**Heart Disease Analysis & Prediction**

**Abstract**

This project aims to analyze heart disease data and predict the risk of heart disease using data analysis and machine learning techniques. The goal is to identify key factors related to heart disease and create a model that can predict the likelihood of a person having heart disease.

**1. Introduction**

**1.1 Background**

Heart disease is a major health problem affecting millions worldwide. Early prediction can help in taking preventive measures and reducing risks. In this project, we use data-driven techniques to analyze heart disease factors and develop a model to predict the risk.

**1.2 Objectives**

* Develop a scalable data pipeline for heart disease data processing.
* Analyze the key factors influencing heart disease.
* Implement advance data engineering techniques to clean and process heart disease data.
* Use machine learning models to predict heart disease risk.
* Visualize important insights through Power BI.

**2. Literature Overview**

There has been extensive research on heart disease prediction. Most studies focus on factors such as age, gender, blood pressure, cholesterol, and lifestyle habits. Various machine learning algorithms (like Logistic Regression and Decision Trees) have been applied to predict the risk of heart disease. Modern data analysis and visualization tools have further enhanced the accuracy and interpretability of these predictions.

**3. Methodology**

**3.1 Design & Architecture**

This project follows a structured approach:

* **Data Collection:** Heart-disease related datasets sourced from public and private datasets. [Dataset]
* **Data Engineering:** Data preprocessing, handling missing values, feature selection, and transformation.
* **Machine Learning Models:** Logistic Regression, Random Forest, and Gradient Boosting; fine-tuning the best-performing model.
* **Real-Time Prediction:** Deploying trained ML models using Apache Kafka to enable real-time heart-disease.
* **Visualization:** Developing Power BI dashboards for better financial insights.
* **Integration:** Using Databricks for data pipeline scalability.
* **Code Implementation:** ETL, streaming data processing and inferencing models a real time environment, Data Bricks.

**3.2 Workflow**

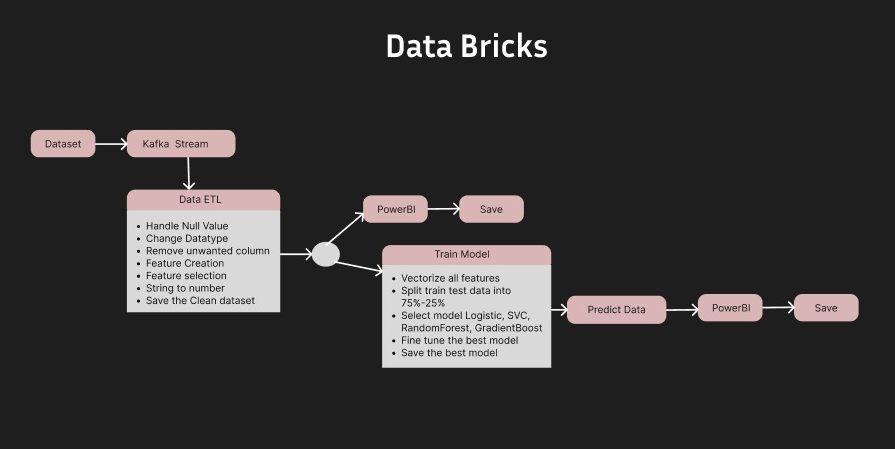


Figure 1: Workflow of the capstone project

**4. Data Insights and Visualization**

I created various visualizations using Power BI to gain insights into heart risk analysis: impact of smoking and gender:

* Age-wise heart disease distribution
* Gender-wise analysis of heart risk
* Cholesterol level comparison with heart disease
* Smoking status and heart disease correlation

**Dashboard overview**

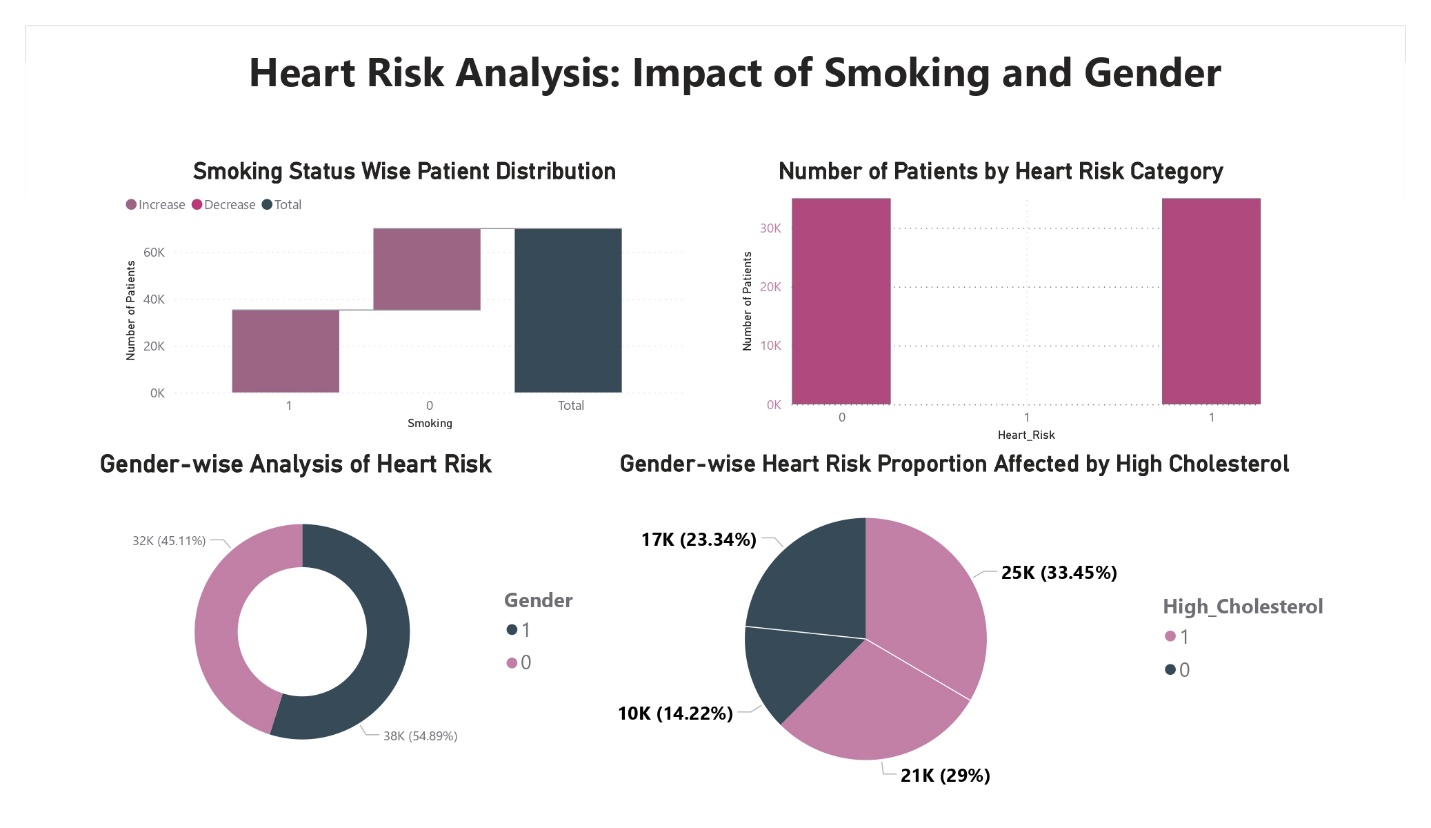


Figure 2: Power BI visualization of heart risks impact of smoking

**Correlation Heatmap**

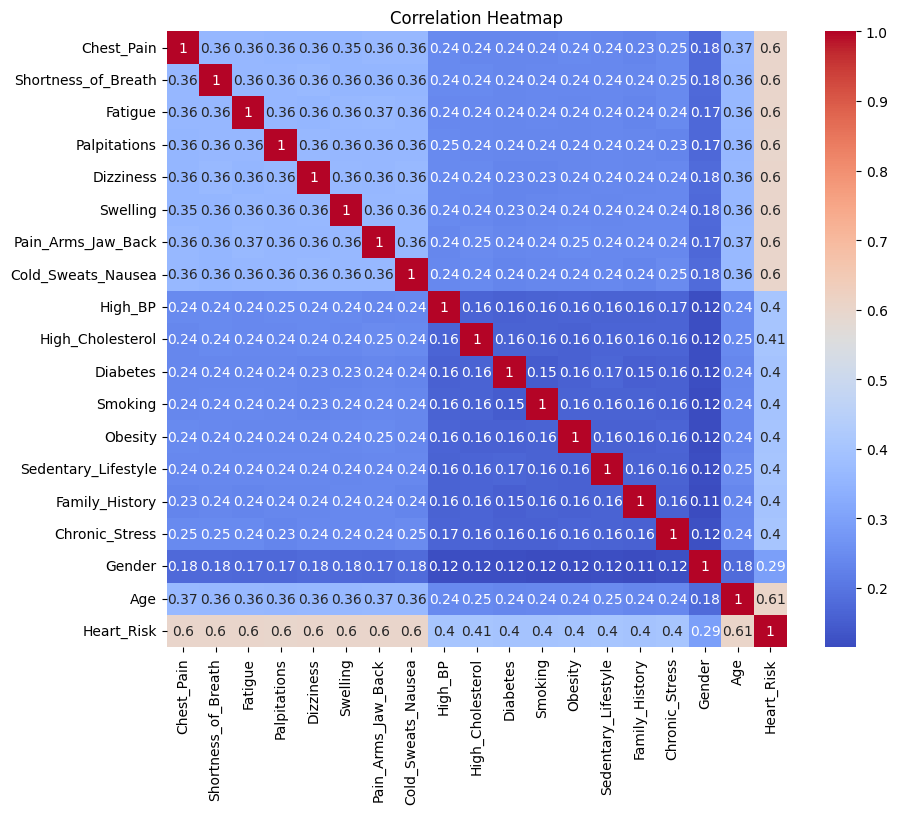
****

Figure 3: Correlation Heatmap

**Confusion Matrix:**

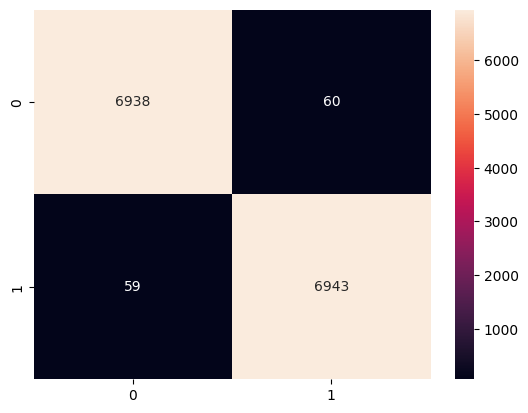
****

Figure 4: Confusion Matrix

**5. Conclusion**

**5.1 Summary**

This project successfully integrates big data engineering, machine learning, and visualization tools to optimize loan assessment. Through Databricks, Kafka, and Power BI, we ensure real-time predictions and scalable processing for health institutions. Also, Early prediction of heart disease can help save lives. By analyzing health data and creating predictive models, we can better understand the risk factors and take preventive measures.

**5.2 Future Work**

* Use more diverse datasets for better accuracy.
* Implement real-time prediction using streaming data.
* Improve the model by adding more health parameters.

**Project Repository and Documentation**

All in here

For further details, please refer to the complete documentations in the linked repository