Capstone Project CMPT 629

An intelligent system for predicting health risks and recommending preventive interventions

Hershil Piplani

Student Id: 301591594

Supervisors: Dr. Steven Bergner and Dr. Zhengjie Miao

Introduction

With chronic diseases on the rise, our healthcare systems are struggling to keep up, often reacting to health crises rather than preventing them. Current diagnostic methods are too slow, leading to delayed interventions and worsening patient outcomes.

What if we could predict health risks such as heart disease, chronic kidney disease, breast cancer, and other diseases before they occur, simply by analyzing patient data? This project aims to use Machine Learning to identify potential health risks early, allowing healthcare providers to act before it's too late. By enabling customized prevention strategies, personalized care, and more efficient resource allocation, this predictive system has the potential to transform our approach to healthcare—shifting from reaction to prevention and ultimately enhancing lives.

Research Background

Comparative Analysis of Logistic Regression, Gradient Boosted Trees, SVM, and Random Forest Algorithms for Prediction of Acute Kidney Injury Requiring Dialysis After Cardiac Surgery

In this research, the authors explore the use of Support Vector Machines (SVM), Random Forest, and Logistic Regression for classifying medical data, particularly in disease diagnosis. They conclude that while Logistic Regression is straightforward and easy to interpret, SVM and Random Forest demonstrate better performance due to their capacity to manage complex, high-dimensional data. The study highlights the benefits and challenges associated with each model and suggests that ensemble methods, such as Random Forest, may offer the best balance between accuracy and interpretability in medical applications.

Data Source

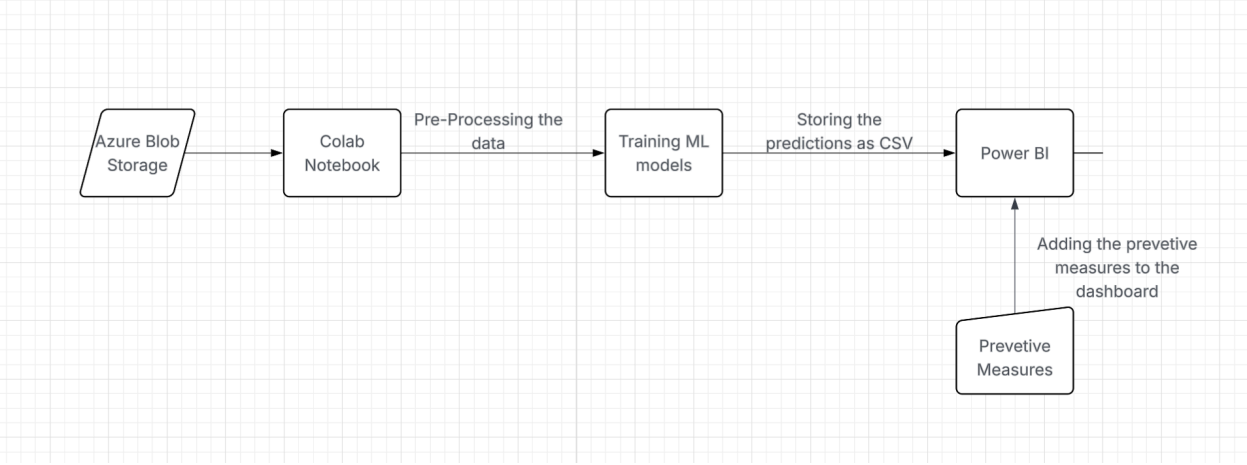
For this project, I utilized the UCI Machine Learning Repository. Finding the right dataset was the main challenge I faced. I also explored potential datasets from other sources, including StatsCan and Kaggle, and searched for quality data on the World Health Organization's Data Collection site, but I found that it was not relevant to my needs.

The UCI Machine Learning Repository offers a variety of well-structured datasets related to diseases, which is why I chose this resource. After reviewing the datasets available, I selected seven that pertain to seven chronic health issues: heart disease, chronic disease, breast cancer, liver disorder, diabetes, maternal health, and obesity.

Problem Statement

1) How can we utilize machine learning to accurately and promptly identify health risks from patient data, ensuring early intervention without compromising the quality of care?
2) Additionally, how can we apply data science to address the challenges posed by varying patient demographics, incomplete medical records, and biases in models? This approach should empower healthcare providers to deliver personalized, preventive care and enhance patient outcomes.

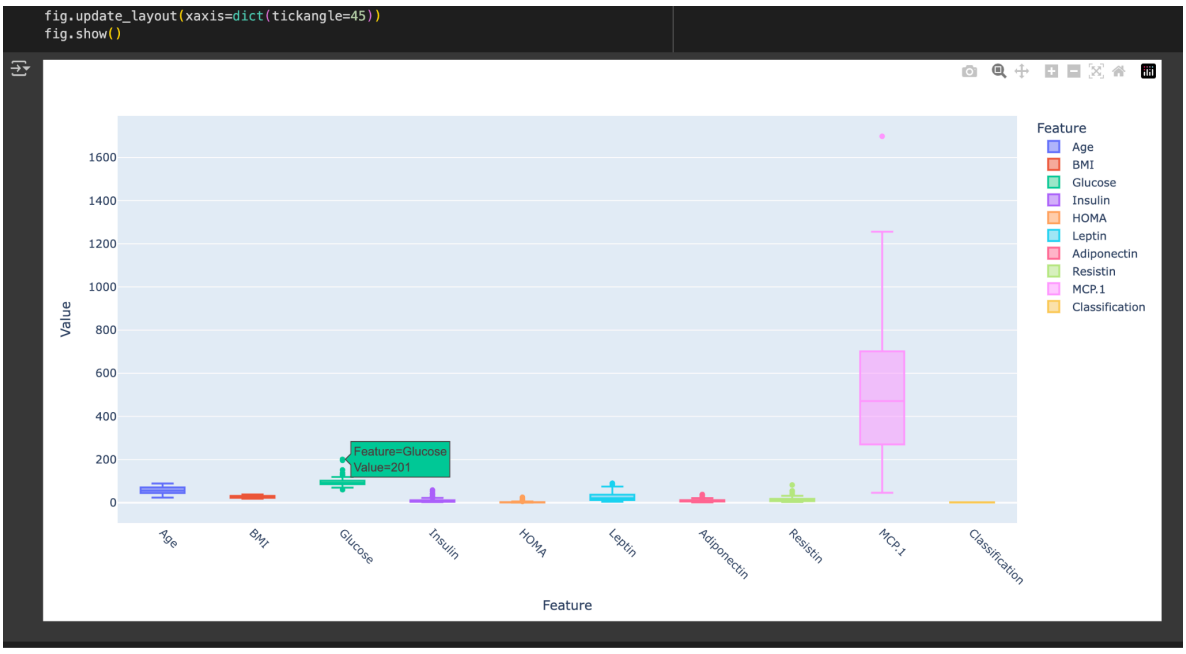
System Architecture



Methodology

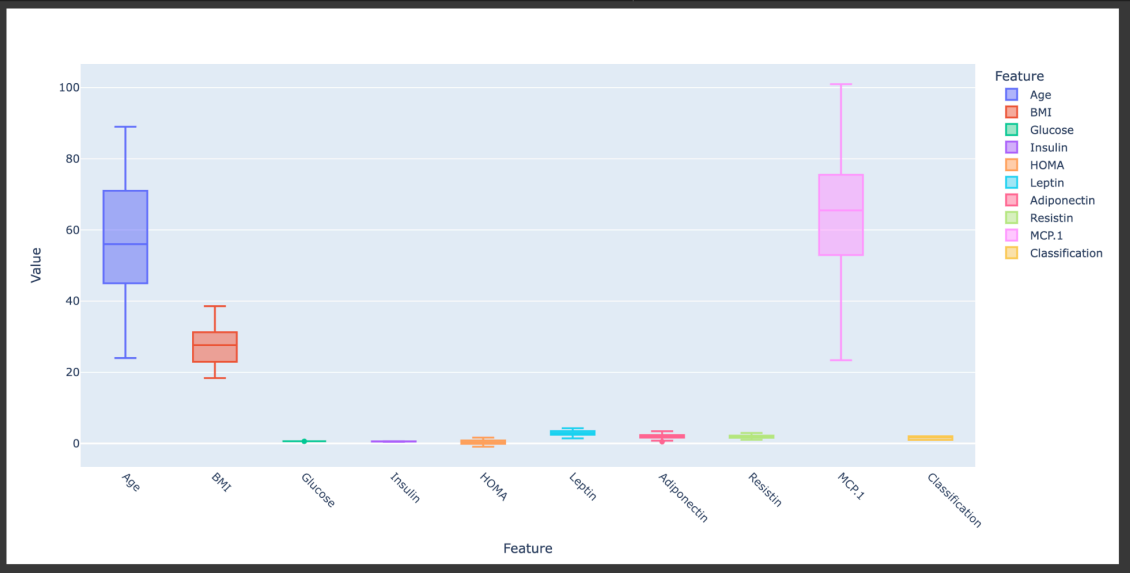
- **Data Storage**: The first step of the system involves storing the data. The data files are stored in Azure Blob Storage in CSV format..
- **Linking the Storage**: Initially, the Azure storage was connected to my working environment on Google Colab. However, to ensure data protection and security, the connection key and other variables were stored as GitHub secrets in the repository. Additionally, there is a GitHub Action file that facilitates the connection between the notebook and the storage.
- **Pre-processing the data**: The datasets used for this project are relatively small but of high quality. The pre-processing began with data cleaning and checking for NaN values, followed by an analysis of skewness and the number of outliers in each column. To address the outliers, I normalized the columns and applied transformations, including Log-Transformation, Yeo-Johnson, and Box-Cox. Below is a screenshot of the plot before and after normalizing the data and applying the transformations.

Before



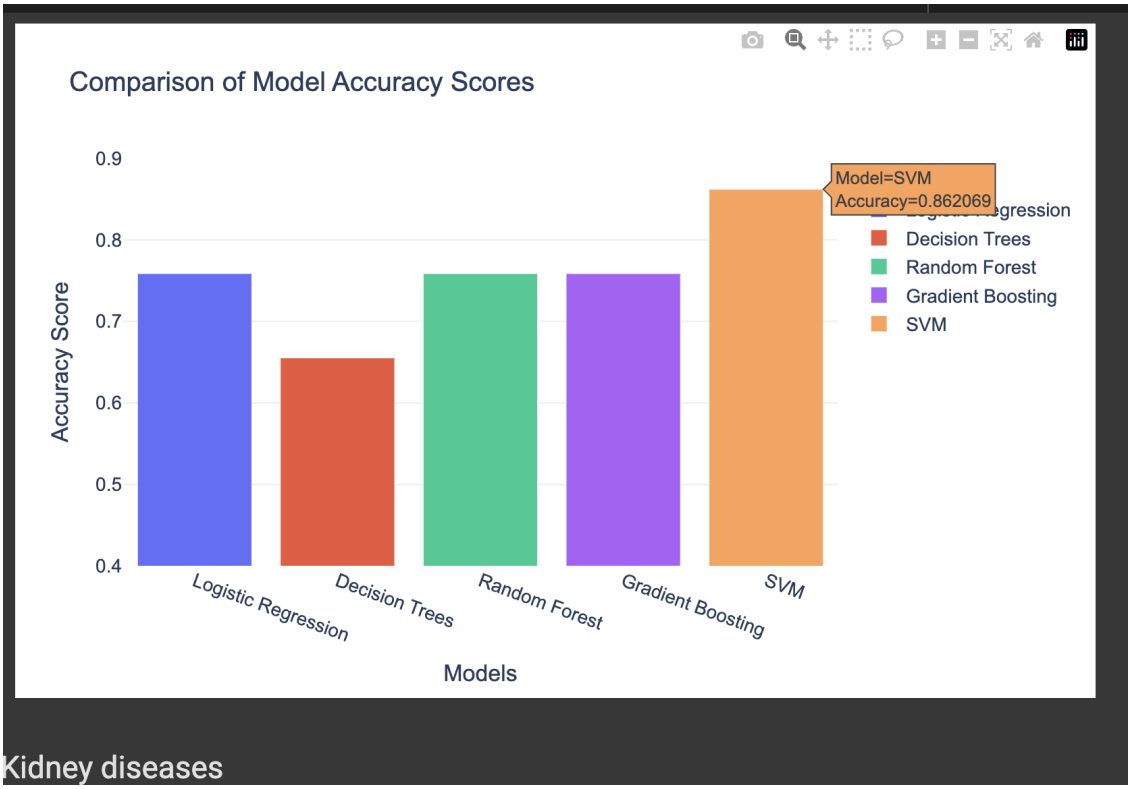
After

```
                height=600)

fig.update_layout(xaxis=dict(tickangle=45))
fig.show()
```



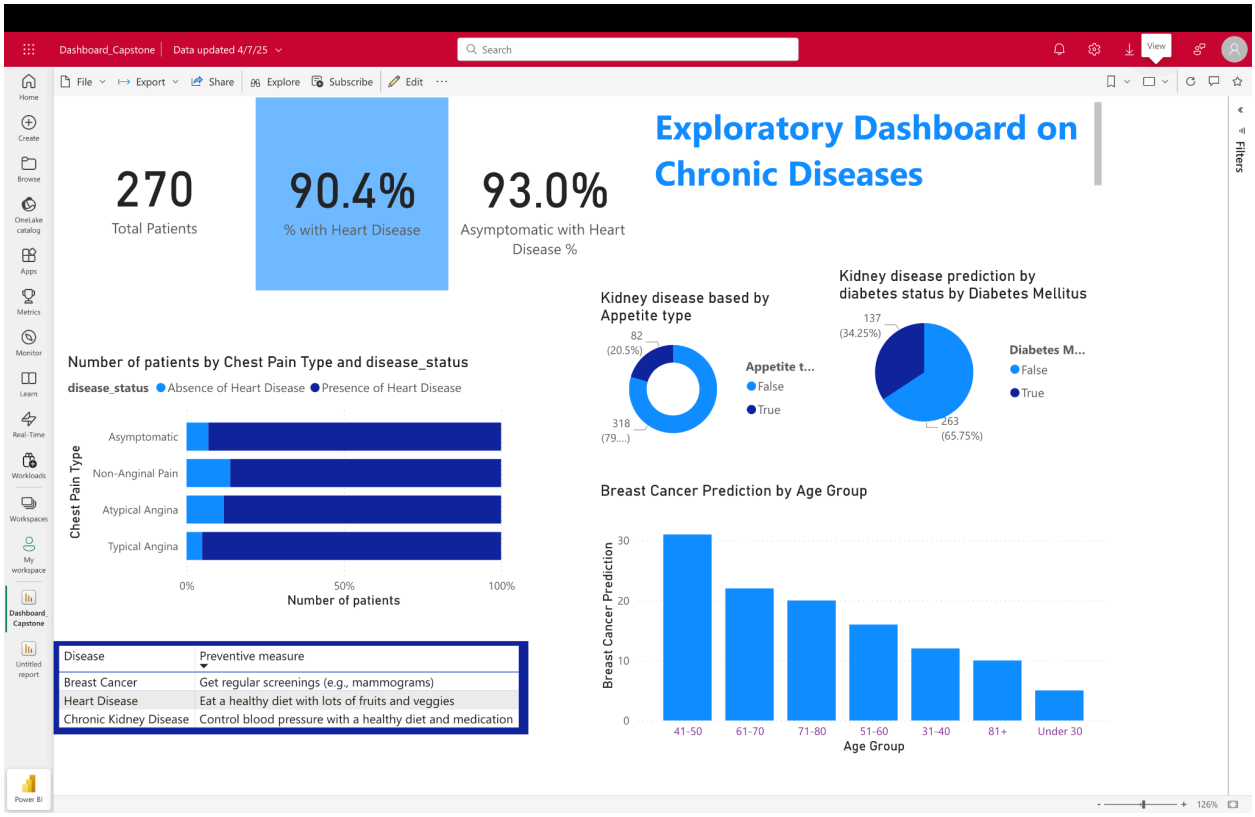

```
[ ] from sklearn.model_selection import train_test_split

    X = df_breastCancer.drop(columns=['Classification'])
```

- **Training ML models:** Since this is a classification problem, I decided to utilize algorithms such as Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). I trained my dataset using these algorithms and compared their accuracy scores. Below is a comparison of these methods applied to the Breast Cancer dataset.



- **Overfitting:** The dataset I used to model kidney diseases exhibited overfitting when I applied Decision Trees and Random Forest algorithms. To address this issue, I trained the model on both the training and testing sets and then calculated the accuracy scores for predictions on both sets. Initially, the model showed overfitting, as the training accuracy was 100% while the testing accuracy was around 86%. However, I improved the model by tuning parameters such as `min_samples_split` and `min_samples_leaf' which helped create a better fit.
- **Data Visualization:** I utilized predictions from the model that had the highest accuracy scores for each disease. For the dashboard, I chose to focus on diseases such as heart disease, chronic kidney disease, and breast cancer. The dashboard was created using Power BI.
- **Preventive measures**: I had to manually enter the top three preventive measures after researching online, as I could not find relevant APIs to scrape data and conduct topic modeling.

Overview of the Dashboard:



References:

1) **Datasources searched:**
    a) https://archive.ics.uci.edu/
    b) https://www150.statcan.gc.ca/n1/en/type/data?text=Health&subject_levels=13
    c) **https://www.kaggle.com/code/chanchal24/diabetes-dataset-eda-prediction-with-7-models**

2) **Research Background:** https://pmc.ncbi.nlm.nih.gov/articles/PMC11283789/