**Project Report**

**On**

**Prediction Of Preoperative Risk**



*Submitted*

*In partial fulfilment*

*For the award of the Degree of*

**PG-Diploma in Big Data Analytics**

**(C-DAC, ACTS (Pune))**

**Guided By: Submitted By:**

Mr. Saurabh Panwar Siddhesh Thakare 230940125048 Bharti Ghildiyal 230940125008

Anjali Kohar 230940125005

Khushi Songirkar 230940125026

Amey Pote 230940125002

**Centre for Development of Advanced Computing (C-DAC), ACTS (Pune- 411008)**

## Acknowledgement

This is to acknowledge our indebtedness to our Project Guide, **Mr. Saurabh Panwar,** C-DAC ACTS, Pune for his constant guidance and helpful suggestion for preparing this project **Prediction of Preoperative Risk.** We express our deep gratitude towards his for inspiration, personal involvement, constructive criticism that he provided us along with technical guidance during the course of this project.

We take this opportunity to thank Head of the department **Mr. Gaur Sunder** for providing us such a great infrastructure and environment for our overall development.

We express sincere thanks to **Mrs. Namrata Ailawar**, Process Owner, for their kind cooperation and extendible support towards the completion of our project.

It is our great pleasure in expressing sincere and deep gratitude towards **Mrs. Risha P R (Program Head)** and **Mrs. Srujana Bhamidi** (Course Coordinator, PG-DBDA) for their valuable guidance and constant support throughout this work and help to pursue additional studies.

Also, our warm thanks to **C-DAC ACTS Pune**, which provided us this opportunity to carry out, this prestigious Project and enhance our learning in various technical fields.

Siddhesh Thakare 230940125048 Bharti Ghildiyal 230940125008

Anjali Kohar 230940125005

Khushi Songirkar 230940125026

Amey Pote 230940125002

## ABSTRACT

In this project, we explore the application of machine learning (ML) techniques in predicting preoperative risk. Preoperative risk prediction plays a crucial role in optimizing patient outcomes and resource allocation in surgical settings. Through a systematic review of relevant literature, we examine the efficacy of ML algorithms in developing predictive models for various perioperative outcomes, including mortality risk, systemic complications, ICU admission, and length of hospital stay. Our analysis highlights the potential of ML to improve risk assessment accuracy compared to traditional scoring systems, paving the way for enhanced patient care and resource utilization in perioperative medicine.

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S. No** | **Title** | **Page No.** |
|  | **Front Page Acknowledgement Abstract**  **Table of Contents** | **I II III**  **IV** |
| **1**  **1.1**  **1.2** | **Introduction**  Introduction  Objective and Specifications | **01-02**  **01**  **02** |
| **2** | **Literature Review** | **03-04** |
| **3** | **Methodology/ Techniques** | **05-15** |
| **3.1** | Datasets | **05** |
| **3.2** | Approach and Methodology/ Techniques | **12** |
| **3.3** | Model Description | **13-15** |
| **4** | **Implementation** | **16** |
| **4.1** | Implementation | **16** |
|  |  |  |
| **5** | **Results** | **21-22** |
| **5.1** | Results | **21** |
| **6** | **Conclusion** | **23** |
| **6.1** | Conclusion | **23** |
| **7** | **References** | **24-25** |
| **7.1** | References | **24** |

**Chapter 1**

**Introduction**

Detection of text regions either from handwritten or printed document images containing various non-textual information is a difficult task, and it can be more challenging to locate the position of the text regions when we deal with a doctor’s prescription.

**1.1 Introduction**

In the era of advanced healthcare systems, the ability to predict patient outcomes accurately plays a pivotal role in enhancing clinical decision-making and improving patient care. Perioperative risk assessment, particularly the prediction of postoperative mortality, is a critical aspect of preoperative evaluation in surgical settings. Traditionally, risk assessment relied on validated prognostic scores and statistical methods. However, with the advent of machine learning (ML) techniques and the availability of large-scale healthcare data, there is a growing interest in leveraging ML models to develop predictive models for perioperative risk stratification.

This project focuses on the development and evaluation of an ML-based model aimed at assessing an individual patient's risk of postoperative mortality based on preoperative data. Leveraging a comprehensive dataset containing various preoperative variables, including demographic information, comorbidities, laboratory values, and procedural details, our objective is to train a predictive model capable of accurately stratifying patients into different risk categories.

The project workflow involves several key steps, starting with data preprocessing and feature engineering to ensure data quality and relevance. Subsequently, various ML algorithms are explored and evaluated to identify the most suitable model for the task. The model is trained on a balanced dataset using techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance issues commonly encountered in medical datasets. Hyperparameter tuning and cross-validation are employed to optimize model performance and ensure generalizability.

Furthermore, interpretability of the model is crucial for clinical adoption and trust. Therefore, feature importance analysis is conducted to identify the most influential factors contributing to postoperative mortality risk prediction. Additionally, model performance metrics such as accuracy, precision, recall, and F1-score are computed, and a detailed analysis of the model's predictive capabilities is provided.

Ultimately, the goal of this project is to develop a robust and interpretable ML model that can assist clinicians in preoperative risk assessment, thereby facilitating informed decision-making and improving patient outcomes in perioperative care settings.

**1.2 Objective**

The objectives of the project work are as -

* Development of a Predictive Model: Create a machine learning model to accurately predict individual postoperative mortality risk based on preoperative data.
* Evaluation of ML Algorithms: Explore and assess various machine learning algorithms to determine the most effective approach for preoperative risk prediction.
* Performance Evaluation: Rigorously validate the developed model using real-world clinical data to ensure its accuracy and reliability in predicting patient outcomes.
* Integration of Interdisciplinary Insights: Integrate insights from healthcare and artificial intelligence disciplines to enhance the preoperative risk assessment process and improve patient care strategies.

**Chapter 2**

**LITERATURE REVIEW**

Shaziz Hafiz, Nicholas Peter Lees [1] The study investigates the role of POSSUM scoring in guiding resource allocation for emergency surgical patients. High-risk patients, identified by POSSUM scores, received enhanced perioperative care, with senior surgeons present during surgery and postoperative ICU/HDU beds allocated as needed. These findings highlight the utility of POSSUM scoring in optimizing care for high-risk surgical patients.

Joanna Abraham, Brian Bartek , Alicia Meng ,S. Avidan [2] The paper examines the implementation of machine learning (ML) in perioperative care to improve risk assessment for surgical patients. Through clinician evaluations, it confirms the alignment between ML-generated risk predictions and manual clinician rankings. The study proposes user-friendly ML visualization formats, facilitating efficient care planning and tailored interventions. Key suggestions include enhancing electronic tool delivery and building trust in ML models. Overall, the research highlights the potential of ML in optimizing perioperative care planning and offers insights for future tool design

Martin Graeßner, Bettina Jungwirth, Elke Frank, Stefan Josef Schaller [3] The study develops an interpretable machine learning model to predict postoperative mortality risk based on preoperative data. Using data from over 66,000 elective surgeries, the model achieves high accuracy (AUROC 0.95, AUPRC 0.109) and identifies key risk factors such as age and C-reactive protein levels. This model enables personalized risk assessment, aiding in preoperative optimization and informed decision-making for patient care.

Chiew Calvin, Liu Nan, Wong Ting Hway, Sim Yilin, Abdullah Hairil [4] The study compares machine learning models to traditional risk calculators (CARES and ASA-PS) in predicting 30-day postsurgical mortality and ICU stay >24 hours. Machine learning models outperform traditional ones, with gradient boosting achieving the best performance (AUPRC 0.23 for mortality, 0.38 for ICU admission). The study, conducted at Singapore General Hospital, demonstrates the potential of machine learning in improving surgical risk prediction, offering insights for clinical shared decision-making and resource allocation.

Hafsa Habehh1 and Suril Gohel [5] The paper provides an overview of machine learning (ML) and artificial intelligence (AI) applications in healthcare, discussing various learning approaches and algorithms. It highlights the advancements in AI technology and its growing role in predicting health emergencies, disease populations, and immune responses. The review covers ML applications in radiology, genetics, electronic health records, and neuroimaging, along with the associated risks and challenges such as privacy and ethical concerns. Overall, it emphasizes the potential of ML-based approaches to improve healthcare outcomes and suggests future applications in the field

Saifur Rahman, Jingwen Zhou, James Jin Kan [6] The paper explores the application of machine learning algorithms in improving healthcare data accuracy and efficiency, particularly focusing on time series healthcare metrics for heart rate data transmission. It reviews various machine learning algorithms, both supervised and unsupervised, and evaluates their feasibility for small and large datasets. The study aims to address the gap in the literature regarding the practical use of machine learning in healthcare, offering contributions in supervised and unsupervised machine learning, and providing a comparative analysis of machine learning models. Overall, the paper underscores the potential of machine learning to enhance healthcare data analysis and suggests directions for future research in the field.

**Chapter 3**

**Methodology and Techniques**

**3.1 Dataset (VitalDB)**

VitalDB serves as the foundational dataset for our project, providing a rich collection of preoperative patient data extracted from electronic health records (EHRs) across various healthcare institutions. This dataset encompasses a wide array of essential variables, including demographic information, medical history, physiological indicators, and details pertaining to surgical procedures. Its comprehensive nature allows for the exploration of diverse risk factors associated with perioperative mortality.

The VitalDB dataset is meticulously curated and partitioned into distinct subsets for training, validation, and testing purposes. Prior to model development, rigorous preprocessing steps are undertaken to ensure data integrity and consistency. These preprocessing steps include handling missing values, encoding categorical variables, and scaling numerical features.

VitalDB have total 5 csv files **clinical\_data.csv, clinical\_parameters.csv, lab\_data.csv , lab\_parameters.csv track\_names.csv**

The availability of VitalDB enables us to develop robust predictive models tailored to assess individualized perioperative risk. By leveraging machine learning techniques on this comprehensive dataset, we aim to enhance our understanding of factors influencing postoperative outcomes and ultimately improve patient care and safety in the perioperative setting.

**3.1 Methodology:**

3.1.1 Data Preprocessing

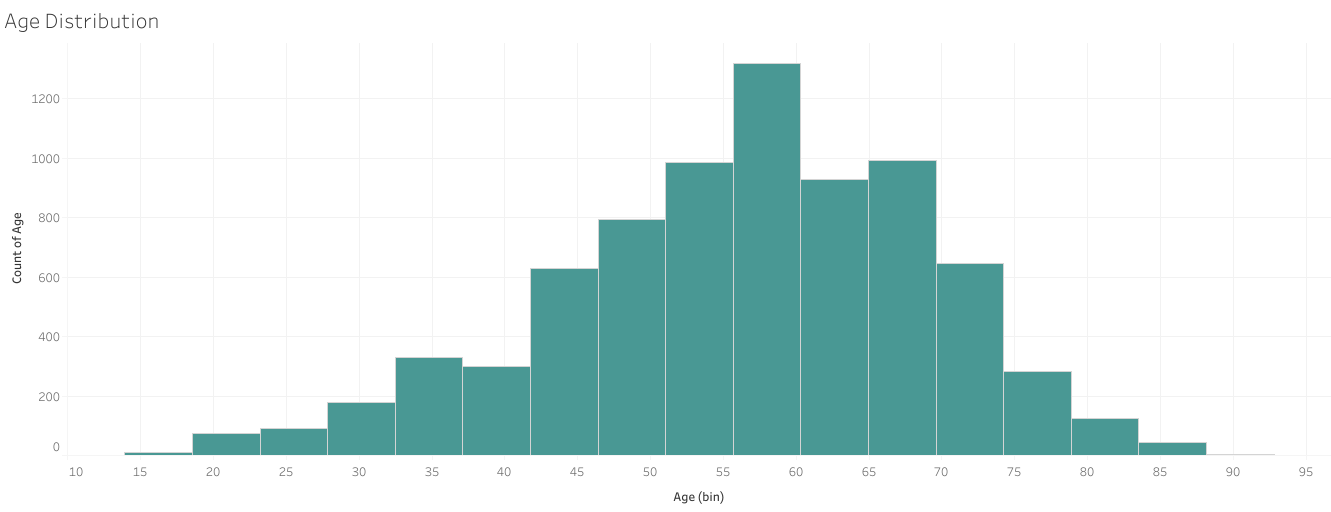
The dataset underwent several preprocessing steps to ensure its suitability for model training and evaluation. These steps included handling missing values through imputation, encoding categorical variables, removing underage and emergency cases, and scaling numerical features using standardization. Additionally, feature engineering techniques were applied to create new features and handle outliers, enhancing the representation of the data for modeling purposes.

Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset and understand its characteristics before proceeding with model building. The analysis began by examining the distribution of key features, including patient demographics, preoperative conditions, and surgical parameters. Histograms and density plots were utilized to visualize the distribution of variables, providing a clear understanding of their spread and potential outliers. Additionally, missing data were identified and addressed through imputation techniques to ensure the integrity of the dataset. Categorical variables were encoded to facilitate their incorporation into machine learning models. Overall, EDA played a crucial role in identifying patterns, trends, and potential challenges within the data, laying the foundation for subsequent modeling efforts aimed at predicting postoperative mortality risk based on preoperative information.

The histogram of the age distribution in the dataset includes a total of 7,736 patient cases. The mean age of the patients is approximately 56.33 years, with a standard deviation of 12.63 years. The minimum age recorded in the dataset is 18 years, while the maximum age is 89 years.

The histogram is skewed to the right, with a larger concentration of patient cases in the 48-65 age range. The median age (50th percentile) is 57 years, and the interquartile range (25th to 75th percentile) spans from 48 to 65 years. This suggests that a significant proportion of the patients in the dataset fall within this range.

Overall, the age distribution in the dataset shows a concentration of patient cases in middle-aged to older individuals, with a relatively broad distribution ranging from 18 to 89 years.

Figure 1.1

The histogram shows the distribution of BMI (Body Mass Index) values in the dataset, which includes 7,736 patient cases. The mean BMI is approximately 23.08, with a standard deviation of 2.94. The minimum BMI recorded in the dataset is 12.9, while the maximum BMI is 43.2.

The histogram is approximately normally distributed, with a peak around the 22-25 BMI range. The median BMI (50th percentile) is 22.8, and the interquartile range (25th to 75th percentile) spans from 21.3 to 24.5. This suggests that a significant proportion of the patients in the dataset have BMIs within this range.

Overall, the BMI distribution in the dataset indicates that the majority of the patients have BMIs in the normal to slightly overweight range, with some variation in both lower and higher BMI values.

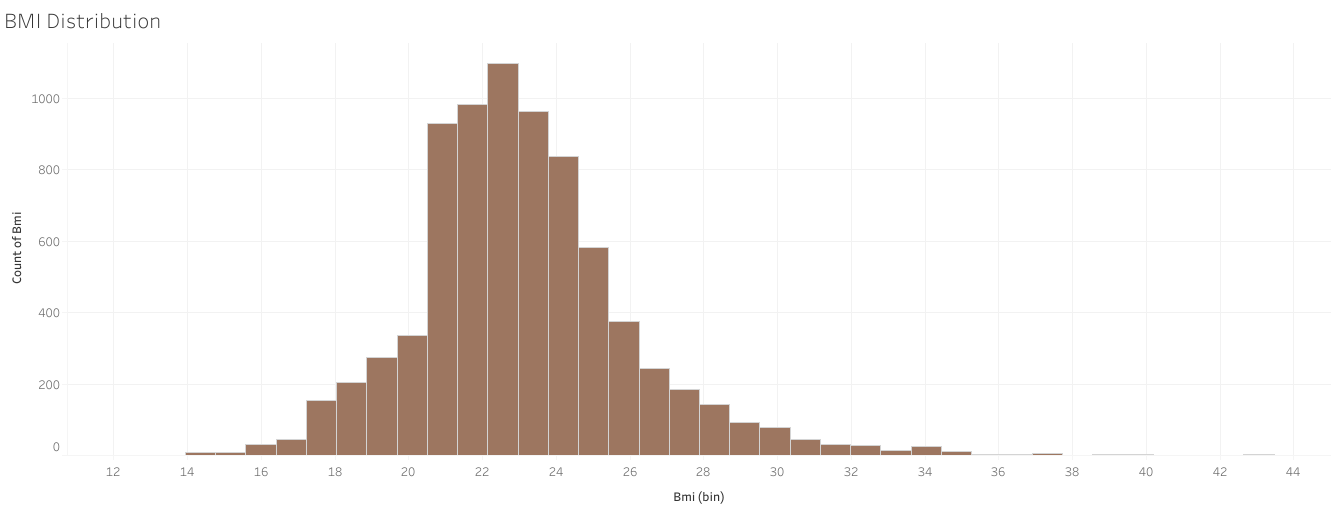


Figure 1.2

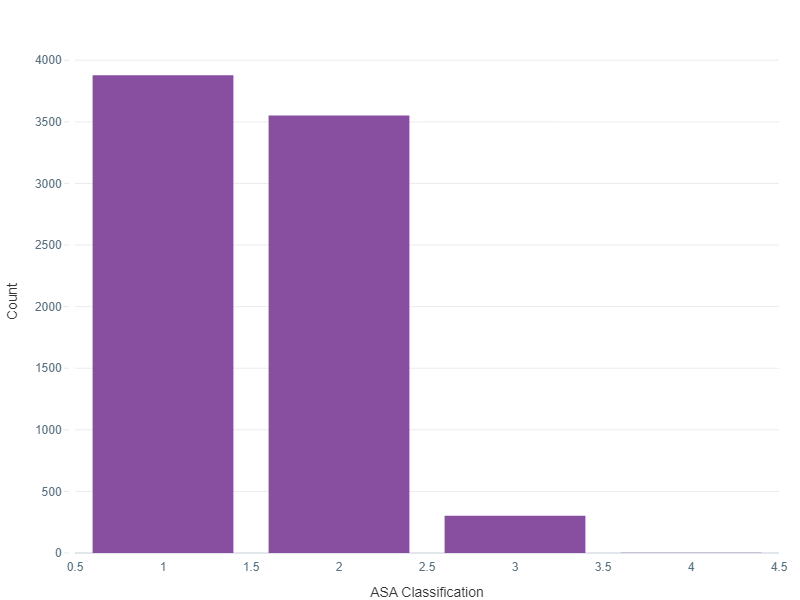


Figure 1.3

The above graph shows the distribution of ASA (American Society of Anesthesiologists) classifications among the patients. Here is a summary of the distribution:

- ASA Class 1: Count - 3879

- ASA Class 2: Count - 3551

- ASA Class 3: Count - 303

- ASA Class 4: Count - 3

The total count of patients included in the analysis is 7736. The summary statistics of the ASA classifications are as follows:

- Mean: 1.54

- Standard Deviation: 0.57

- Minimum: 1.0

- 25th Percentile: 1.0

- Median (50th Percentile): 1.0

- 75th Percentile: 2.0

- Maximum: 4.0

Based on this distribution, the majority of the patients fall into ASA Class 1 and Class 2, indicating relatively low to moderate surgical risk. There are a smaller number of patients in ASA Class 3, indicating a higher surgical risk, and only a few patients in ASA Class 4, indicating a significantly high surgical risk

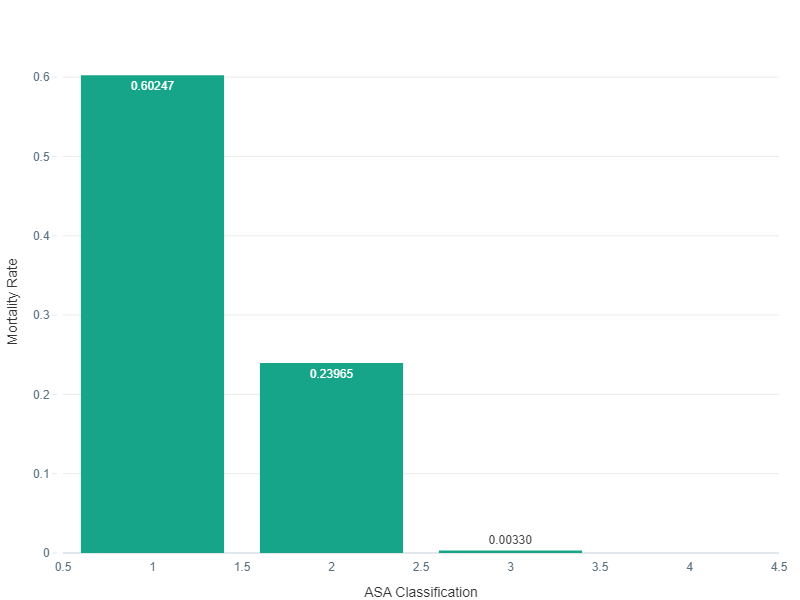


Figure 1.4

Mortality Across ASA Classifications:

The ASA classification with the highest mortality rate is 1.0, with a mortality rate of 60.25%. The ASA classification with the lowest mortality rate is 4.0, with a mortality rate of 0%.In conclusion, the mortality rates vary across different ASA classifications. ASA classification 1.0 has the highest mortality rate, while ASA classification 4.0 has the lowest mortality rate.

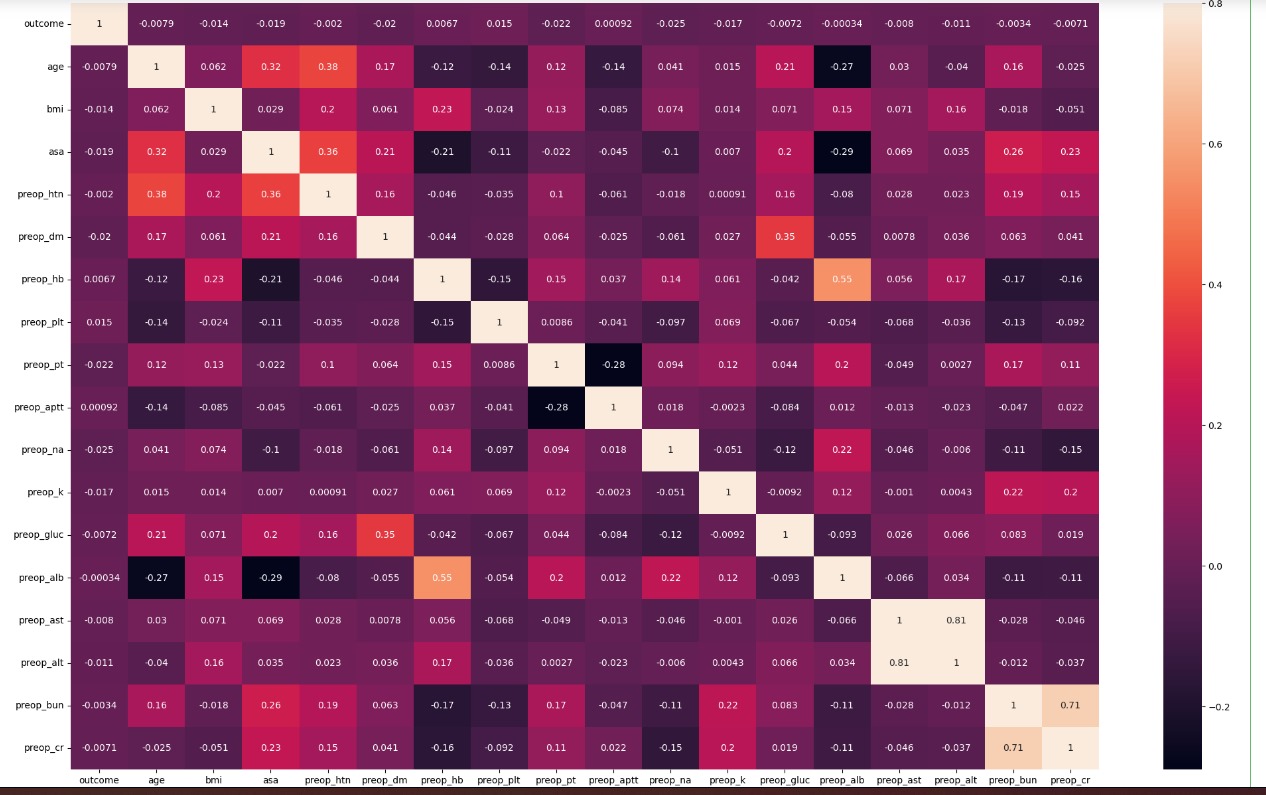


Figure 1.5

The correlation heatmap provides a visual representation of the relationships between specific numerical features in the dataset. Here are the key observations and conclusions from the heatmap:

1. Positive Correlations:

- Preoperative Platelet Count (preop\_plt) and Preoperative Hemoglobin (preop\_hb) show a moderate positive correlation.

- Preoperative Creatinine (preop\_cr) and Blood Urea Nitrogen (preop\_bun) exhibit a strong positive correlation.

2. Negative Correlations:

- There are no significant negative correlations observed among the selected features.

3. Weak Correlations:

- Most of the other selected features show weak correlations with each other, indicating a lack of strong linear relationships

3.1.2 Model Training

Various machine learning classifiers were considered for predicting perioperative risk based on preoperative data. The dataset was split into training, validation, and test sets using a stratified approach to preserve class distribution. Techniques such as oversampling using SMOTE were applied to address class imbalance, ensuring that the models were trained on a balanced dataset. Models were trained using algorithms such as ExtraTreesClassifier, Logistic Regression, and Support Vector Machines (SVM), leveraging the scikit-learn library in Python.

3.1.3 Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the performance of the selected classifiers. Grid search, coupled with cross-validation, was employed to search for the best combination of hyperparameters, such as the number of estimators, maximum depth, and minimum samples split and leaf. The goal was to identify hyperparameters that maximize model performance while avoiding overfitting.

3.1.4 Model Evaluation

The trained models were evaluated using multiple metrics, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). Confusion matrices were analyzed to assess the models' performance in predicting different classes of perioperative risk. Cross-validation was also performed to assess the models' generalization performance and mitigate the risk of overfitting.

3.1.5 Feature Importance Analysis

Feature importance analysis was conducted to identify the most influential features in predicting perioperative risk. Techniques such as permutation importance or SHAP (SHapley Additive exPlanations) values were used to quantify the impact of individual features on model predictions. This analysis provided insights into the factors contributing most significantly to the prediction outcome and helped prioritize features for further investigation.

3.1.6 Model Interpretability

Efforts were made to interpret the trained models and understand the underlying decision-making process. Visualizations such as decision trees, partial dependence plots, and feature importance plots were utilized to gain insights into how the models make predictions. This interpretability analysis enhanced the transparency and trustworthiness of the models' predictions, facilitating their adoption in clinical decision-making.

3.1.7 Model Deployment Considerations

While not directly part of the methodology for model development, considerations for deploying the trained model in a real-world setting were addressed. Factors such as model scalability, latency requirements, and integration with existing healthcare systems were taken into account to ensure the seamless transition from model development to deployment in clinical practice.

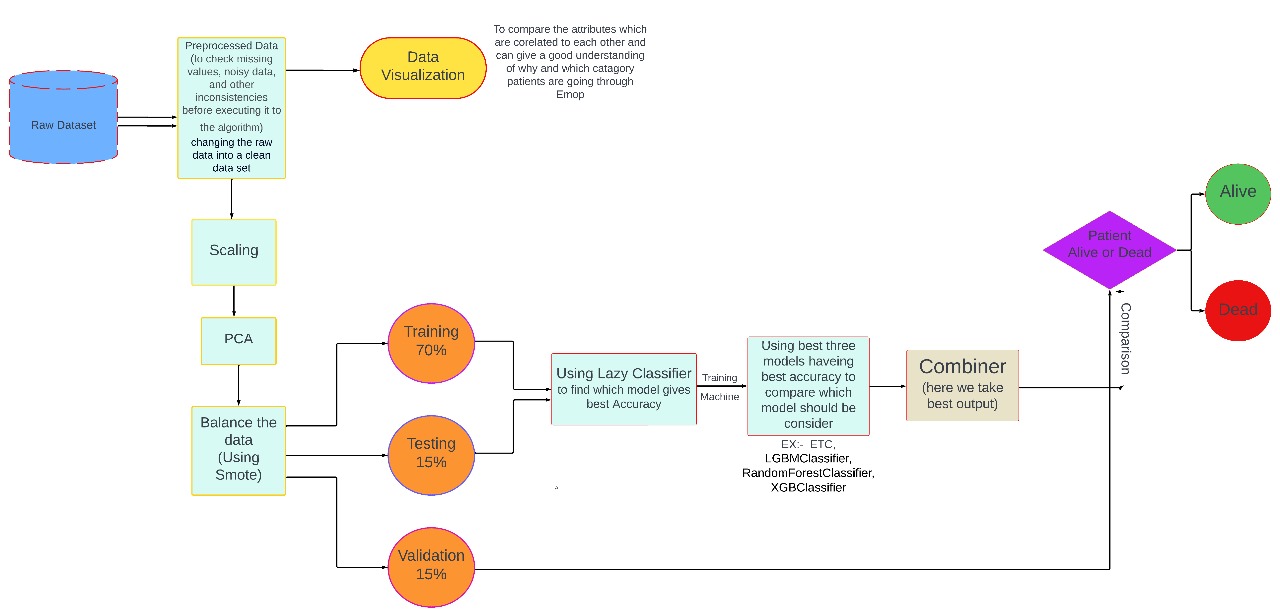


Figure 1.1

**3.3 Model Description:**

LazyPredict serves as a pivotal component in the initial stages of model selection and evaluation. By employing LazyPredict, we have efficiently explored various machine learning algorithms without the need for extensive manual configuration

Lazy Predict is a Python library that automates the model selection and training process in machine learning.It provides a quick way to evaluate and compare multiple machine learning models without the need for extensive manual tuning.The library takes a "lazy" approach by using default hyperparameters for each algorithm and provides a baseline performance comparison.Lazy Predict will automatically evaluate a variety of classifiers on your dataset and provide a summary of their performance metrics.

The predictive model for assessing perioperative risk based on preoperative data employs various machine learning algorithms trained on the VitalDB dataset. The model utilizes an ensemble of classifiers, including ExtraTreesClassifier, LGBMClassifier, RandomForestClassifier.

The ExtraTreesClassifier, a variant of decision tree algorithms, is utilized as one of the primary classifiers. It constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or the mean prediction (regression) of individual trees. This classifier is well-suited for handling imbalanced datasets and capturing complex relationships between features and target labels. Its ability to mitigate overfitting and provide robust predictions makes it a valuable component of the ensemble. It is an ensemble learning method that belongs to the family of decision tree algorithms. It builds multiple decision trees on random subsets of the training data and combines their predictions through averaging or voting. Unlike Random Forest, it selects random thresholds for each feature rather than searching for the best possible thresholds. This randomness in selecting splits often results in faster training times compared to traditional decision trees. It is effective for both classification and regression tasks and is known for its robustness to overfitting.

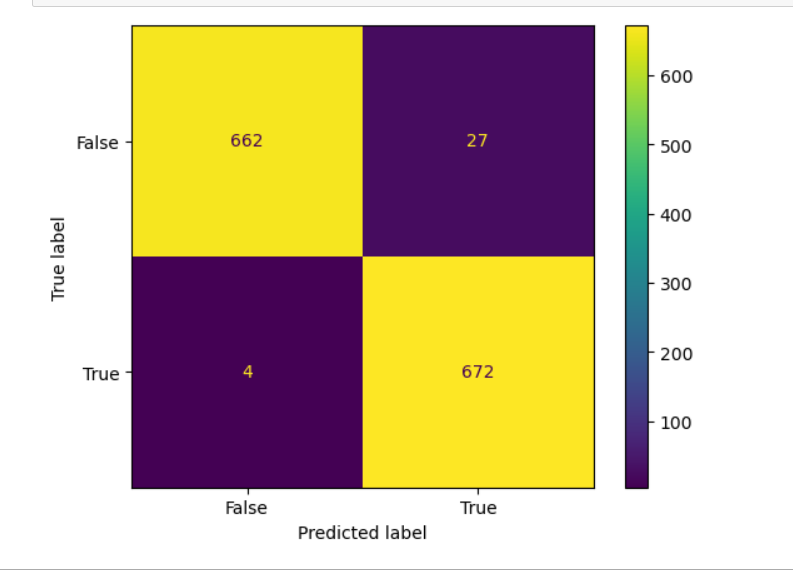


Figure 3.1

LGBMClassifier (LightGBM):

LGBMClassifier is a gradient boosting framework that uses tree-based learning algorithms. It is designed for distributed and efficient training of large-scale datasets.

LightGBM focuses on leaf-wise tree growth, where it grows the tree level-wise rather than level-wise like traditional methods. It achieves faster training speeds and lower memory usage compared to other gradient boosting frameworks. LightGBM supports parallel and GPU learning, making it suitable for large datasets and high-dimensional feature spaces.

It often provides better accuracy with less computation time compared to other boosting algorithms.

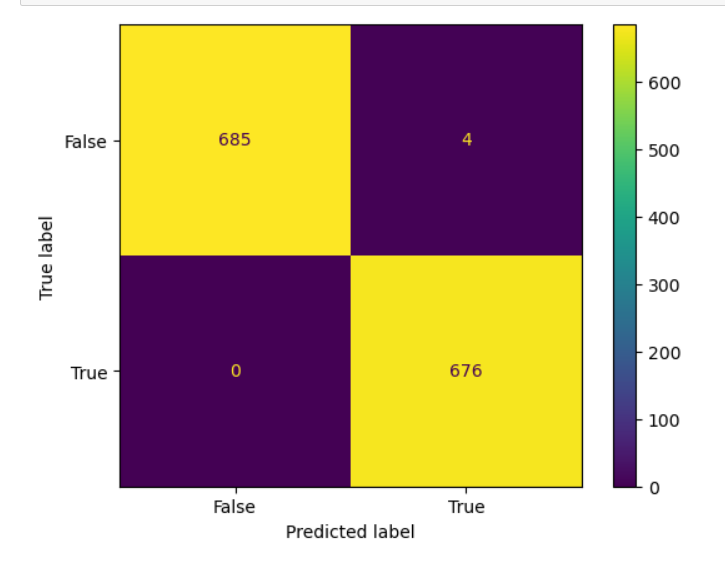


Figure 3.2

RandomForestClassifier:

RandomForestClassifier is another ensemble learning method based on decision trees.

It builds multiple decision trees using bootstrap samples of the training data and random feature subsets.The final prediction is made by averaging or voting over the predictions of individualtrees.Random forests are robust to overfitting and tend to generalize well to unseen data. They can handle high-dimensional feature spaces and are effective for both classification and regression tasks. RandomForestClassifier offers good performance out-of-the-box with minimal hyperparameter tuning

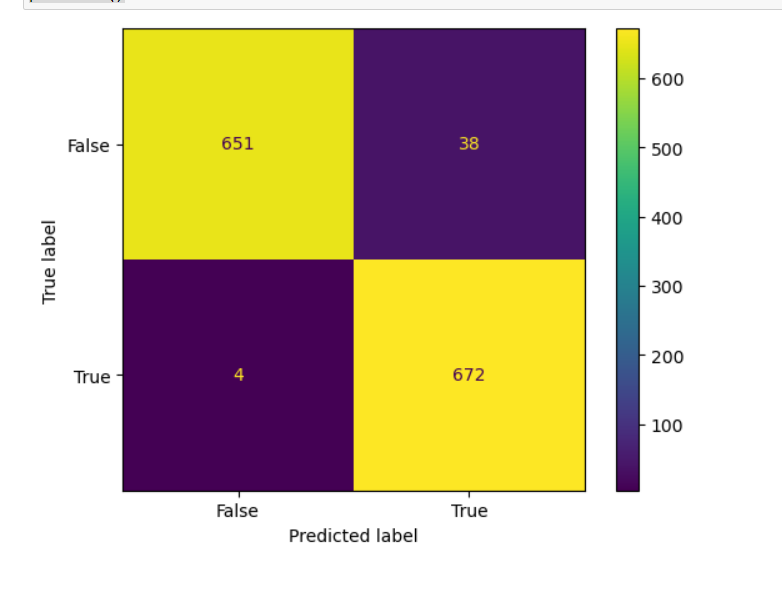


Figure 3.3

**Chapter 4**

**Implementation**

1. Use of Python Platform for writing the code with **scikit-learn**
2. Hardware and Software Configuration:

Hardware Configuration:

* + CPU: 8 GB RAM, Quad core processor
  + GPU: 16GB RAM **Nvidia's GTX 1080Ti**

Software Required:

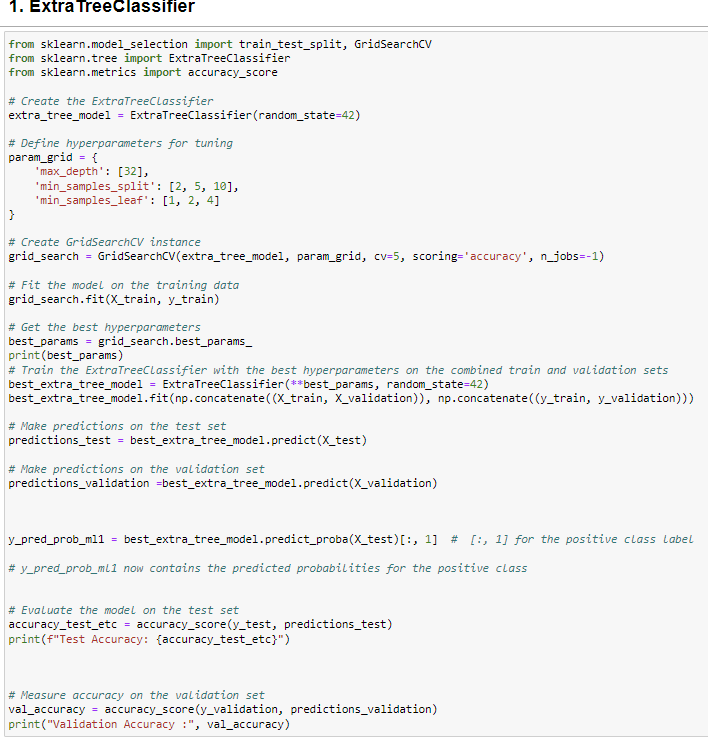
* + **Anaconda**: It is a package management software with free and open-source distribution of the Python and R programming language for scientific computations (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify deployment.
  + **Jupyter Notebook**:

Jupyter is a web-based interactive development environment for Jupyter notebooks, code, and data.

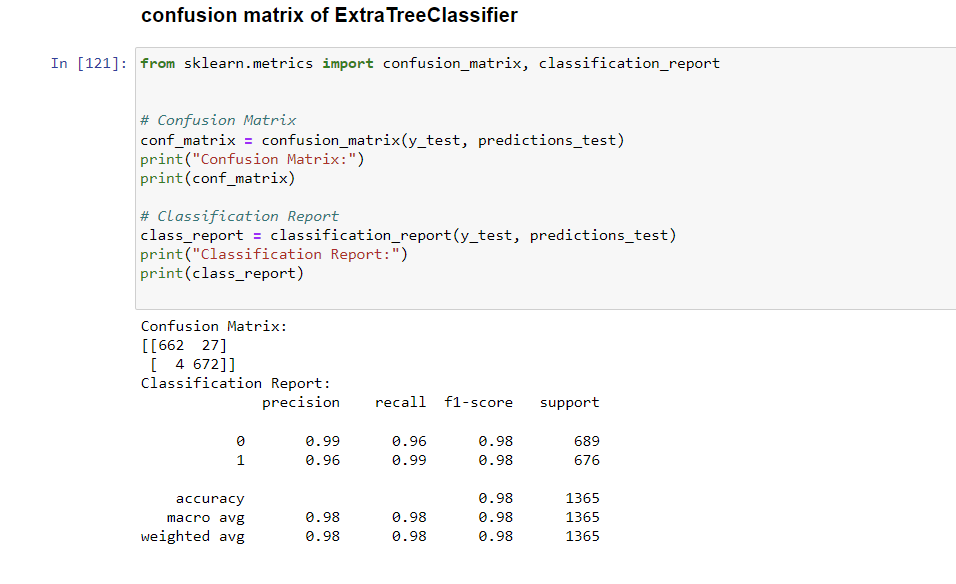
Jupyter is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning.

Jupyter is extensible and modular: write plugins that add new components and integrate with existing ones.

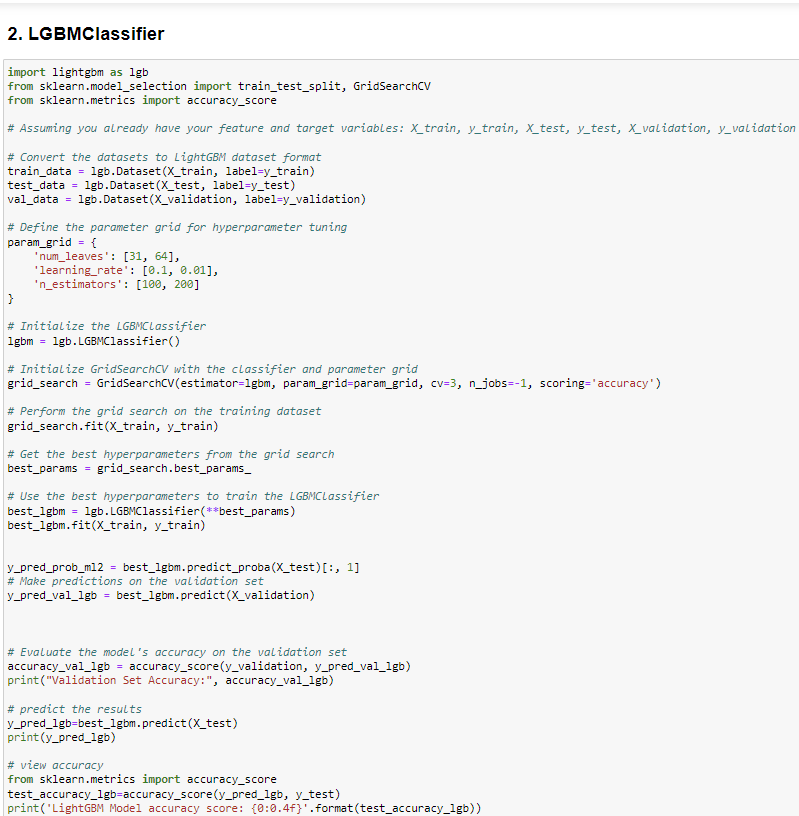
**1.ExtraTreeClassifier:**

****

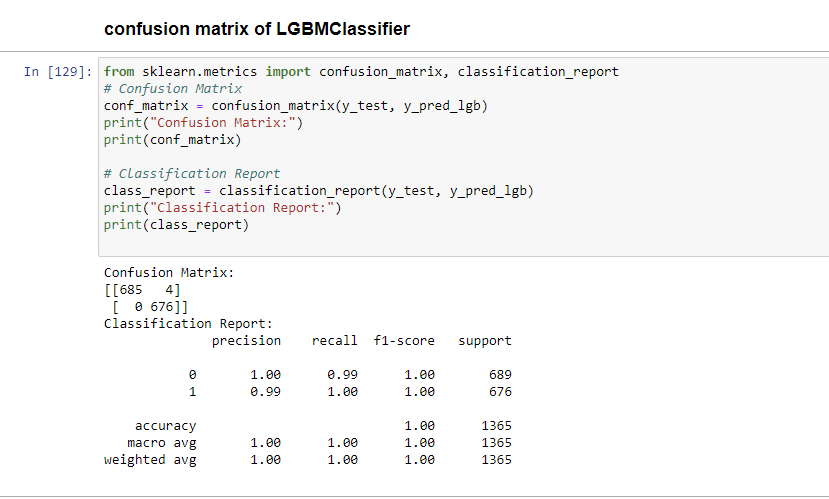
**Model Summary –**

****

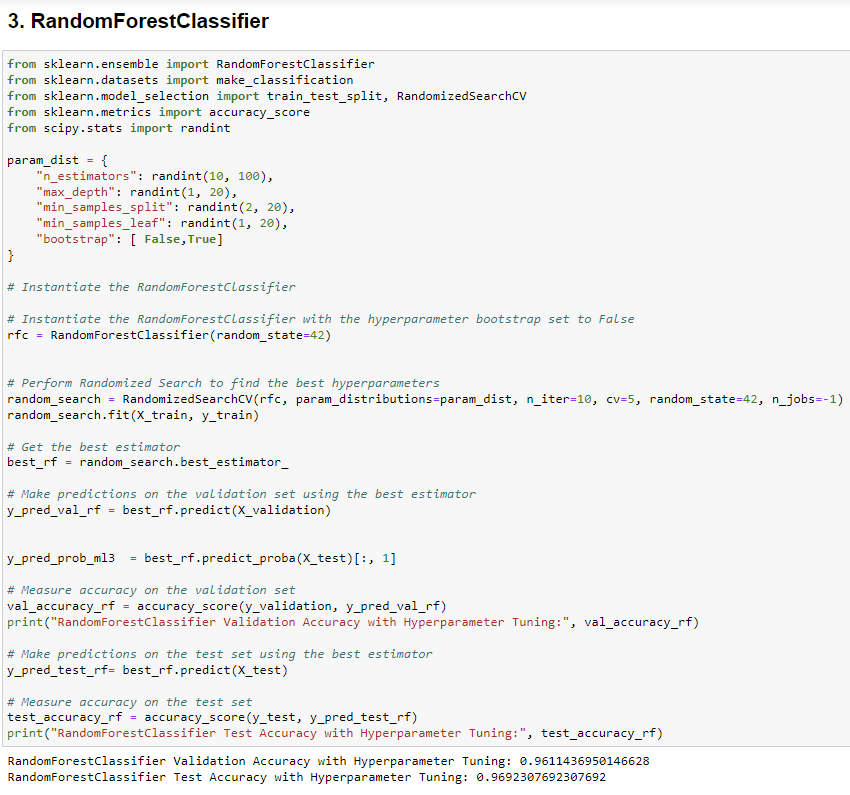
**2. LGBMClassifier:**

****

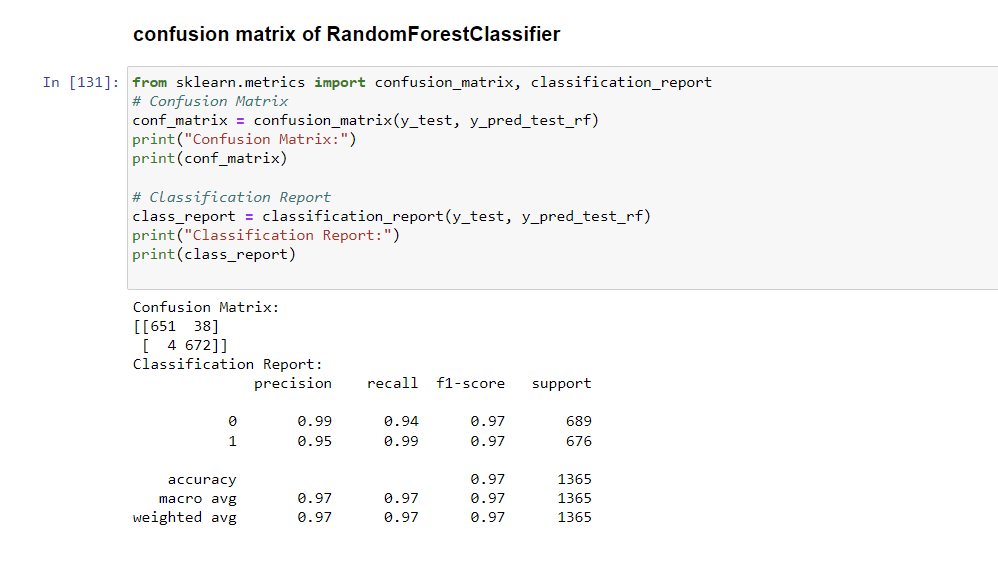
**Model Summary:**

****

1. **RandomForestClassifier :**

****

**Model Summary:**

****

**Chapter 5**

**Results**

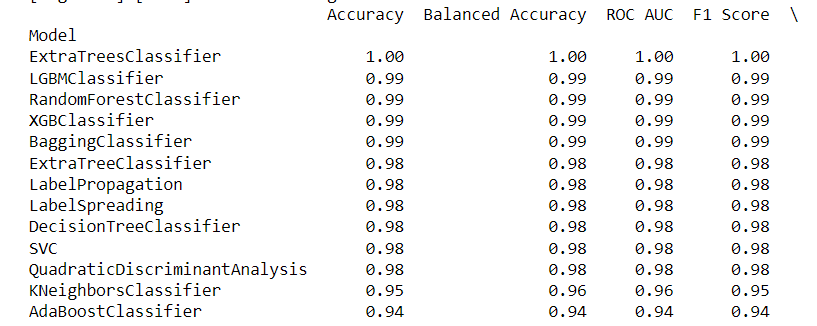
## ROC-AUC curve:

## 

**ROC\_AUC Score –**

|  |  |
| --- | --- |
| **ML\_Model** | **ROC\_AUC SCORE** |
| Extra Tree Classifier | 0.9772 |
| LGBM Classifier | 0.9971 |
| RandomForest Classifier | 0.9692 |

**Accuracy :**



**Chapter 6**

**Conclusion**

**6.1 Conclusion**

* The preoperative prediction of Machine Learning (ML) models is crucial for enhancing surgical outcomes and patient safety. These models underscore the significance of accurate information dissemination to facilitate proactive planning and minimize surgical risks. Further research in this area is imperative to refine predictive algorithms, optimize preoperative protocols, and ultimately improve patient care and surgical outcomes. It is essential to prioritize data transparency and dissemination of validated findings while refraining from spreading unverified information to maintain trust and confidence in preoperative prediction models.

**Chapter 7**

**References**

[1] Shaziz Hafiz, Nicholas Peter Lees “Coordinating perioperative care for the 'high risk' general surgical patient using risk prediction scoring”

[2] Joanna Abraham, Brian Bartek, Alicia Meng, S. Avidan “Integrating machine learning predictions for perioperative risk management”

[3] Martin Graeßner, Bettina Jungwirth, Elke Frank, Stefan Josef Schaller “Enabling personalized perioperative risk prediction by using a machine‑learning model based on preoperative data”

[4] Chiew Calvin, Liu Nan, Wong Ting Hway, Sim Yilin, Abdullah Hairil “Utilizing Machine Learning Methods for Preoperative Prediction of Postsurgical Mortality and Intensive Care Unit Admission”

[5] Hafsa Habehh1 and Suril Gohel1 “Machine Learning in Healthcare”

[6] Saifur Rahman,Jingwen Zhou ,James Jin Kan “A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges”

[6] <https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c>

[7] <https://github.com/brianhill11/PreopMortalityPrediction>

[8] <https://towardsdatascience.com/what-is-the-c4-5-algorithm-and-how-does-it-work-2b971a9e7db0>

[9] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7668340/>