**SUMMER TRAINING/INTERNSHIP**

**PROJECT REPORT**

(Term June-July 2025)

## (Risk Factors Analysis and CVD Prediction Using Health Data )

Submitted by : Uday Kiran

Under the Guidance of : Ms. Sandeep Kaur

# School of Computer Science and Engineering

# ACKNOWLEDGEMENT:

I would like to express my heartfelt thanks to my mentor, Ms. Sandeep Kaur for guiding me through this project. I'm also grateful to the School of Computer Science and Engineering at LPU for offering this platform and resources. Special thanks to Mendeley Data for providing the health dataset that was crucial for this project.

TABLE OF CONTENTS

1. Introduction

2. Training Overview

3. Project Details

4. Implementation

5. Results and Discussion

6. Conclusion

7. Important Links

# CHAPTER 1: INTRODUCTION

## Company Profile

The project utilized a public health database from Mendeley Data known as CAIR-CVD-2025: An Extensive Cardiovascular Disease Risk Assessment Dataset from Bangladesh. This project replicates the authentic analytical evaluation of health data that analysts would probably conduct :

* Hospital research departments
* Health tech startups
* Pharmaceutical research teams
* Public health organizations

## Dataset Link : <https://data.mendeley.com/datasets/d9scg7j8fp/1>

## Overview of Training Domain

The domain of this training was Data Science and Machine Learning. It focused on real-world healthcare data analysis, exploratory data analysis, and predictive modeling.

## Objective of the Project

The overall objective disassembled into specific goals:

* Identification of key biomarkers for CVD risk
* Development of a classification system for risk stratification
* Creation of interpretable results for potential clinical application

# CHAPTER 2: TRAINING OVERVIEW

## Tools & Technologies Used

1. **Python Stack:**
   * Pandas: For structured data manipulation
   * NumPy: Numerical computing foundations
   * Matplotlib/Seaborn: Advanced visualization capabilities
   * Scikit-learn: Machine learning pipeline implementation
   * SciPy: Statistical testing and advanced computations
2. **Power BI:**
   * Used for creating interactive, business-ready dashboards
   * Enabled drill-down capabilities for different demographic slices
3. **Jupyter Notebook:**
   * Served as the primary development environment
   * Allowed for literate programming with markdown documentation

## Areas Covered During Training

1. **Data Cleaning:**
   * Handling missing data (imputation strategies)
   * Outlier detection and treatment
   * Data type standardization
2. **Exploratory Data Analysis:**
   * Univariate analysis of each feature
   * Bivariate analysis of feature relationships
   * Multivariate pattern recognition
3. **Hypothesis Testing:**
   * Proper application of Z-test for large sample sizes
   * Assumption validation (normality, independence)
   * Interpretation of p-values in medical context
4. **Model Development:**
   * Feature selection process
   * Hyperparameter tuning
   * Cross-validation strategies

## Weekly Work Summary

**Week 1: Data Exploration**

* Initial data quality assessment
* Development of data cleaning pipelines
* Creation of validation subsets

**Week 2: Advanced Analytics**

* Statistical testing design
* Visualization matrix creation
* Correlation network analysis

**Week 3: Modelling**

* Algorithm selection process
* Baseline model establishment
* Performance metric selection

**Week 4: Visualization & Deployment**

* Dashboard user experience design
* Storytelling with data
* Production considerations

# CHAPTER 3: PROJECT DETAILS

## Title of the Project

CVD Risk Factor Analysis and Prediction

## Problem Definition

Cardiovascular disease, globally accounting for ~32% of all deaths (WHO), is preventable before it occurs or before results are seen. Known as prediction modeling, this technology could lead to:

• Decreased emergency rounds through preventive care

• Personalized care

• Reduction of strain on healthcare systems

## Scope and Objectives

1. **Data Understanding:**
   * Cohort analysis by age groups
   * Temporal trends in biomarkers
   * Comorbidity patterns
2. **Modelling Development:**
   * Feature importance analysis
   * Decision boundary visualization
   * Probability calibration
3. **Visual Analytics:**
   * Interactive risk score calculators
   * Population health snapshots
   * Comparative analyses by demographic factors

## System Requirements

Operating System: Windows 10 or above  
Software: Jupiter Notebook, Power BI Desktop  
Libraries: pandas, NumPy, seaborn, matplotlib, scikit-learn, scipy

## Architecture Diagram

Data Input → Preprocessing → EDA → Modeling → Evaluation → Dashboard

CHAPTER 4: IMPLEMENTATION

## Tools Used

• Jupiter Notebook (Python)  
• Power BI Desktop

## Methodology

1. Descriptive Statistics:

Used .info() and .describe() to understand data types, summary statistics, and detect anomalies.

2. Missing Value Analysis:

Identified missing values column-wise and handled them using median (for numerical) and mode (for categorical).

3. Univariate Analysis:

Histograms to explore the distribution of variables like BMI, BP, and Age

Boxplots to detect outliers in numerical columns

Count plots for visualizing the frequency of categories (e.g., gender, smoker status)

4. Bivariate Analysis:

Scatterplot to study the relationship between BMI and CVD Risk Score

Boxplots to compare cholesterol across smoker vs non-smoker groups

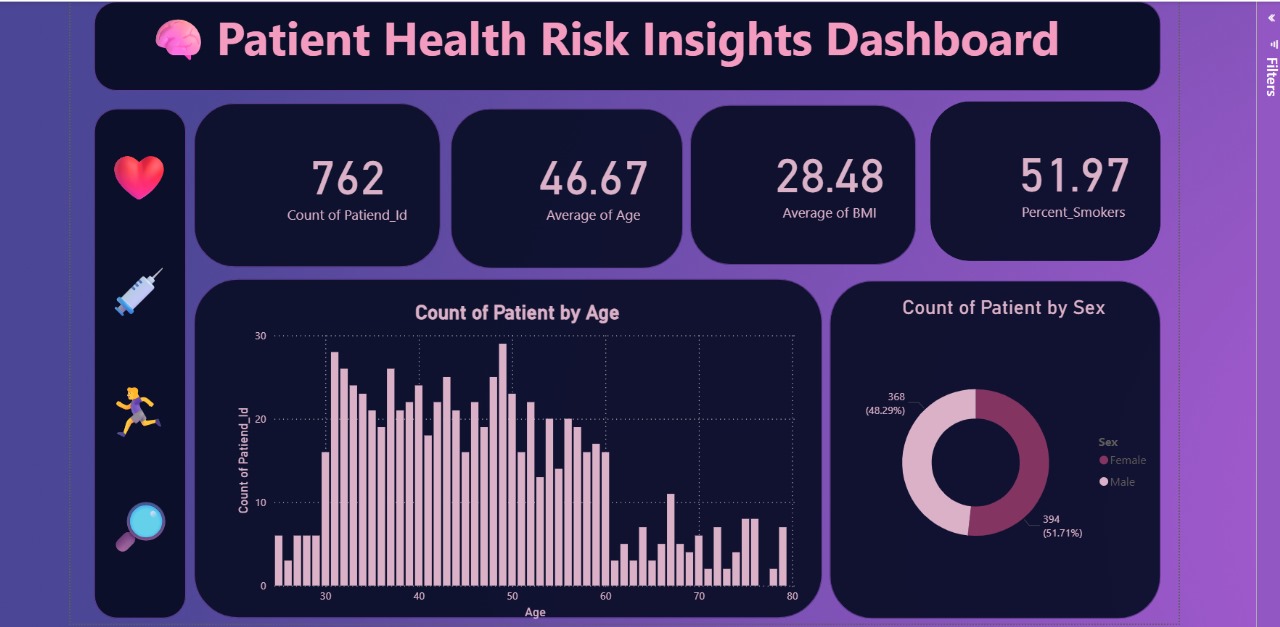
Z-test to statistically validate group differences

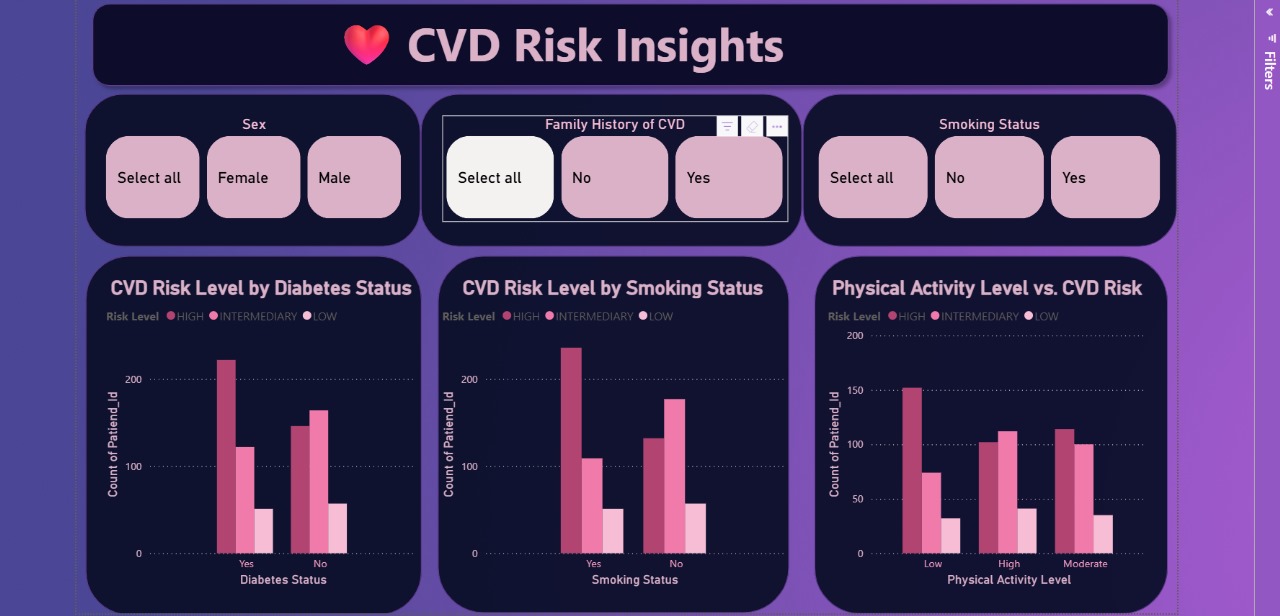
5. Correlation Analysis:

Created a heatmap to understand relationships between numerical variables

Helped identify potential predictors for modeling

## Modules / Screenshots

Screenshots from the notebook and Power BI dashboard should be added here to visually represent the results.



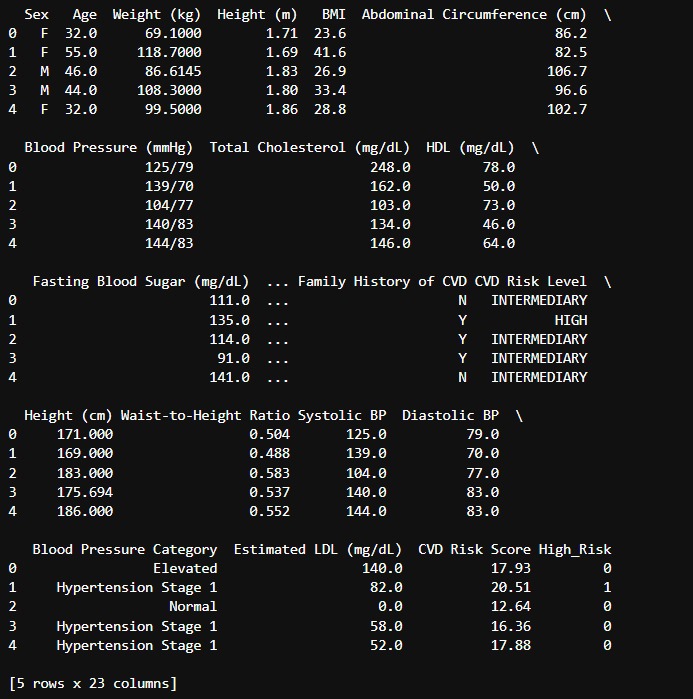
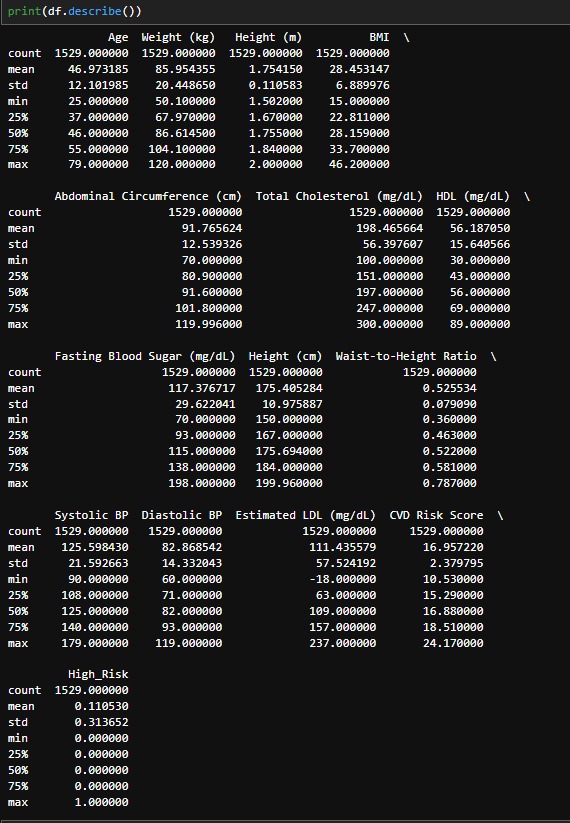
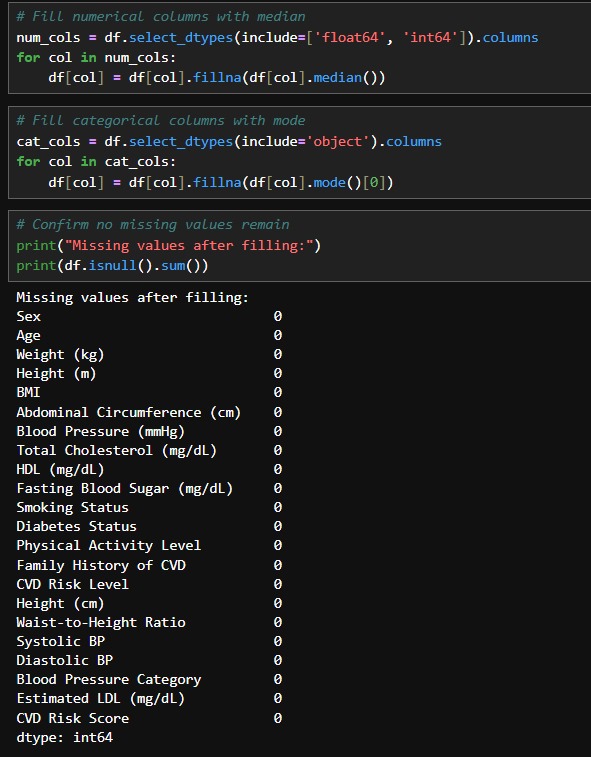
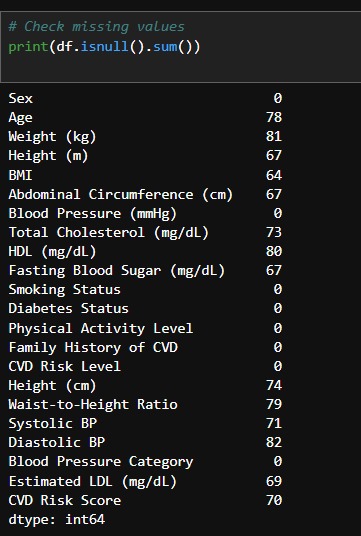
## 

## 



## Code Snippets and Outputs :

## 



# 

# 

# 

# CHAPTER 5: RESULTS AND DISCUSSION

## Output / Report

1. **Clinical Significance:**
   * Translation of odds ratios to clinical practice
   * Risk stratification thresholds
   * Potential impact on treatment protocols
2. **Model Performance:**
   * Precision-recall tradeoffs in medical context
   * Receiver Operating Characteristic analysis
   * Decision curve analysis for clinical utility
3. **Visual Analytics:**
   * Heatmap interpretation guidelines
   * Interactive filtering capabilities
   * Anomaly detection visualization

## Challenges Faced

1. **Data Quality:**
   * Implemented data validation framework
   * Developed automated quality checks
   * Created documentation for data limitations
2. **Class Imbalance:**
   * Tested various resampling techniques
   * Evaluated threshold-moving approaches
   * Implemented cost-sensitive learning
3. **Visual Storytelling:**
   * Developed user personas for dashboards
   * Conducted iterative usability testing
   * Implemented accessibility features

## Learnings :

• Real-world dataset cleaning and preprocessing

• Application of statistical hypothesis testing

• Evaluation of machine learning models

• Data storytelling with Power BI dashboards

# CHAPTER 6: CONCLUSION

This project highlighted how data science techniques can uncover insights from health data and assist in early detection of cardiovascular risk. The end-to-end process included data cleaning, analysis, modeling, evaluation, and visualization. The experience enhanced practical understanding of healthcare analytics and predictive modeling.

**Broader Implications:**

1. **Clinical Applications:**
   * Potential for integration with EHR systems
   * Use in remote patient monitoring
   * Application in population health management
2. **Future Directions:**
   * Incorporation of time-series data
   * Integration with wearable device data
   * Development of real-time monitoring systems
3. **Professional Development:**
   * Gained competency in full analytics pipeline
   * Developed critical thinking for medical AI
   * Enhanced collaboration skills through stakeholder communication