



Big Data Analytics for Diabetes Prediction

Deepika Venkatesan (dv2260), Dawood Ghauri
(dg4140), Ratik Vig (rv2292)

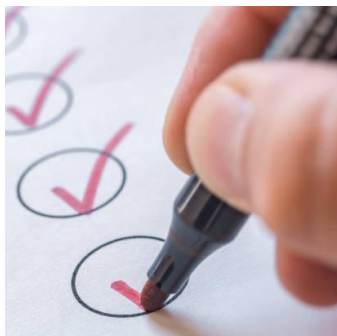
Big Data - Section D - S24
05/07/2024

Agenda



Problem Statement

Why is this important?



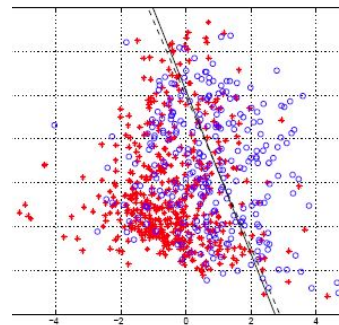
Objectives

What are our goals? How will we accomplish our goals?



Data Sources and Tools

What will help us accomplish our goals?



Results

What were our findings?

Problem Statement

Why is this important?

01

Problem Statement

- The World Health Organization lists diabetes among the top 10 causes of death globally; early detection through past data analysis is crucial for prevention.
- Utilize big data analytics to enhance early detection and management of diabetes through a comprehensive analysis of healthcare data from CDC.
- Develop a data ingestion pipeline for efficient Extract, Transform, and Load (ETL) of diverse healthcare data, focusing on patient demographics and lifestyle factors.

Problem Statement


- Produce visualizations to identify and understand critical factors influencing diabetes, supporting informed decision-making by healthcare professionals.
- Create advanced machine learning models to predict individual diabetes risk, using sophisticated algorithms to assess various risk factors and their interactions.

Objectives

What are our goals? How will we accomplish our goals?

02

Objectives

- 
1. Develop a Robust Data Ingestion Pipeline
 - a. Identify Relevant Diabetes Correlations
 - i. Diabetes, BMI, High Blood Pressure, Cholesterol Checked, Tobacco Use, Heavy Alcohol Consumption, etc.
 - b. Implement Systematic Data Cleaning and Preprocessing
 - c. Establish MongoDB Storage for Processed Data
 2. Develop API for Data Retrieval
 3. Visualize Critical Risk Factors through Web UI
 4. Implement Machine learning models using Keras and Tensorflow

Data Sources and Tools

What will help us accomplish our goals?

03

Data Sources and Tools

1. Technologies

- a. Data Processing: Spark Core
- b. Library: PySpark, pymongo
- c. Programming Language: Python3
- d. Backend Framework: Flask
- e. Frontend: HTML, JS
- f. Visualization: Matplotlib, Highcharts
- g. Processed Data Storage: MongoDB
- h. Machine Learning: Tensorflow, Keras

2. Datasets (2015, 2017, 2019, 2021):

https://www.cdc.gov/brfss/annual_data/annual_data.htm **(All data is loaded in s3 bucket, and keys are added to code. No additional setup required)**

3. Pipeline

Github: <https://github.com/ratik-vig/healthc>
[are_pipeline](https://github.com/ratik-vig/healthc)

4. Front-End

Github: <https://github.com/ratik-vig/healthc>
[are_vis](https://github.com/ratik-vig/healthc)

Results

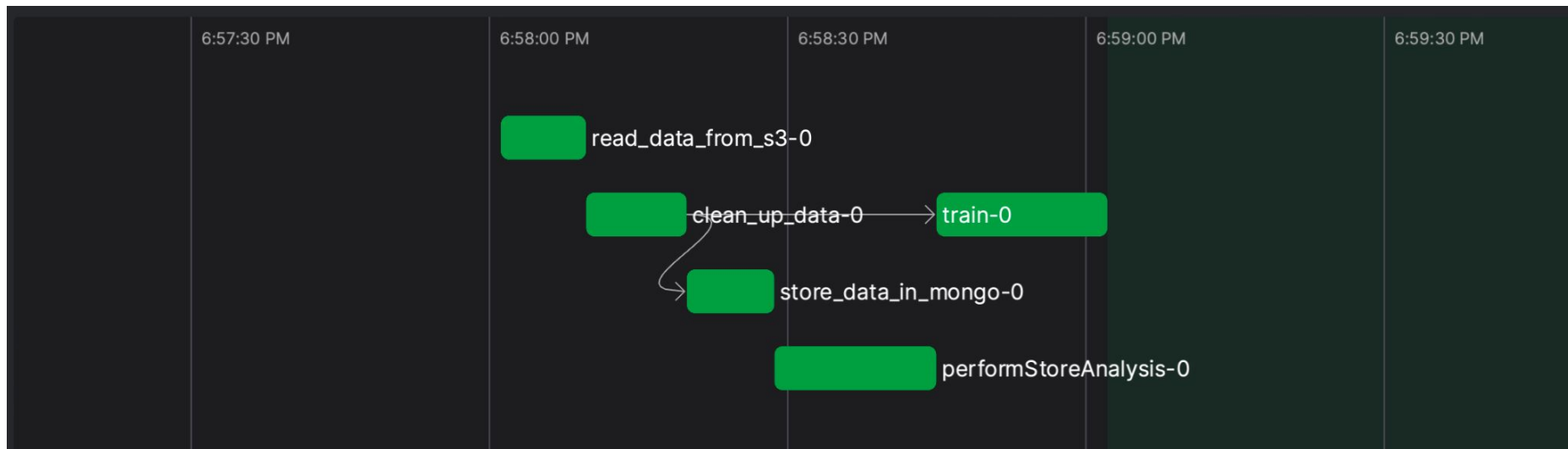
—
What were our findings?

04

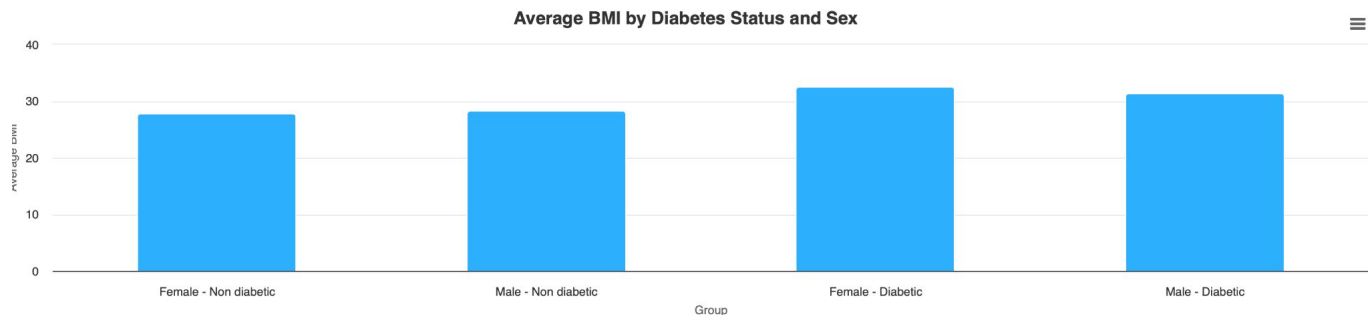
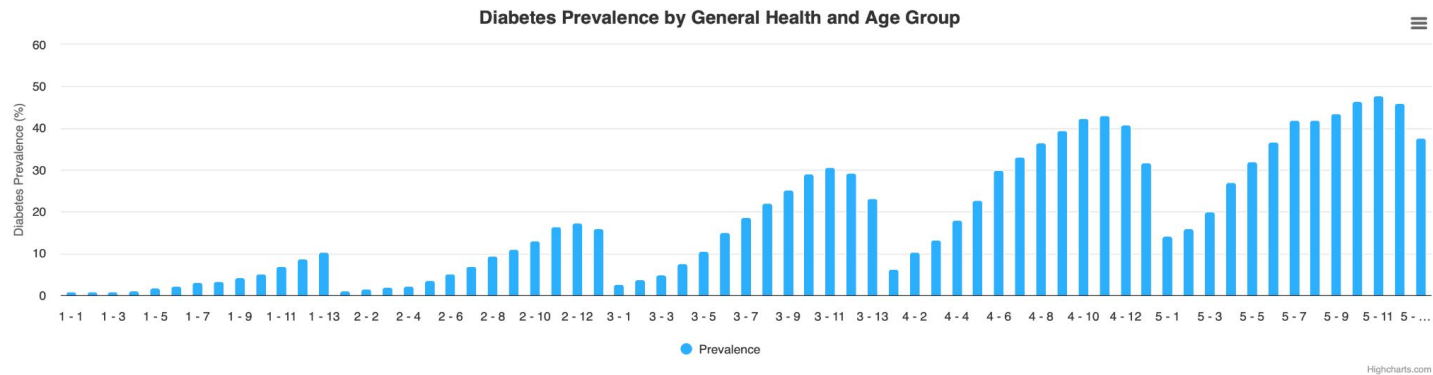
Results

- Created an batch data ingestion pipeline and stored clean and processed data to MongoDB
- Use Highcharts to analyze clean data through Flask endpoint
- Achieved a testing accuracy of 79%

Pipeline Execution

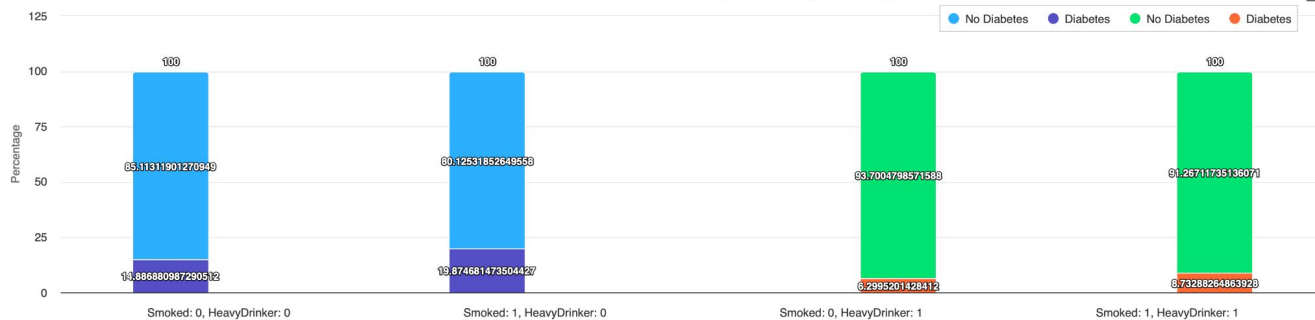


Visualizations

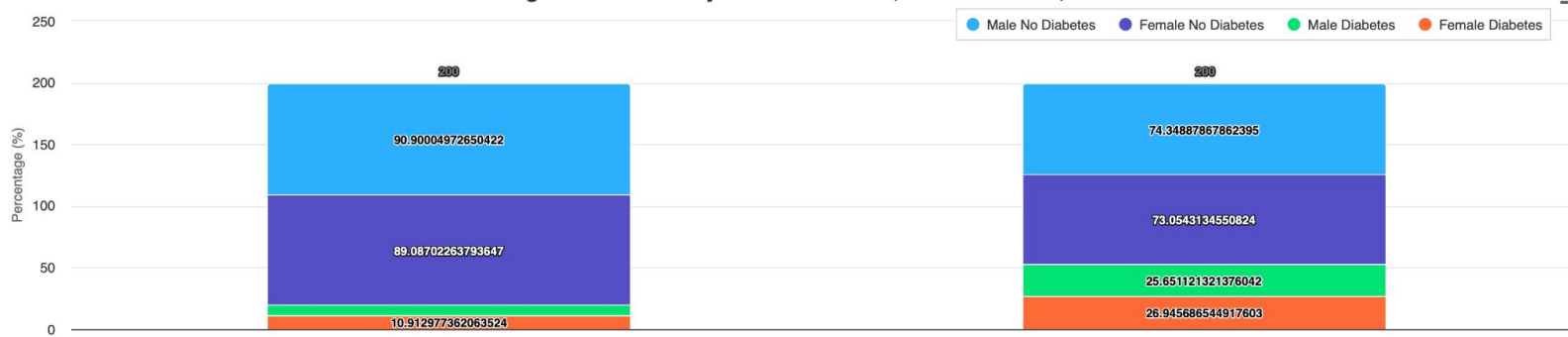


Visualizations

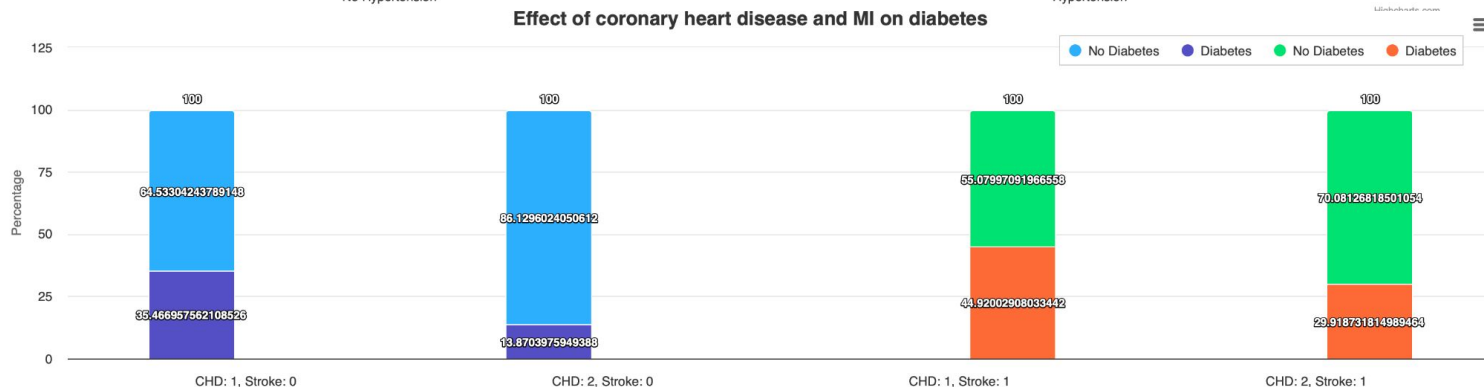
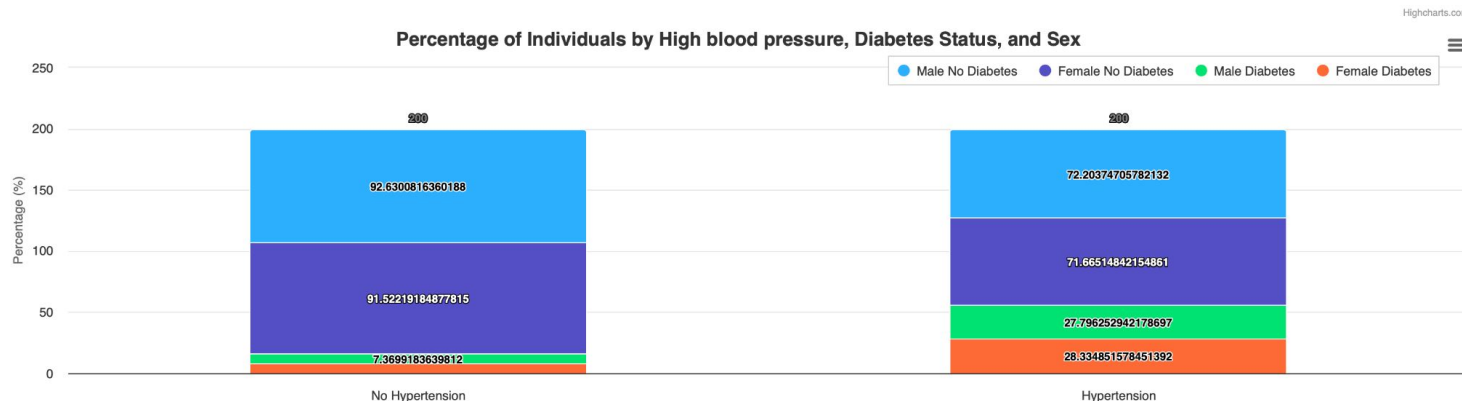
Diabetes Status Based on Smoking and Heavy Drinking



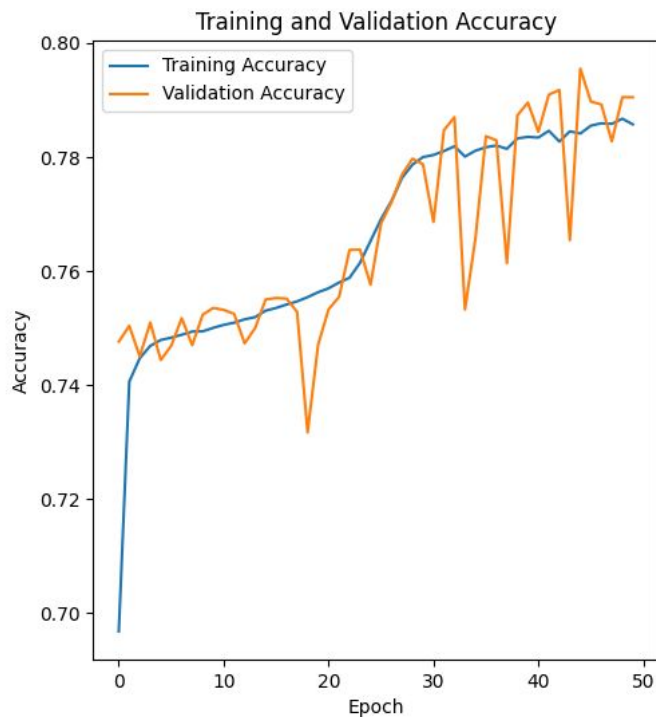
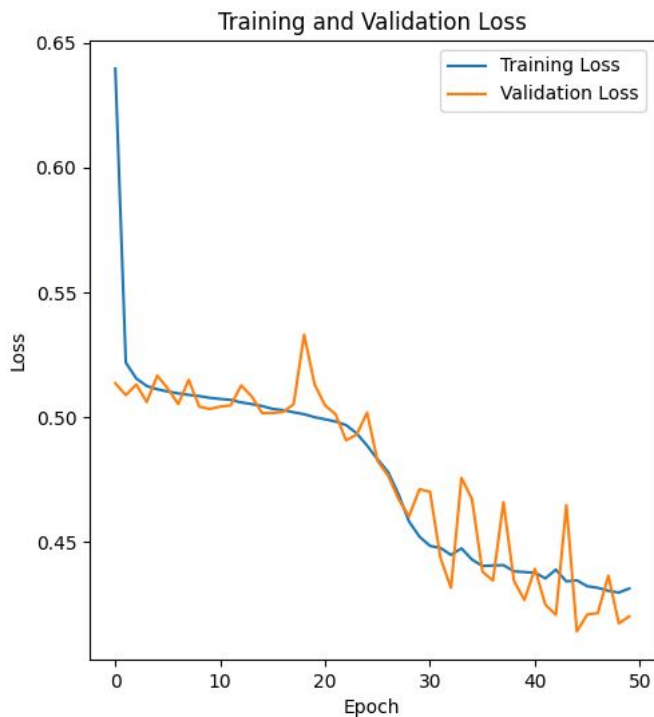
Percentage of Individuals by Cholesterol Level, Diabetes Status, and Sex



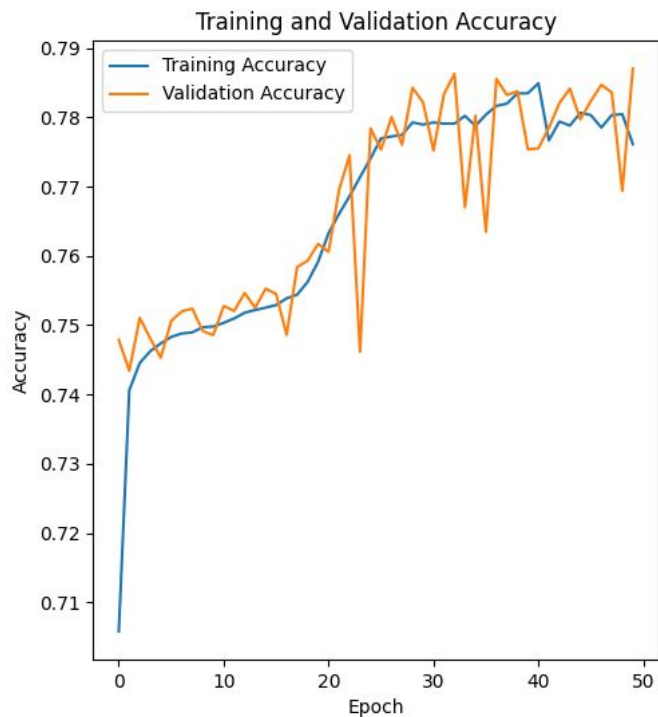
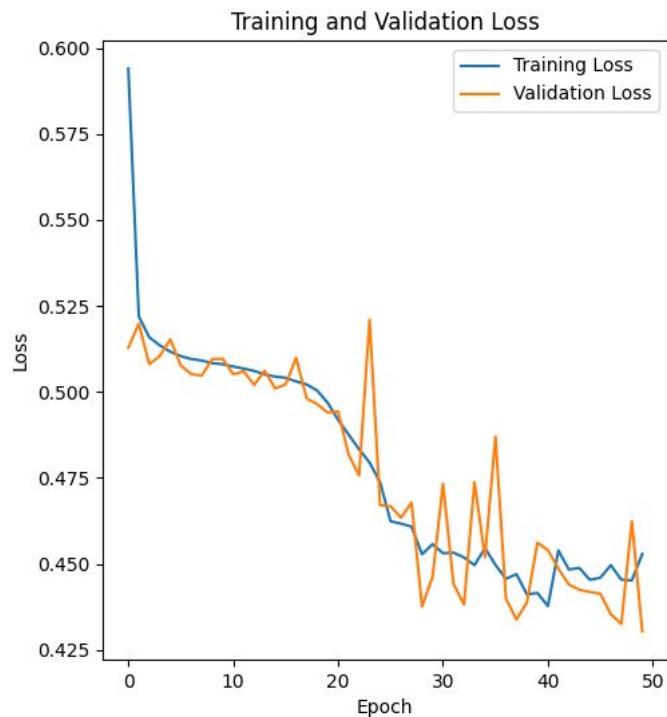
Visualizations



Model Training - Model 1



Model Training - Model 2



Conclusion and Lessons Learned

- Understanding of how to create Data ingestion and machine learning pipelines
- Experimented with various tools such as Airflow, Luigi and Prefect
- Trained ML models from scratch using Tensorflow and Keras
- Tried techniques to address class imbalance and finally implemented SMOTE

Questions?

- Deepika Venkatesan (dv2260@nyu.edu)
- Dawood Ghauri (dq4140@nyu.edu)
- Ratik Vig (rv2292@nyu.edu)