Annexure-I

CVD Risk Factor Analysis and Prediction

LPU SKILL DEVELOPMENT

A training report

Submitted in partial fulfilment of the requirements for the award of degree of

Bachelors of Technology

(Computer science and Engineering)

Submitted to

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA, PUNJAB

From 10/06/25 to 18/07/25

SUBMITTED BY

SITARAM MANEPELLI

12308166

M.N.S. Sita Ram

Annexure-II: Student Declaration

To whom so ever it may concern

I, **Sitaram Manepalli, 12308166,** hereby declare that the work done by me on "**CVD Risk Factor Analysis and Prediction**" from **June, 2025** to **July, 2025**, is a record of original work for the partial fulfilment of the requirements for the award of the degree, **Bachelors of Technology (Computer science and Engineering).**

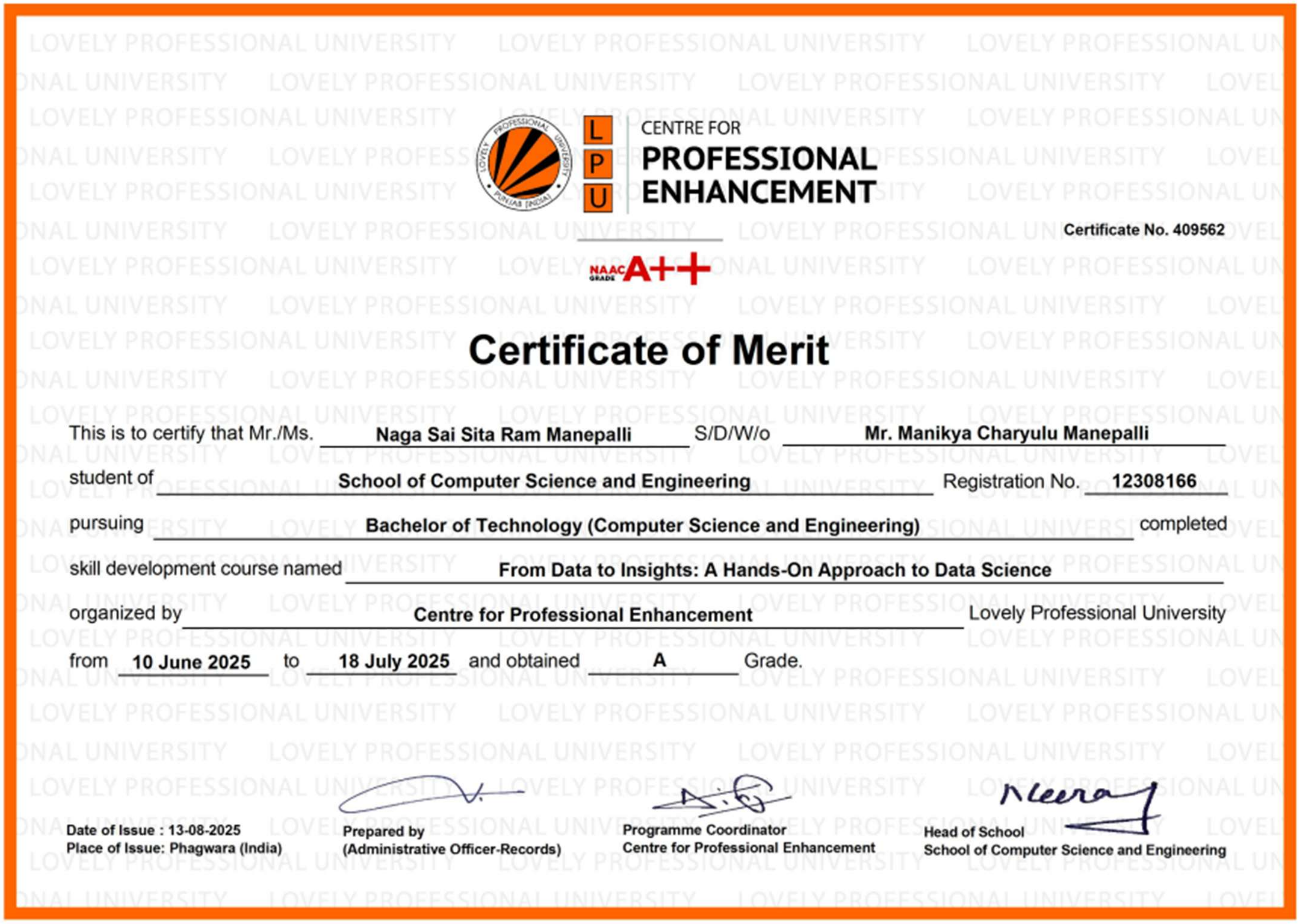
Uday Kiran (12315696)

M.N.S.Sita Ram

Signature of the student

Dated: 31/08/2025

Certificate



LPU | CENTRE FOR PROFESSIONAL ENHANCEMENT

NAAC GRADE A++

Certificate No. 409562

Certificate of Merit

This is to certify that Mr./Ms. Naga Sai Sita Ram Manepalli S/D/W/o Mr. Manikya Charyulu Manepalli

student of School of Computer Science and Engineering Registration No. 12308166

pursuing Bachelor of Technology (Computer Science and Engineering) completed

skill development course named From Data to Insights: A Hands-On Approach to Data Science

organized by Centre for Professional Enhancement Lovely Professional University

from 10 June 2025 to 18 July 2025 and obtained A Grade.

Date of Issue : 13-08-2025
Place of Issue: Phagwara (India)

Prepared by
(Administrative Officer-Records)

Programme Coordinator
Centre for Professional Enhancement

Neeraj
Head of School
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 and LPU skill development for offering this platform and resources. Special thanks to Mendeley Data for providing the health dataset that was crucial for this project

M.N.S. Sita Ram

TABLE OF CONTENTS

Chapter 1: Introduction

- Company profile
- Overview of training domain
- Objective of the project

Chapter 2: Training Overview

- Tools & technologies used
- Areas covered during training
- Daily/weekly work summary

Chapter 3: Project Details

- Title of the project
- Problem definition
- Scope and objectives
- System Requirements
- Architecture Diagram (if any)
- Data flow / UML Diagrams

Chapter 4: Implementation

- Tools used
- Methodology
- Modules / Screenshots
- Code snippets (if needed)

Chapter 5: Results and Discussion

- Output / Report
- Challenges faced
- Learnings

Chapter 6: Conclusion

- Summary

Chapter 6: Refrences

1. INTRODUCTION

1.1 Company Profile

This project was conducted as an independent data analysis initiative, simulating the work of a data analyst in a public health or research organization. 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 analytical workflow that data analysts would typically perform in settings such as:

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

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

1.2 Overview of Training Domain

The domain of this training was **Data Science and Machine Learning**, with a specific focus on healthcare analytics. The training emphasized real-world application, covering the entire data lifecycle from data cleaning and exploratory data analysis to predictive modeling and results visualization.

1.3 Objective of the Project

The primary objective of this project was to analyze a comprehensive health dataset to identify key risk factors associated with Cardiovascular Disease (CVD) and develop a predictive model for risk assessment.

The specific goals were:

1. To conduct a thorough Exploratory Data Analysis (EDA) to understand the dataset and uncover relationships between various health metrics.
2. To identify the most significant biomarkers and lifestyle factors contributing to CVD risk.
3. To build and evaluate a machine learning model capable of classifying patients into different risk levels.
4. To create an interactive and user-friendly Power BI dashboard to visualize the findings and make the data accessible to a non-technical audience.

1.4 Role and Profile

During this project, my role was to design, develop the dashboard in the Power Bi .

Data Cleaning: Sourcing, cleaning, and transforming the raw dataset into a usable format.
Power Bi : Translating analytical findings into a clear and interactive visual dashboard using Power Bi. Interpreting results and presenting insights.

This project was undertaken as part of the
From Data to Decisions: A Hands-On Approach to Data Science *from* ***LPU SKILL DEVELOPMENT***, completed between **June 10, 2025 to July 18, 2025**, comprising **39 days** of guided learning

Chapter 2: Training Overview

2.1 Tools & Technologies Used

- **Python Stack:**
    - **Pandas & NumPy:** For data manipulation, cleaning, and numerical computation.
    - **Matplotlib & Seaborn:** For static data visualization to generate plots like histograms, boxplots, and heatmaps.
    - **Scikit-learn:** For implementing the machine learning pipeline, including data splitting and logistic regression modeling.
    - **SciPy / Statsmodels:** For performing statistical hypothesis testing (Z-test and T-test).
- **Power BI:**
    - Used for creating an interactive, business-ready dashboard for dynamic data exploration.
    - Enabled drill-down capabilities for analyzing different demographic and health segments.
- **Jupyter Notebook:**
    - Served as the primary Integrated Development Environment (IDE) for coding, analysis, and documenting the workflow.

2.2 Areas Covered During Training

- **Data Cleaning:** Handling missing data through imputation (median for numerical, mode for categorical), outlier detection using Z-scores, and ensuring data type consistency.
- **Exploratory Data Analysis (EDA):** Performing univariate analysis (histograms, boxplots) to understand distributions, bivariate analysis (scatter plots) to explore relationships, and multivariate analysis (correlation heatmap) to identify patterns.

- **Hypothesis Testing:** Applying Z-tests and T-tests to statistically validate the significance of differences between high-risk and low-risk patient groups for key features like BMI and Blood Pressure.
- **Model Development:** Implementing a full machine learning pipeline, including feature selection, data splitting (train/test), model training (Logistic Regression), and performance evaluation.
- **Data Visualization & Storytelling:** Designing an intuitive and insightful Power BI dashboard to communicate complex data findings effectively.

2.3 Weekly Work Summary

- **Week 1: Data Acquisition and Cleaning** ○ Downloaded and understood the CVD dataset. ○ Performed initial data quality assessment, identifying missing values and data types. ○ Developed and applied data cleaning pipelines, including imputation strategies for numerical and categorical columns.
- **Week 2: Exploratory Data Analysis & Hypothesis Testing**
  - ○ Conducted univariate and bivariate analysis to visualize data distributions and relationships. ○ Created a correlation matrix and heatmap to identify multicollinearity.
  - ○ Designed and executed Z-tests to statistically compare key metrics between risk groups.
- **Week 3: Machine Learning Modeling** ○ Selected features based on EDA findings and domain knowledge. ○ Prepared the data for modeling by splitting it into training and testing sets.

○ Trained a Logistic Regression model to predict a "High Risk" target variable. ○ Evaluated the model using accuracy, classification reports, and a confusion matrix.

- **Week 4: Dashboard Creation and Reporting** ○ Imported the cleaned dataset into Power BI. ○ Designed and built an interactive dashboard with multiple pages for different insights (Overall, Lifestyle, Blood Pressure). ○ Finalized the project report, summarizing the methodology, findings, and conclusions.

Chapter 3: Project Details

3.1 Title of the Project:

Risk Factors Analysis and CVD Prediction Using Health Data

3.2 Problem Definition

Cardiovascular disease (CVD) is a leading cause of mortality worldwide. Early identification of at-risk individuals is crucial for implementing preventive measures and reducing the burden on healthcare systems. This project addresses the need for an accessible and data-driven approach to:

1. Identify the key clinical and lifestyle factors that contribute most significantly to CVD risk.
2. Develop a reliable predictive model to classify individuals based on their risk profile.
3. Present these complex findings in a clear, interactive format that can be understood by healthcare professionals and patients alike.

3.3 Scope and Objectives

The project's scope covers the end-to-end process of data analysis, from raw data to actionable insights.

Objectives:

- **Data Understanding:** Perform a deep dive into the dataset to analyze patient demographics, clinical measurements (like cholesterol and blood pressure), and lifestyle factors (like smoking and physical activity).
- **Modeling Development:** Build a classification model that predicts whether a patient is at high risk for CVD, achieving a high level of accuracy (>90%).
- **Visual Analytics:** Create a multi-page Power BI dashboard that provides:

  ○ A high-level summary of patient health metrics. ○ Detailed insights into how lifestyle choices correlate with CVD risk. ○ An analysis of blood pressure and cholesterol levels across different patient segments.

3.4 System Requirements

- **Operating System:** Windows 10 or above
- **Software:** Jupyter Notebook, Power BI Desktop
- **Python Libraries:** pandas, numpy, seaborn, matplotlib, scikit-learn, scipy, statsmodels

3.5 Architecture Diagram

The project follows a standard data science workflow:

Data Input (CSV) → Data Preprocessing (Jupyter) → Exploratory Data Analysis (Jupyter) → Modeling & Evaluation (Jupyter) → Visualization (Power BI)

Chapter 4: Implementation 4.1 Tools Used

- **Jupyter Notebook (Python):** For all data cleaning, analysis, and modeling tasks.
- **Power BI Desktop:** For creating the interactive data visualization dashboard.

4.2 Methodology

1. **Descriptive Statistics:** Used .info() and .describe() in pandas to get an initial overview of the dataset's structure, data types, and summary statistics.

2. **Missing Value Analysis:** Identified missing values using .isnull().sum() and imputed them using the median for numerical columns and the mode for categorical columns to maintain data integrity.

3. **Univariate and Bivariate Analysis:**

   ○ Used histograms and boxplots to explore the distribution and identify outliers in variables like BMI, Blood Pressure, and Age. ○ Used a scatterplot to visualize the relationship between BMI and CVD Risk Score. ○ Generated a correlation heatmap to understand the linear relationships between all numerical variables.

4. **Hypothesis Testing:** Conducted Z-tests and T-tests to determine if there were statistically significant differences in key metrics (like BMI and Systolic BP) between high-risk and low-risk patient groups.

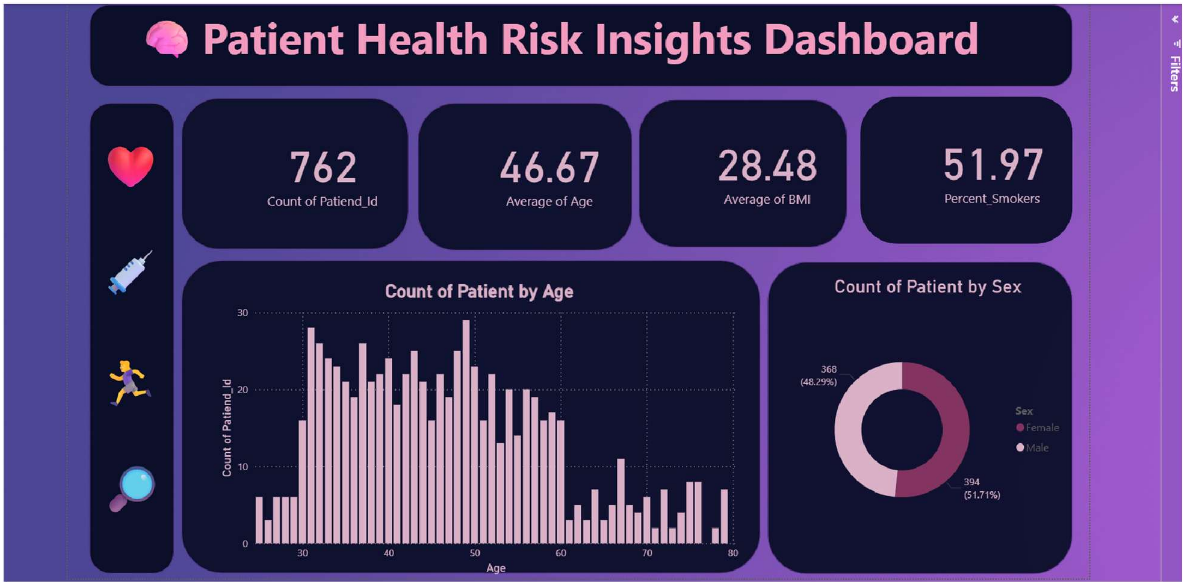
5. **Predictive Modeling:**

   ○ Engineered a binary target variable High_Risk where a CVD Risk Score of 20 or higher was classified as high risk.

○ Split the data into an 80% training set and a 20% test set. ○
Trained a Logistic Regression model on the training data. ○
Evaluated the model's performance on the unseen test data using accuracy, precision, recall, and F1-score.

4.3 Modules / Screenshots

Overview



Patient Health Risk Insights Dashboard

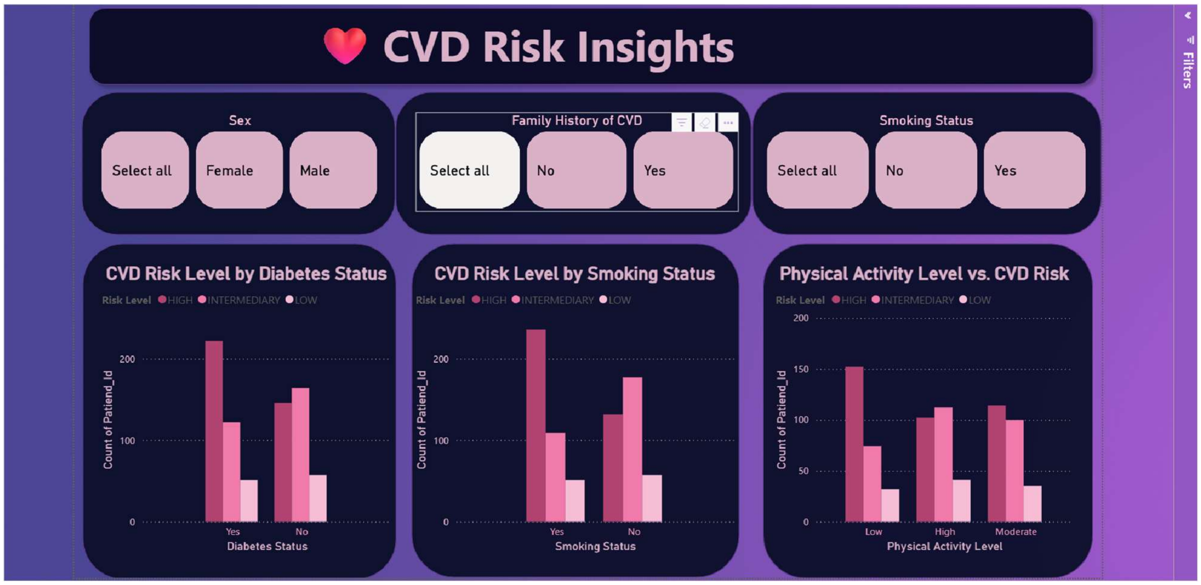
762
Count of Patiend_Id

46.67
Average of Age

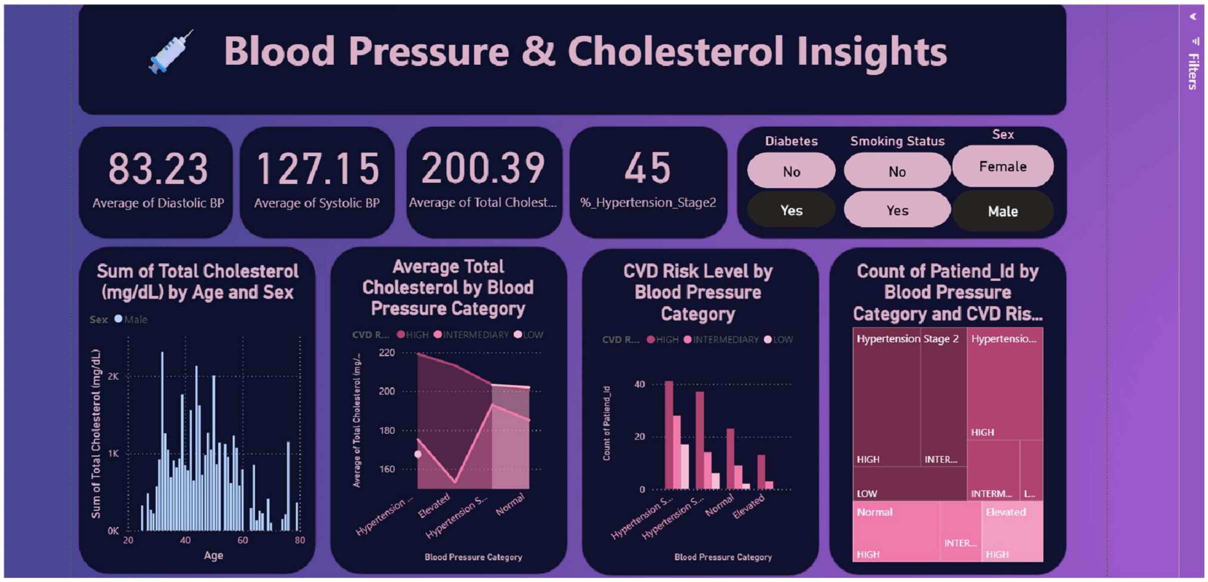
28.48
Average of BMI

51.97
Percent_Smokers

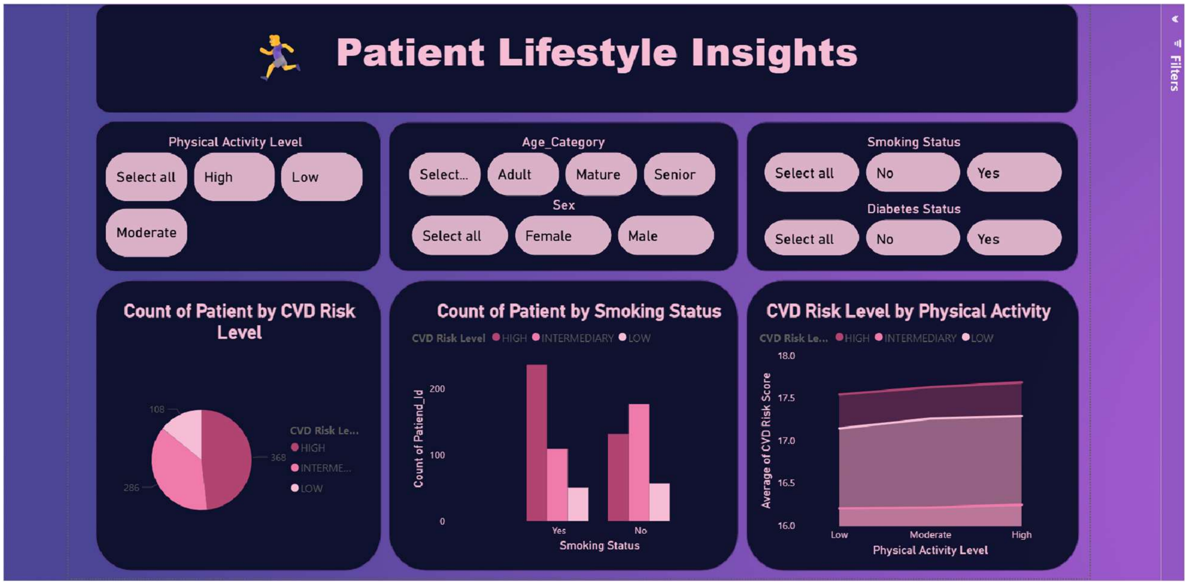
CVD Risk Insights



Blood Pressure & Cholesterol Overview



Lifestyle & Behavioral Risk Factors



Patient Lifestyle Insights

Physical Activity Level: Select all | High | Low | Moderate

Age_Category: Select... | Adult | Mature | Senior

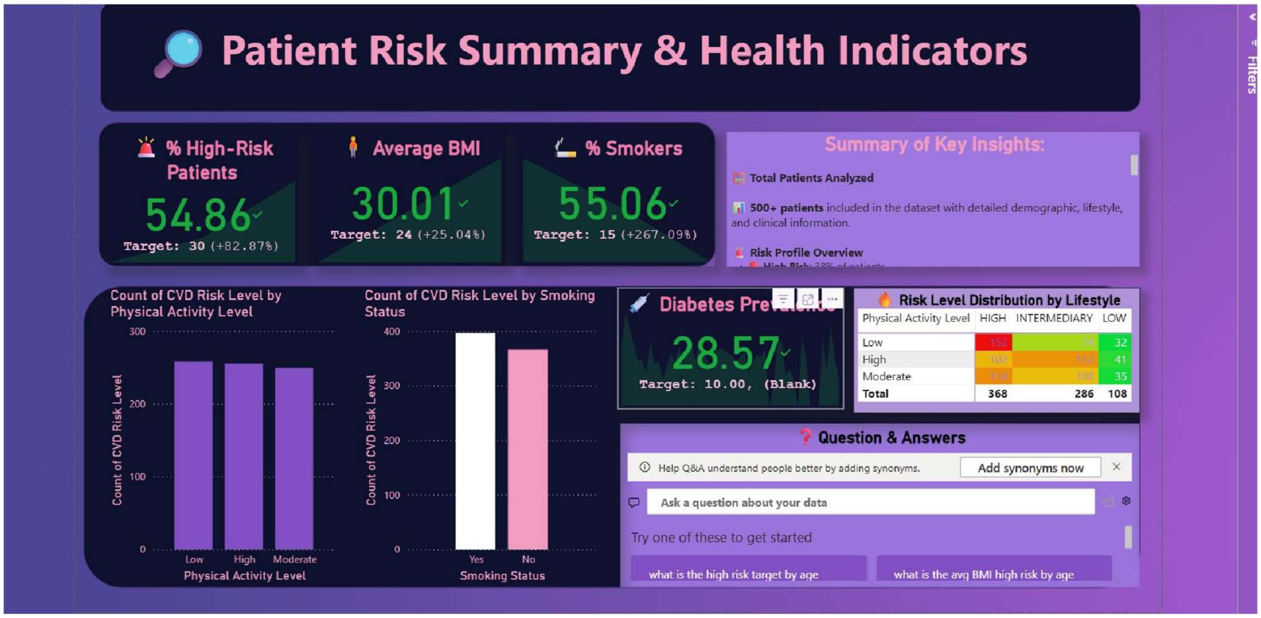
Sex: Select all | Female | Male

Smoking Status: Select all | No | Yes

Diabetes Status: Select all | No | Yes

Filters
Count of Patient by CVD Risk Level
108
368
286
CVD Risk Le...
● HIGH
● INTERME...
● LOW
Count of Patient by Smoking Status
CVD Risk Level ● HIGH ● INTERMEDIARY ● LOW
Count of Patiend_Id
200
100
0
Yes
No
Smoking Status
CVD Risk Level by Physical Activity
CVD Risk Le... ● HIGH ● INTERMEDIARY ● LOW
Average of CVD Risk Score
18.0
17.5
17.0
16.5
16.0
Low
Moderate
High
Physical Activity Level

Patient Risk Summary & Health Indicators



Patient Risk Summary & Health Indicators

% High-Risk Patients: 54.86 ✓ Target: 30 (+82.87%)

Average BMI: 30.01 ✓ Target: 24 (+25.04%)

% Smokers: 55.06 ✓ Target: 15 (+267.09%)

Summary of Key Insights:

Total Patients Analyzed

500+ patients included in the dataset with detailed demographic, lifestyle, and clinical information.

Risk Profile Overview

- High Risk: 38% of patients

Diabetes Prevalence: 28.57 ✓ Target: 10.00, (Blank)

Risk Level Distribution by Lifestyle

Physical Activity Level	HIGH	INTERMEDIARY	LOW
Low	152	74	32
High	102	112	41
Moderate	114	100	35
Total	**368**	**286**	**108**

Question & Answers

Help Q&A understand people better by adding synonyms. Add synonyms now ×

Ask a question about your data

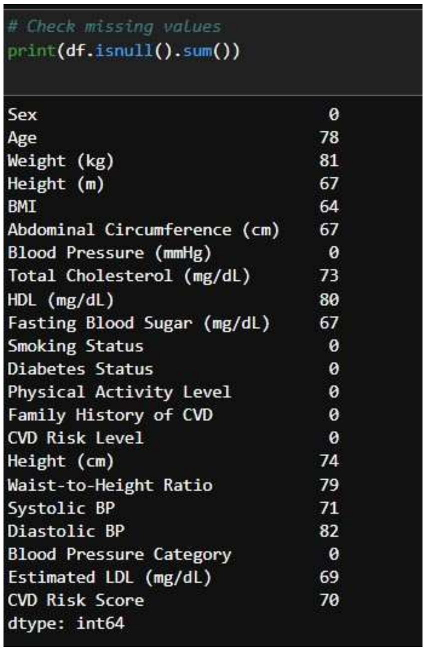
Try one of these to get started

what is the high risk target by age

what is the avg BMI high risk by age

4.4 Code Snippets

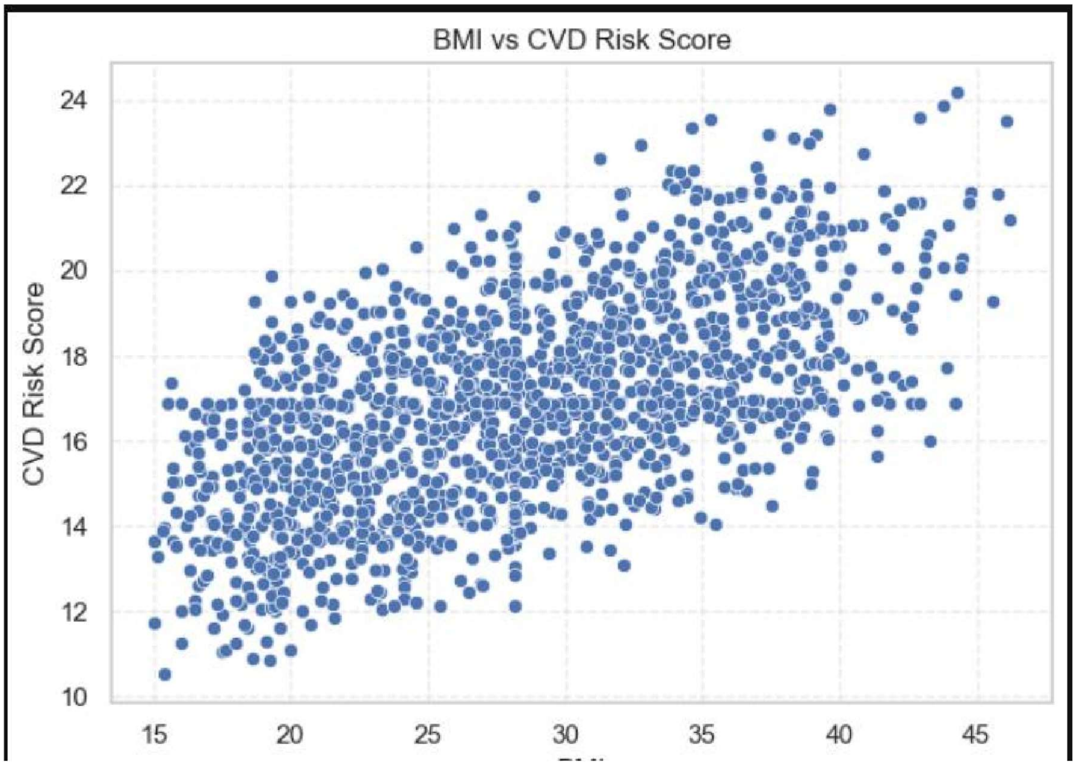
Missing values



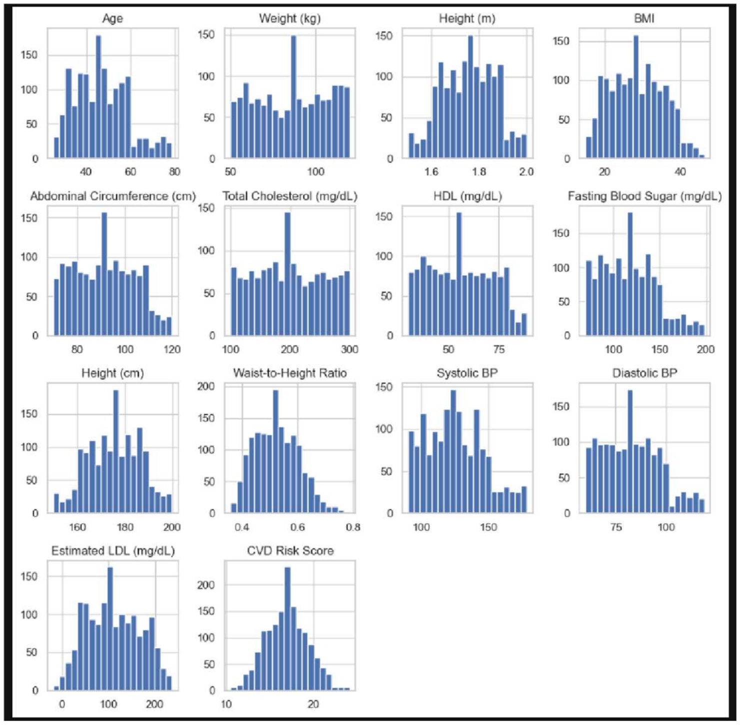
```
# Check missing values
print(df.isnull().sum())
```

```
Sex                            0
Age                           78
Weight (kg)                   81
Height (m)                    67
BMI                           64
Abdominal Circumference (cm)  67
Blood Pressure (mmHg)          0
Total Cholesterol (mg/dL)     73
HDL (mg/dL)                   80
Fasting Blood Sugar (mg/dL)   67
Smoking Status                 0
Diabetes Status                0
Physical Activity Level        0
Family History of CVD          0
CVD Risk Level                 0
Height (cm)                   74
Waist-to-Height Ratio         79
Systolic BP                   71
Diastolic BP                  82
Blood Pressure Category        0
Estimated LDL (mg/dL)         69
CVD Risk Score                70
dtype: int64
```

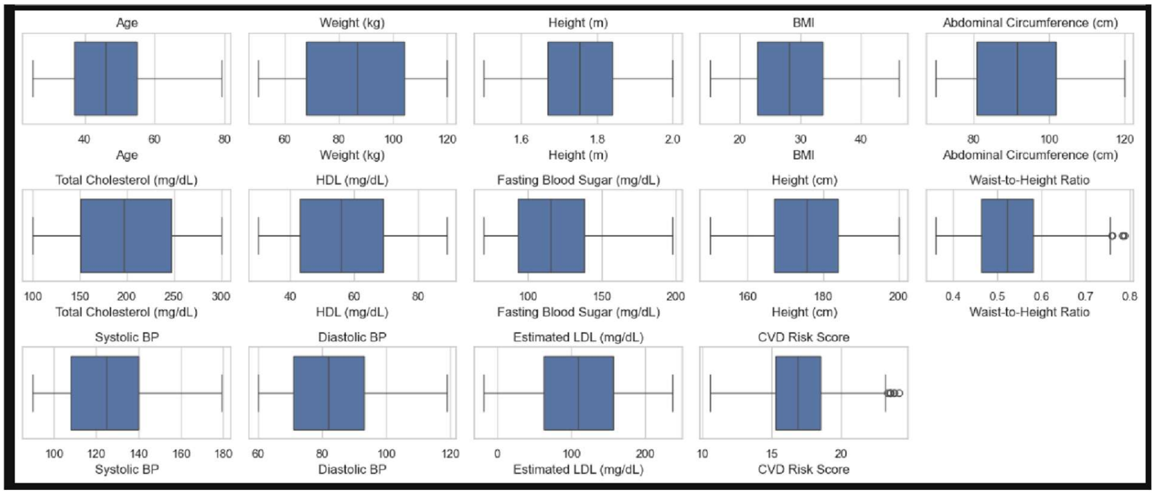
BMI VS CVD RISK SCORE



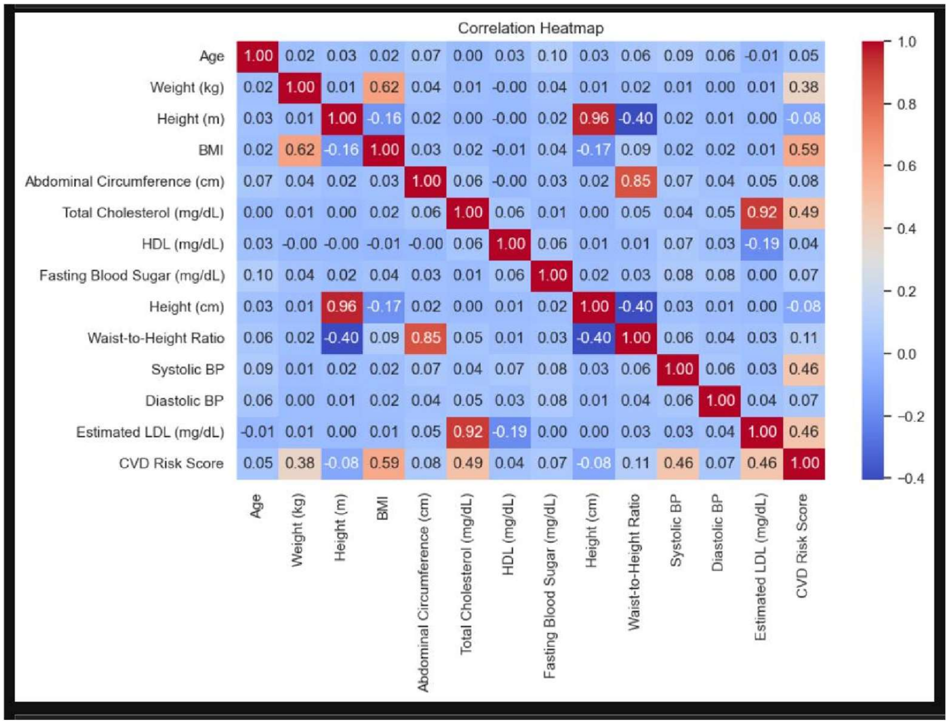
Histograms



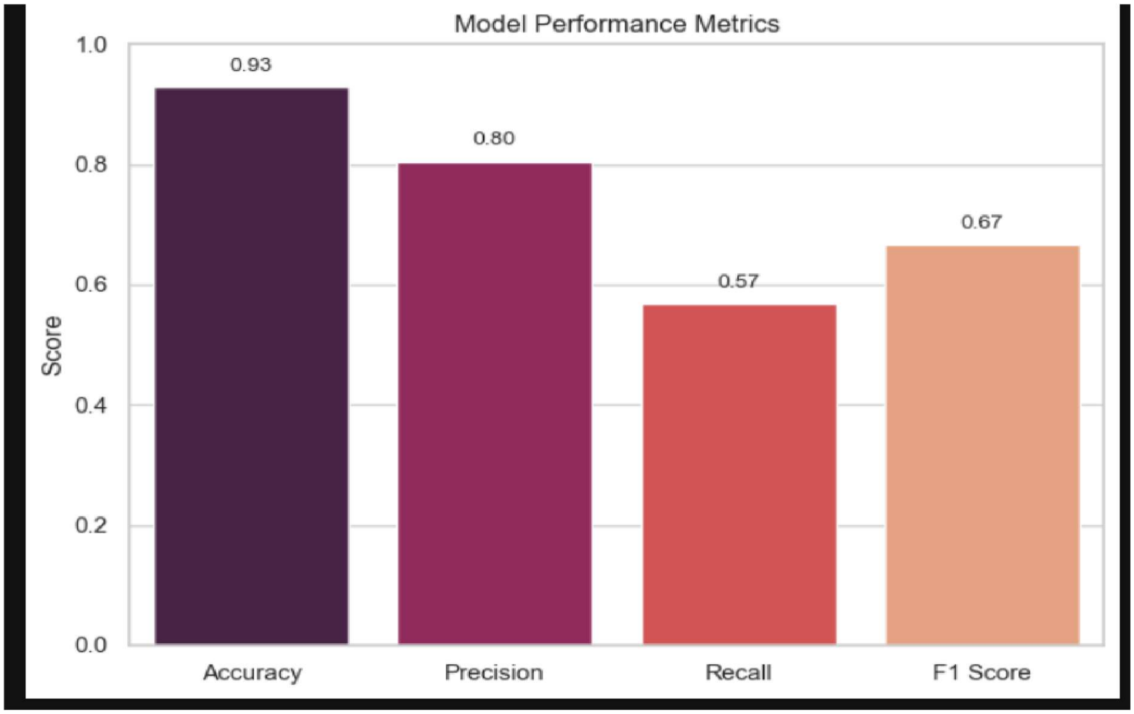
Boxplots



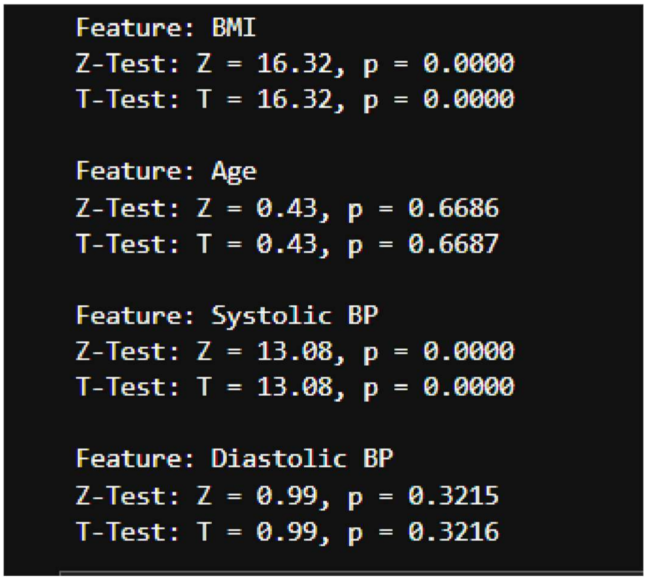
Correlation Heatmap



Model Performance Matrix



Hypothesis Testing: Z-Test & T-Test



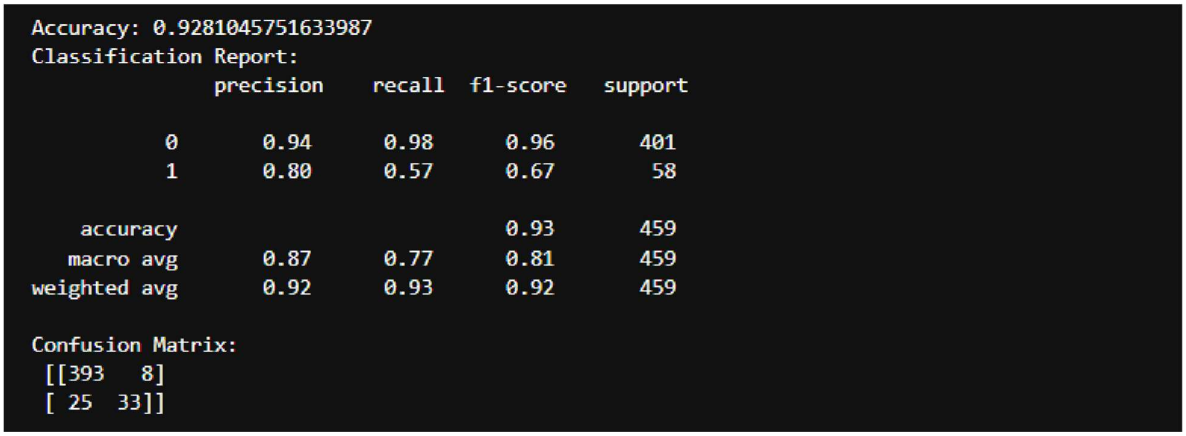
```
Feature: BMI
Z-Test: Z = 16.32, p = 0.0000
T-Test: T = 16.32, p = 0.0000

Feature: Age
Z-Test: Z = 0.43, p = 0.6686
T-Test: T = 0.43, p = 0.6687

Feature: Systolic BP
Z-Test: Z = 13.08, p = 0.0000
T-Test: T = 13.08, p = 0.0000

Feature: Diastolic BP
Z-Test: Z = 0.99, p = 0.3215
T-Test: T = 0.99, p = 0.3216
```

Logistic Regression (Classification)



```
Accuracy: 0.9281045751633987
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96       401
           1       0.80      0.57      0.67        58

    accuracy                           0.93       459
   macro avg       0.87      0.77      0.81       459
weighted avg       0.92      0.93      0.92       459

Confusion Matrix:
 [[393   8]
 [ 25  33]]
```

Chapter 5: Results and Discussion 5.1 Output / Report

The analysis and modeling yielded several key insights:

- **Model Performance:** The Logistic Regression model achieved an impressive **accuracy of approximately 93%** on the test set. The model demonstrated high precision (94%) for identifying low-risk patients and strong recall (98%) for the same class. For the high-risk class, the precision was 80%, indicating that when the model predicts high risk, it is correct 80% of the time.
- **Significant Risk Factors:** The hypothesis tests confirmed that **BMI** and **Systolic BP** showed a statistically significant difference ($p < 0.0001$) between high-risk and low-risk groups. This suggests these are strong indicators of CVD risk. Age and Diastolic BP did not show a statistically significant difference in this specific test.
- **Visual Insights:** The Power BI dashboard effectively visualizes these findings. For instance, the "Blood Pressure & Cholesterol Insights" page shows that patients with Hypertension Stage 2 have a notably higher concentration of high-risk cases. Similarly, the "Patient Lifestyle Insights" page illustrates the correlation between physical activity levels and CVD risk scores.

5.2 Challenges Faced

- **Data Quality:** The dataset contained a significant number of missing values across multiple columns (e.g., Age, Weight, BMI). This required a careful imputation strategy (using median and mode) to avoid introducing bias while retaining as much data as possible.
- **Class Imbalance:** The "High_Risk" category was a minority class (169 high-risk vs. 1360 non-high-risk patients). This can lead to models that are biased towards predicting the majority class. While our model performed well, techniques like SMOTE or class weighting could be explored to further enhance performance for the minority class.

5.3 Learnings

- Gained hands-on experience in the complete data science pipeline, from data cleaning to model deployment in a dashboard.
- Developed a deeper understanding of applying statistical tests (Ztest, T-test) to validate analytical findings.
- Learned to interpret machine learning evaluation metrics (accuracy, precision, recall) in the context of a real-world problem.
- Strengthened skills in data storytelling and creating intuitive, interactive dashboards using Power BI to communicate results to both technical and non-technical stakeholders.

Chapter 6: Conclusion 6.1 Summary

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

Chapter 7: References

https://pandas.pydata.org/

https://scikit-learn.org/stable/

https://matplotlib.org/

https://www.lpu.in/skilldevelopment/