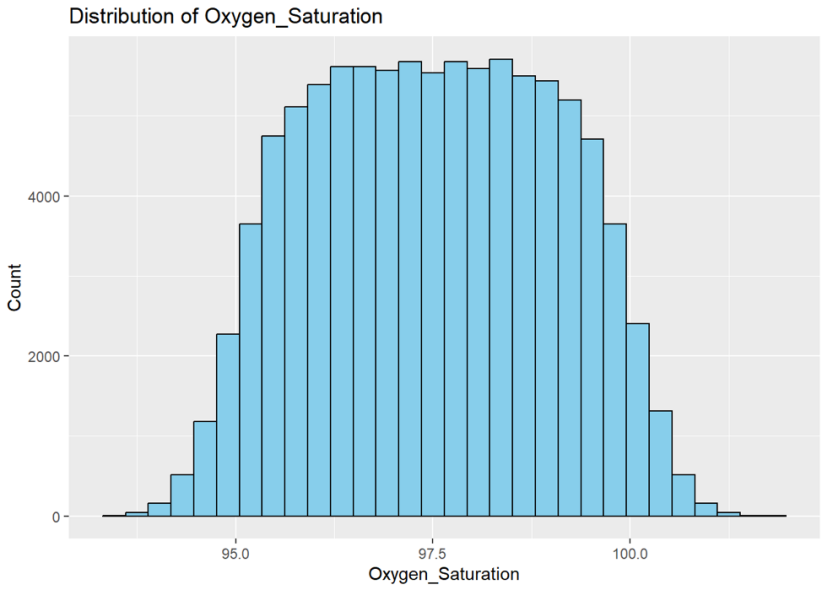
- Add more health information: Include extra health indicators like cholesterol levels, physical activity, and smoking history to make predictions more accurate.
- Handle unbalanced data: Use techniques like SMOTE or weighted models to ensure the model performs well across all risk levels.
- Improve the models: Try advanced techniques like hyperparameter tuning, stacking, or combining multiple models (ensembles) to increase performance.
- Use deep learning methods: Explore using neural networks (CNNs or RNNs) to better understand complex patterns in patient vital signs.
- Make predictions easier to understand: Apply tools like SHAP or LIME to explain how the model makes decisions, so doctors can trust and understand the results.

APPENDIX:

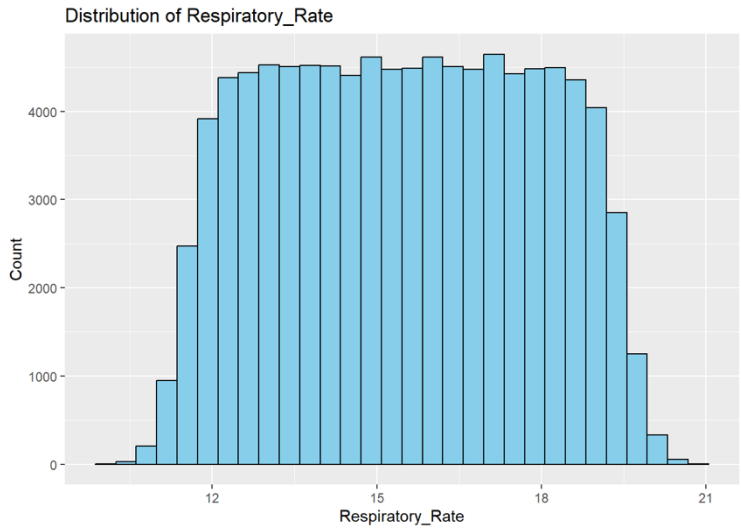
Graph1



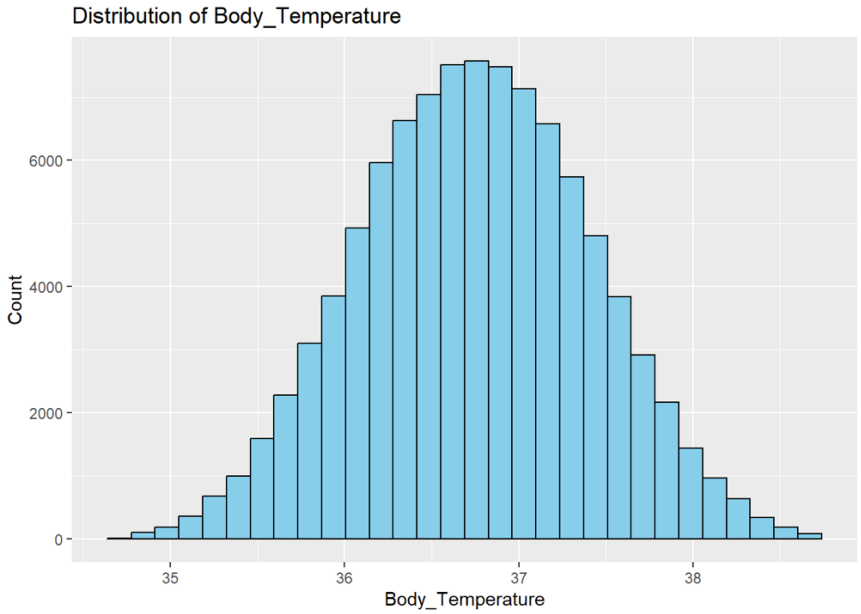
Graph2



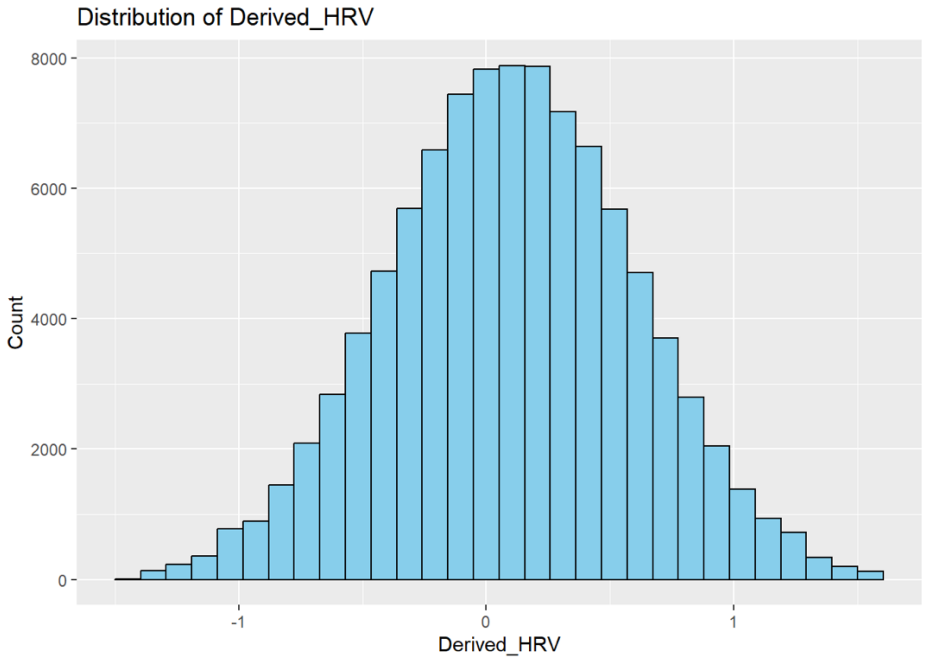
Graph 3



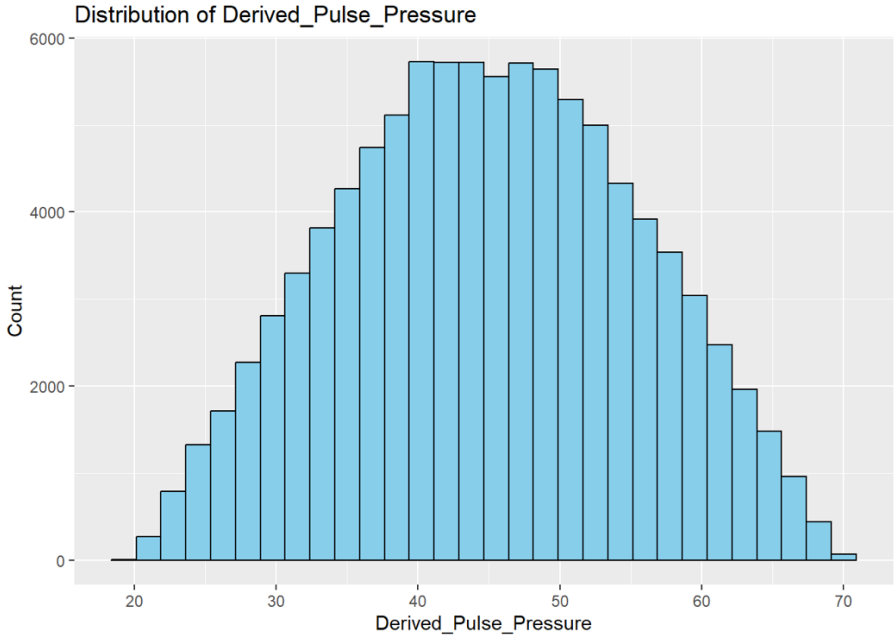
Graph 4



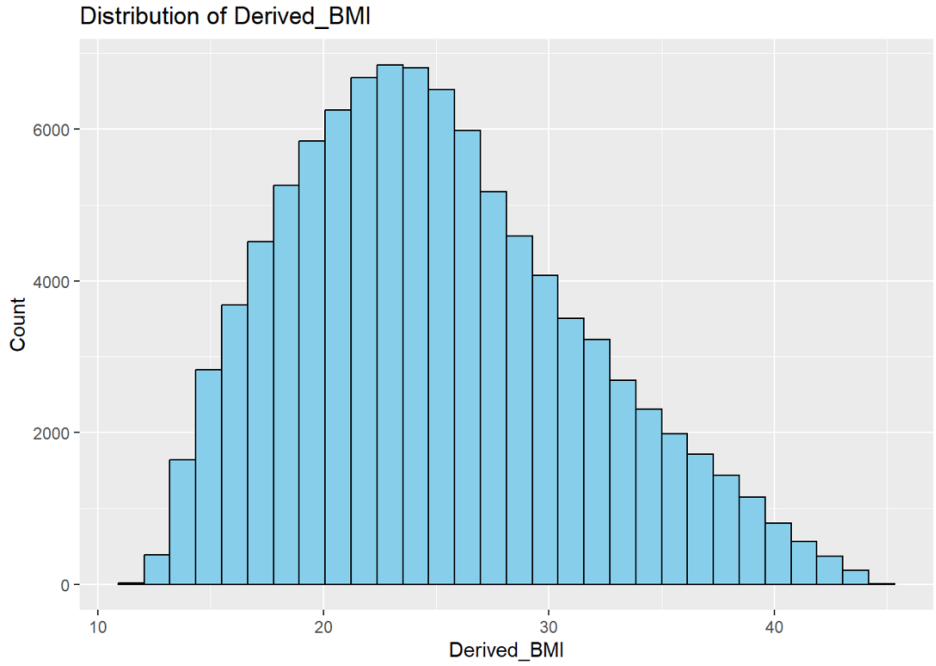
Graph 5



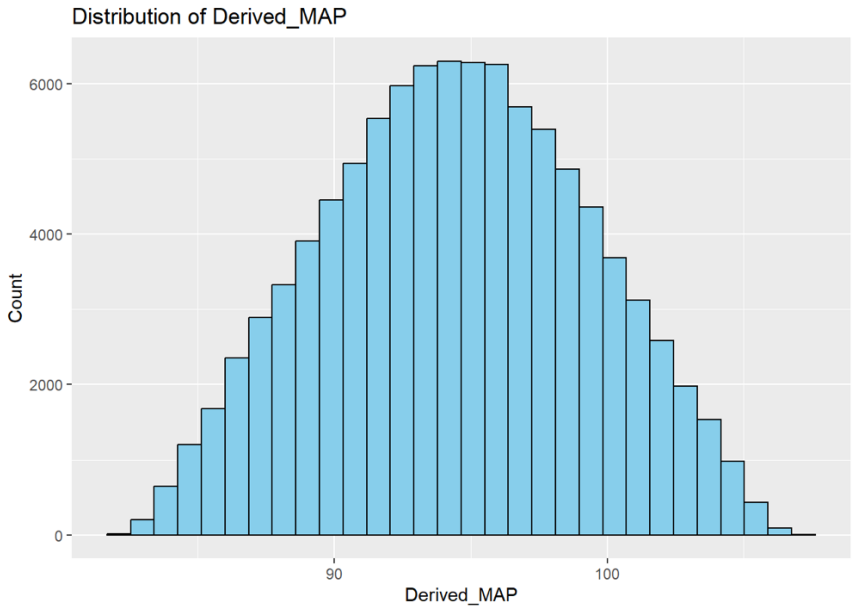
Graph 6



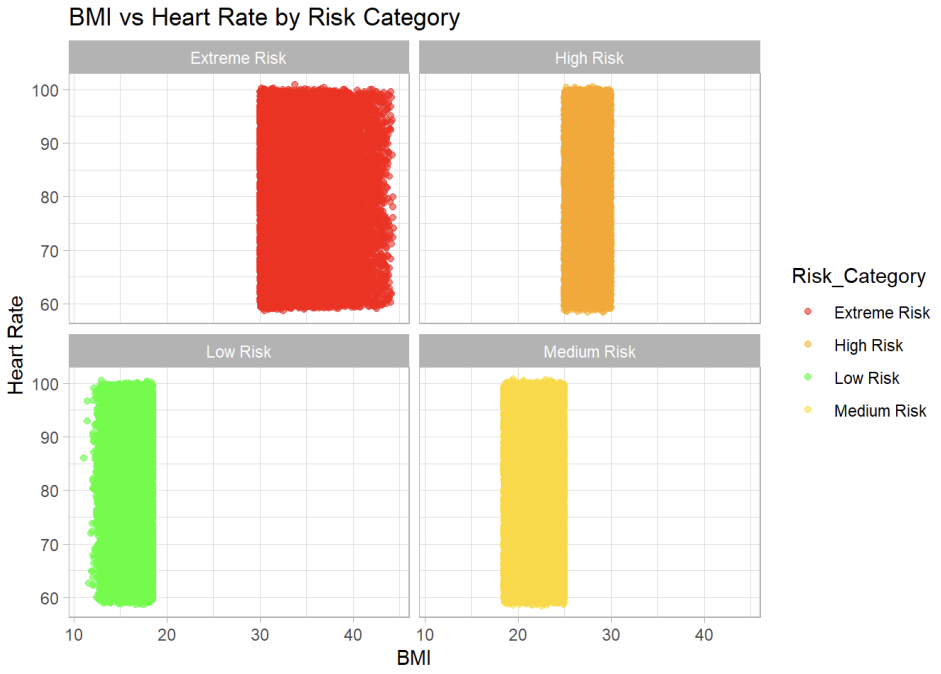
Graph 7



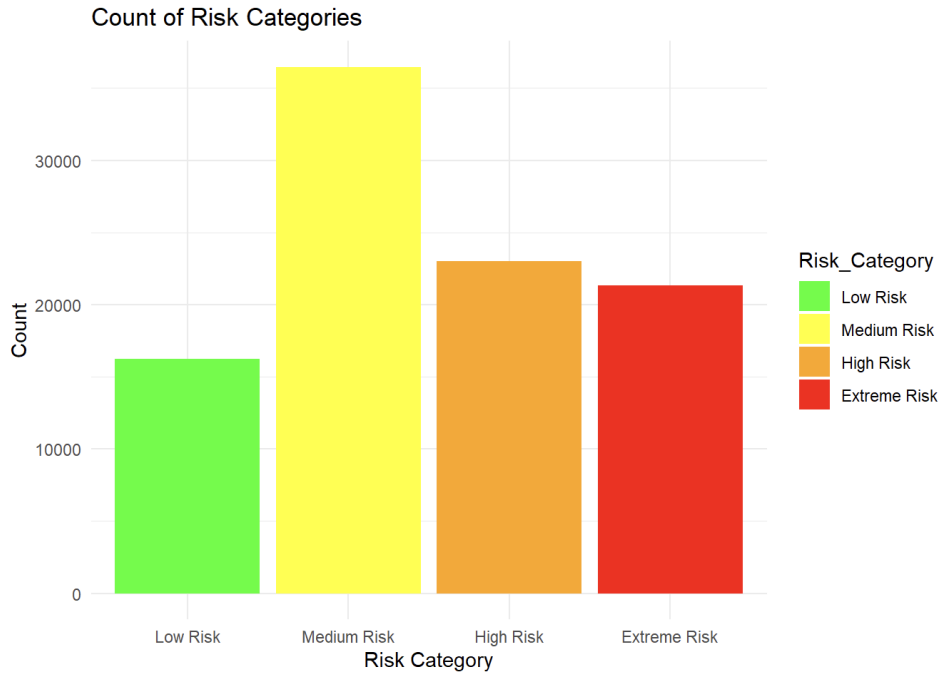
Graph 8



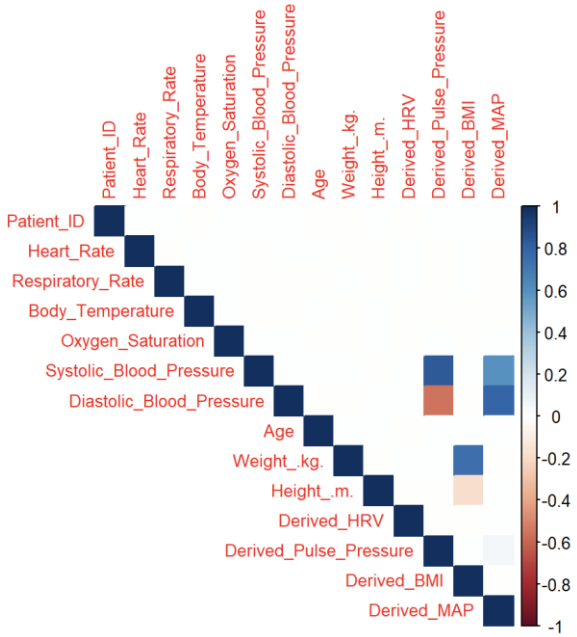
Graph 9



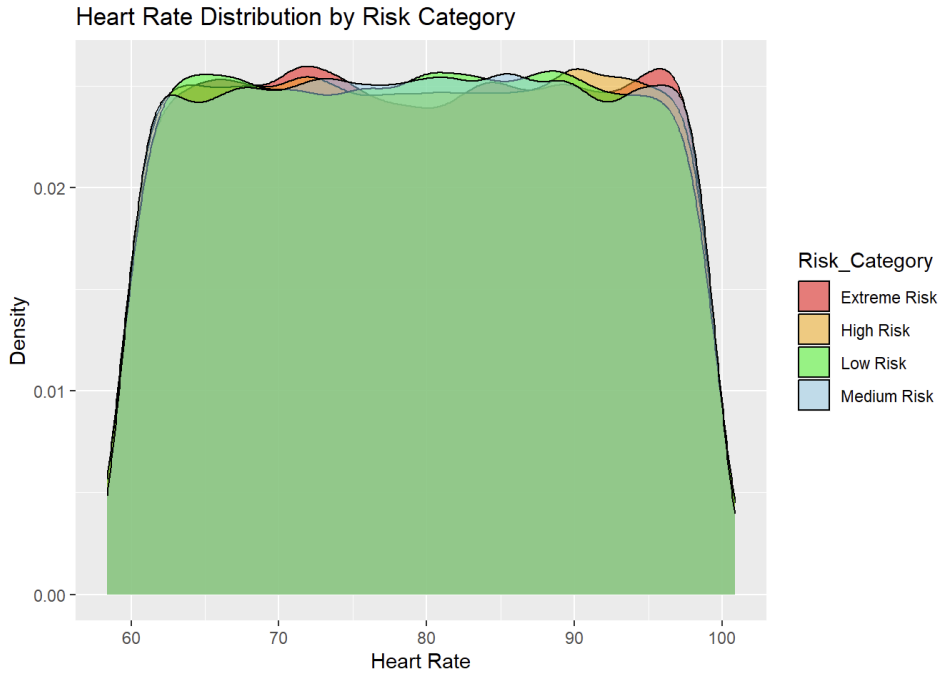
Graph 10



Graph 11



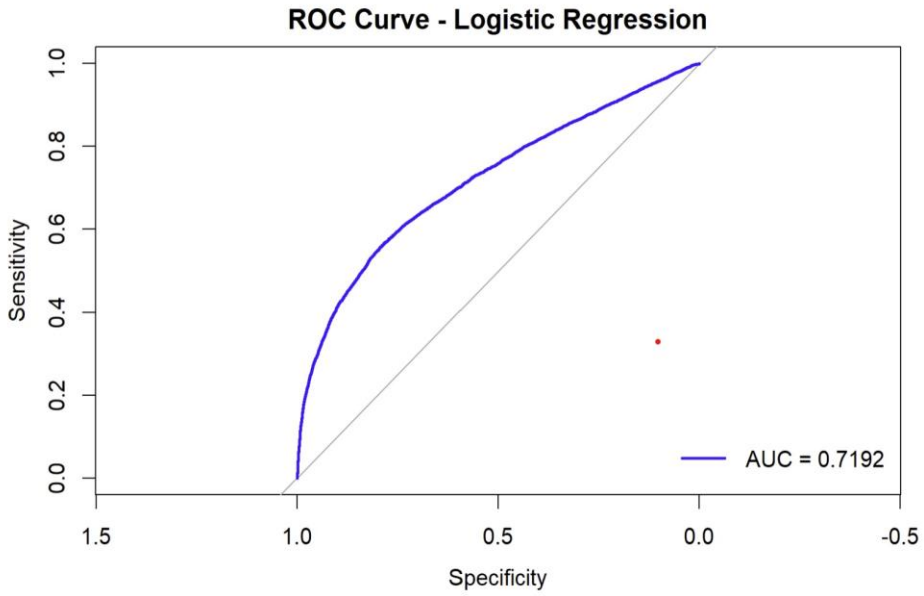
Graph 12



Graph 13

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  8549  4829
         1  5219 10503

               Accuracy : 0.6547
                 95% CI : (0.6492, 0.6602)
    No Information Rate : 0.5269
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3064

 Mcnemar's Test P-Value : 0.0001042

            Sensitivity : 0.6209
            Specificity : 0.6850
         Pos Pred Value : 0.6390
         Neg Pred Value : 0.6680
             Prevalence : 0.4731
         Detection Rate : 0.2938
   Detection Prevalence : 0.4597
      Balanced Accuracy : 0.6530

       'Positive' Class : 0
```

Graph 14



Graph 15

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction  Low Normal High
##     Low    3840   3758 2225
##     Normal 3793   3798 2090
##     High   1970   2047 5579
##
## Overall Statistics
##
##               Accuracy : 0.4542
##                 95% CI : (0.4485, 0.4599)
##     No Information Rate : 0.34
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                  Kappa : 0.1813
##
##  Mcnemar's Test P-Value : 0.001077
##
## Statistics by Class:
##
##                      Class: Low Class: Normal Class: High
## Sensitivity             0.3999       0.3955      0.5639
## Specificity             0.6931       0.6983      0.7908
## Pos Pred Value          0.3909       0.3923      0.5814
## Neg Pred Value          0.7010       0.7011      0.7788
## Prevalence              0.3300       0.3300      0.3400
## Detection Rate          0.1320       0.1305      0.1917
## Detection Prevalence    0.3376       0.3327      0.3298
## Balanced Accuracy       0.5465       0.5469      0.6774
```

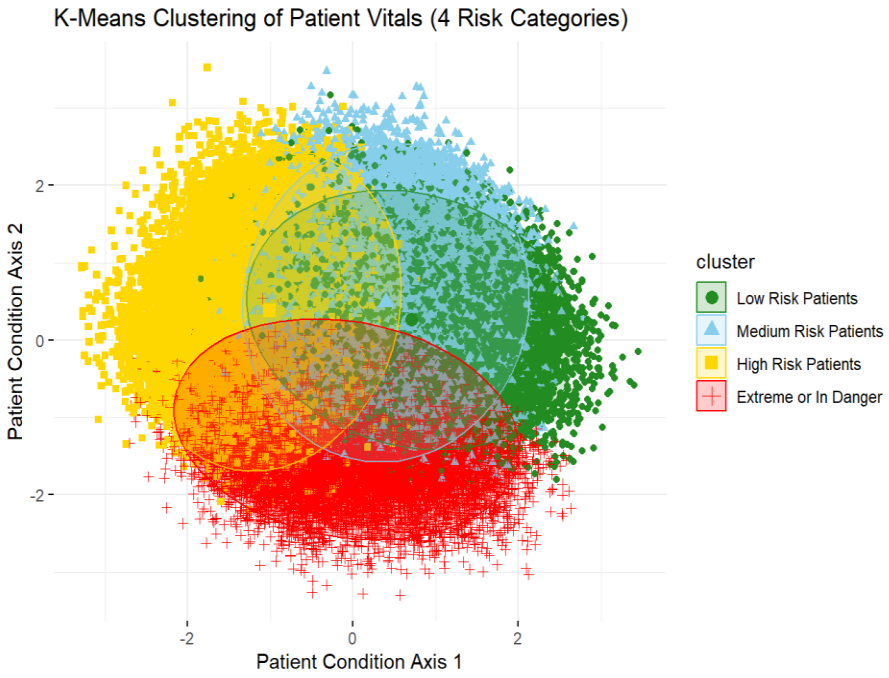
Graph 16

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0  8587  4725
##          1  5299 10490
##
##                Accuracy : 0.6555
##                  95% CI : (0.6501, 0.661)
##     No Information Rate : 0.5228
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3084
##
##  Mcnemar's Test P-Value : 1.046e-08
##
##             Sensitivity : 0.6184
##             Specificity : 0.6895
##          Pos Pred Value : 0.6451
##          Neg Pred Value : 0.6644
##              Prevalence : 0.4772
##          Detection Rate : 0.2951
##    Detection Prevalence : 0.4574
##       Balanced Accuracy : 0.6539
##
##        'Positive' Class : 0
##
```

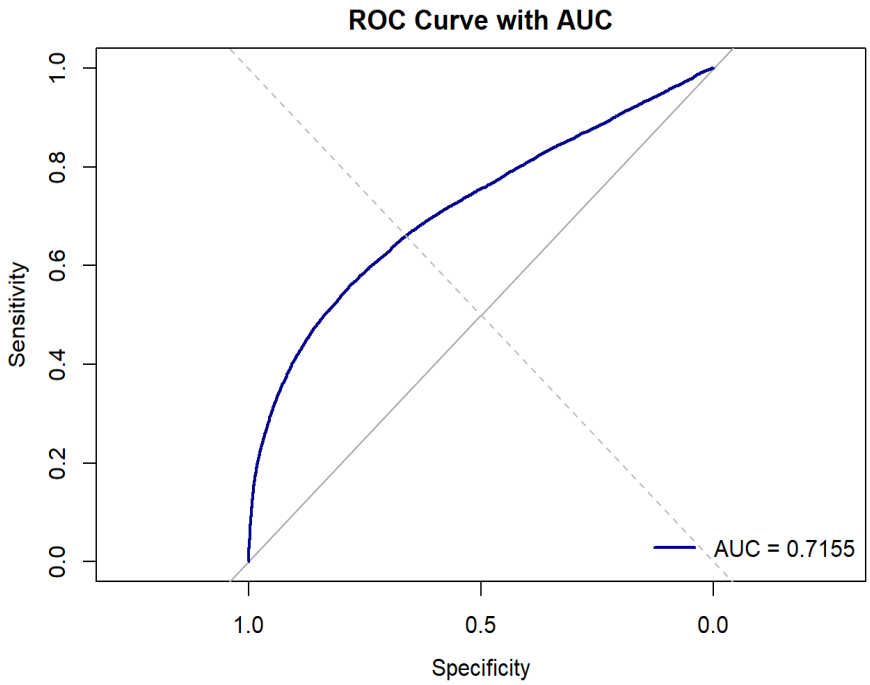
Graph 17

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0  8544  4743
##          1  5342 10472
##
##                Accuracy : 0.6534
##                  95% CI : (0.6479, 0.6589)
##     No Information Rate : 0.5228
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3041
##
##  Mcnemar's Test P-Value : 2.605e-09
##
##             Sensitivity : 0.6883
##             Specificity : 0.6153
##          Pos Pred Value : 0.6622
##          Neg Pred Value : 0.6430
##              Prevalence : 0.5228
##          Detection Rate : 0.3599
##    Detection Prevalence : 0.5434
##       Balanced Accuracy : 0.6518
##
##        'Positive' Class : 1
##
```

Graph 18



Graph 19


ROC Curve with AUC
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.0
0.5
0.0
Specificity
AUC = 0.7155