Data Exploration

Developer: Ethan Nolet

Sources:

- Kaggle: https://www.kaggle.com/search?q=heart+data
- UCI repo: https://archive.ics.uci.edu/datasets/?search=Heart+Disease
- Github: https://github.com/caravanuden/cardio?tab=readme-ov-file
- Publication: https://www.nature.com/articles/s41598-024-80919-9#Sec24

Resources Explored

- Cardiovascular_Disease_Dataset
- CDC Diabetes Health Indicators
- Heart Attack Risk & Prediction Dataset In India
- Heart Disease
- Risk Factors for Cardiovascular Heart Disease
- https://link.springer.com/article/10.1007/s10916-016-0536-z
- https://www.nature.com/articles/s41598-024-80919-9#Sec24
- Hospital Admissions Data: Two Year Hospital Admissions and Discharge Data from Hero DMC Heart Institute

Data to use for the project

Hospital Admissions Data: Two Year Hospital Admissions and Discharge Data from Hero DMC Heart Institute

Data Exploration Notes

1. https://link.springer.com/article/10.1007/s10916-016-0536-z and https://www.nature.com/articles/s41598-024-80919-9#Sec24 : From these, I established an initial list of risk factors for CVD, HTN, and CAD:

   CVD: Radiation exposure, age, ambient noise, irregular work hours, body weight and BMI, elevated triglyceride, history of HTN and CVD, elevated FBS levels, poor sitting posture, dust exposure, and smoking.

HTN: Continuous work shifts, BMI, body weight, family history of HTN, age, improper posture, elevated cholesterol levels, low HDL levels, gas exposure, prolonged sitting, smoking, dust exposure, and a history of HTN.

CAD: Smoking, Diabetes Mellitus, High Density Lipoprotein, Duke Treadmill Score, and Duration Recovery (from Duke Treadmill test)

2. Cardiovascular_Disease_Dataset: This focuses on clinical data obtained from patient exercise tests, which is the most useful for identifying CAD. I believe this would be best applied to the Plan A+ for identifying additional diseases.

3. CDC Diabetes Health Indicators: This database does contain a field HeartDiseaseorAttack, which tracks the presence of a coronary heart disease or myocardial infarction. Additional factors include high blood pressure, high cholesterol, BMI, smoker status, history of stroke, and physical activity, with all but the BMI stored as a binary value. I believe the physical activity field would be useful based on the relationship between sedentary lifestyle and cardiovascular diseases.

4. (Potential Best) Heart Attack Risk & Prediction Dataset In India: This database records the presence of HTN, but does not give heart disease status directly; it provides only an evaluation for 'Heart Attack Risk'. Additional factors include obesity status, smoking status, physical activity, cholesterol level, triglyceride level, LDL level, HDL level, blood pressure, exposure to air pollution, family history, and stress level. Across all of the datasets I found, this has the greatest overlap of factors with the publication, particularly due to including triglyceride levels, HDL levels, and exposure to air pollution.

5. (Potential Best if Verifiable) Heart Disease: This database records heart disease status directly. Additional factors include blood pressure, cholesterol level, exercise habits, smoker status, family heart disease history, diabetes, BMI, high blood pressure, low HDL, high LDL, stress level, sleep hours, sugar consumption, triglyceride level, CRP level, and homocysteine level. As with the [Heart Attack Risk & Prediction Dataset in India], this database has a great overlap of factors with the publication. However, I cannot find any information about the origin of this data, and the use of an AI-generated cover image increases my doubts about its integrity.

6. Risk Factors for Cardiovascular Heart Disease: This dataset records cardiovascular disease status directly. Additional factors include weight, blood pressure, cholesterol level, glucose level, smoker status, alcohol consumption, and active status. This data is notable for tracking 70,000 individuals, which I expect could be valuable.

7. Hospital Admissions Data: Two Year Hospital Admissions and Discharge Data from Hero DMC Heart Institute: This database contains information collected from patients admitted to the Hero DMC Heart Institute, such as their age, sex, patient history, smoking status, alcohol use, hemoglobin, total lymphocyte count, platelets, glucose, urea, creatinine, brain natriuretic peptide, raised cardiac enzymes, and ejection fraction.

Model Exploration

Developer: Mariama Njie

Data to use for the project

Hospital Admissions Data: Two Year Hospital Admissions and Discharge Data from Hero DMC Heart Institute

Language/Tools to use for the project

Python

Jupiper Notebooks

DB to use for the project

SQlite

Details about Dataset

This dataset contains information collected from patients admitted to the Hero DMC Heart Institute, such as their age, sex, patient history, smoking status, alcohol use, hemoglobin, total lymphocyte count, platelets, glucose, urea, creatinine, brain natriuretic peptide, raised cardiac enzymes, and ejection fraction.

Attributes Analysis

ALL ATTRIBUTES

SERIAL NUMBER, ADMISSION NUMBER, DATE OF ADMISSION, DATE OF DISCHARGE, AGE, GENDER, RURAL(R) /URBAN(U), TYPE OF ADMISSION-EMERGENCY/OPD, MONTH YEAR, DURATION OF STAY, DURATION OF INTENSIVE UNIT STAY, OUTCOME, SMOKING, ALCOHOL, DIABETES MELLITUS, HYPERTENSION, CORONARY ARTERY DISEASE, CARDIOMYOPATHY, CHRONIC KIDNEY DISEASE, HAEMOGLOBIN, TOTAL LEUKOCYTES COUNT, PLATELETS, GLUCOSE, UREA, CREATININE, B-TYPE NATRIURETIC PEPTIDE, RAISED CARDIAC ENZYMES, EJECTION FRACTION, SEVERE ANAEMIA, ANAEMIA, STABLE ANGINA, ACUTE CORONARY SYNDROME, ST ELEVATION MYOCARDIAL INFARCTION, ATYPICAL CHEST PAIN, HEART FAILURE, HEART FAILURE WITH REDUCED EJECTION FRACTION, HEART FAILURE WITH NORMAL EJECTION FRACTION, VALVULAR HEART DISEASE, COMPLETE HEART BLOCK, SICK SINUS SYNDROME, ACUTE KIDNEY INJURY, CEREBROVASCULAR ACCIDENT INFRACT, CEREBROVASCULAR ACCIDENT BLEED, ATRIAL FIBRILATION, VENTRICULAR TACHYCARDIA, PAROXYSMAL SUPRA VENTRICULAR TACHYCARDIA, CONGENITAL HEART DISEASE, URINARY TRACT INFECTION, NEURO CARDIOGENIC SYNCOPE, ORTHOSTATIC, INFECTIVE ENDOCARDITIS, DEEP VENOUS THROMBOSIS, CARDIOGENIC SHOCK, SHOCK, PULMONARY EMBOLISM, CHEST INFECTION

EXCLUDED ATTRIBUTES

SERIAL NUMBER, ADMISSION NUMBER, DATE OF ADMISSION, DATE OF DISCHARGE, RURAL(R) /URBAN(U), TYPE OF ADMISSION-EMERGENCY/OPD, MONTH YEAR, DURATION OF STAY, DURATION OF INTENSIVE UNIT STAY, OUTCOME

REMAINING ATTRIBUTES

AGE, GENDER, SMOKING, ALCOHOL, DIABETES MELLITUS, HYPERTENSION, CORONARY ARTERY DISEASE, CARDIOMYOPATHY, CHRONIC KIDNEY DISEASE, HAEMOGLOBIN, TOTAL LEUKOCYTES COUNT, PLATELETS, GLUCOSE, UREA, CREATININE, B-TYPE NATRIURETIC PEPTIDE, RAISED CARDIAC ENZYMES, EJECTION FRACTION, SEVERE ANAEMIA, ANAEMIA, STABLE ANGINA, ACUTE CORONARY SYNDROME, ST ELEVATION MYOCARDIAL INFARCTION, ATYPICAL CHEST PAIN, HEART FAILURE, HEART FAILURE WITH REDUCED EJECTION FRACTION, HEART FAILURE WITH NORMAL EJECTION FRACTION, VALVULAR HEART DISEASE, COMPLETE HEART BLOCK, SICK SINUS SYNDROME, ACUTE KIDNEY INJURY, CEREBROVASCULAR ACCIDENT INFRACT, CEREBROVASCULAR ACCIDENT BLEED, ATRIAL FIBRILATION, VENTRICULAR TACHYCARDIA, PAROXYSMAL SUPRA VENTRICULAR TACHYCARDIA, CONGENITAL HEART DISEASE, URINARY TRACT INFECTION, NEURO CARDIOGENIC SYNCOPE, ORTHOSTATIC, INFECTIVE ENDOCARDITIS, DEEP VENOUS THROMBOSIS, CARDIOGENIC SHOCK, SHOCK, PULMONARY EMBOLISM, CHEST INFECTION

Prediction Types

1. Presence of the disease
2. Type of diseasel. 5 types that can be predicted are:
    a. Heart Disease Types Present
    b. HYPERTENSION
    c. CORONARY ARTERY DISEASE
    d. CARDIOMYOPATHY
    e. HEART FAILURE
    f. CONGENITAL HEART DISEASE

Current Algorithm accuracy

[33]:
```
#Test models after Feature selection
test_models(cvd_class_data_30_df)
```

	Class1	Class2
Descision Tree Classification	72.27	64.4
RadomForest	32.26	35.55
Perceptron	72.03	50.48
GaussianNB	54.78	22.46
KNN	74.06	64.35
All Data	(6897, 27)	(6897, 27)
X-Train	(4827, 25)	(4827, 25)
Y-Train	(4827,)	(4827,)
X-Test	(2070, 25)	(2070, 25)
X-Test	(2070,)	(2070,)

UI Exploration

Developer: Ethan Nolet, Mariama Njie

Language to use for the project

Javascript

Python,

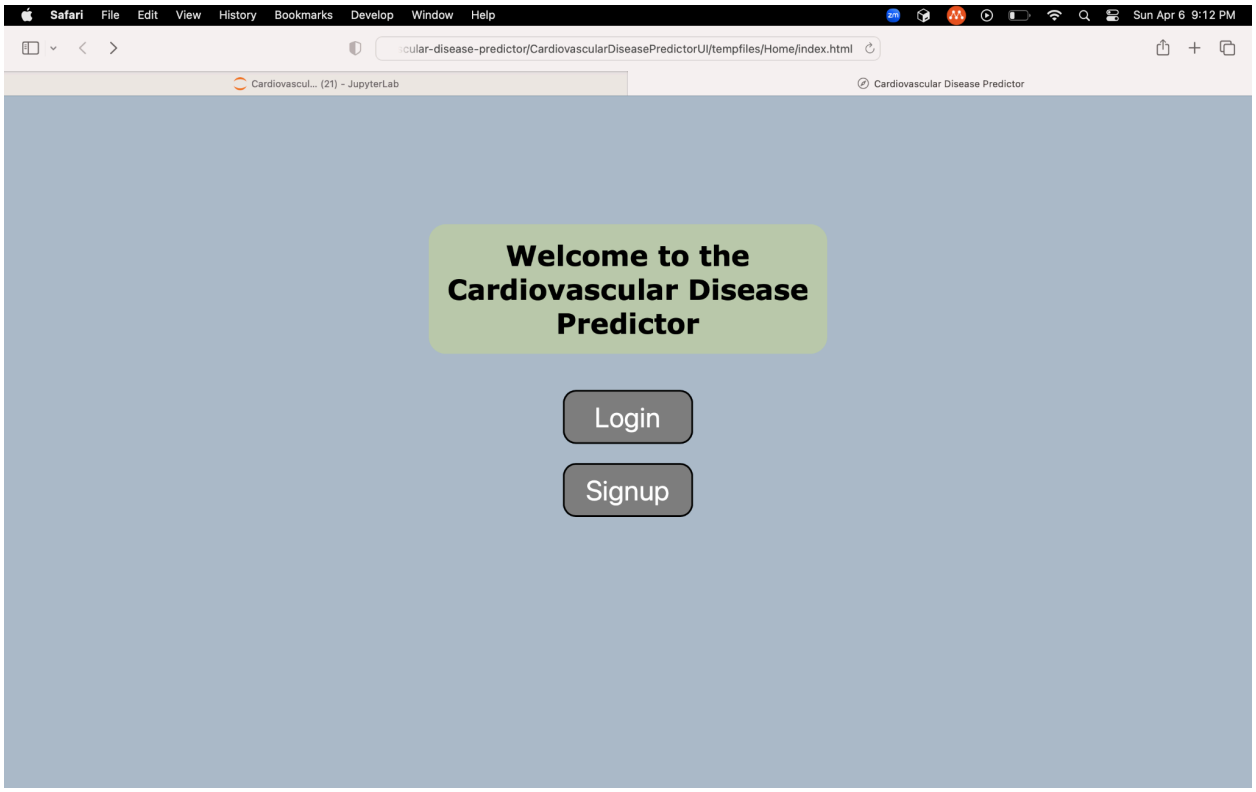
Flask,

Html,

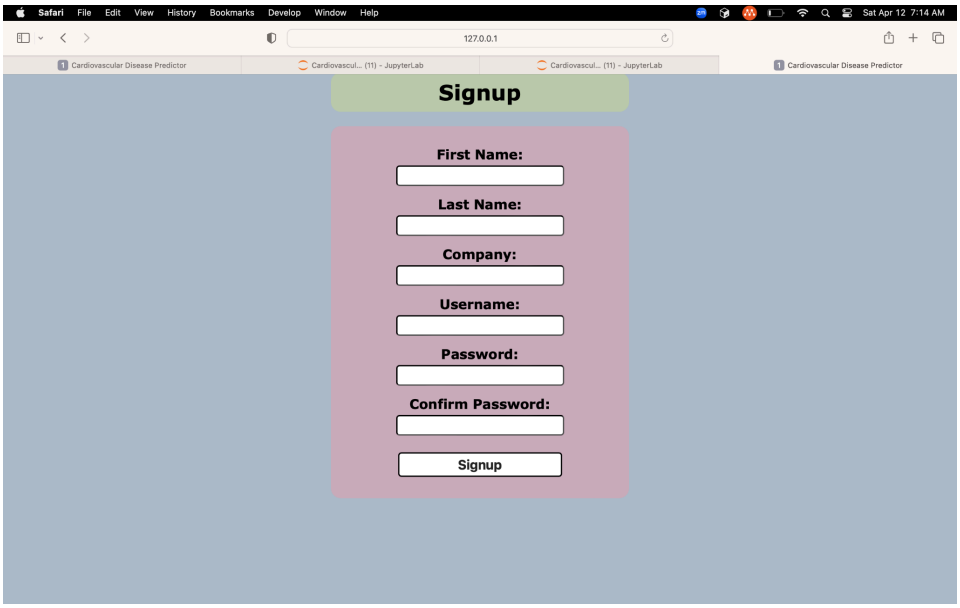
css

Implementation

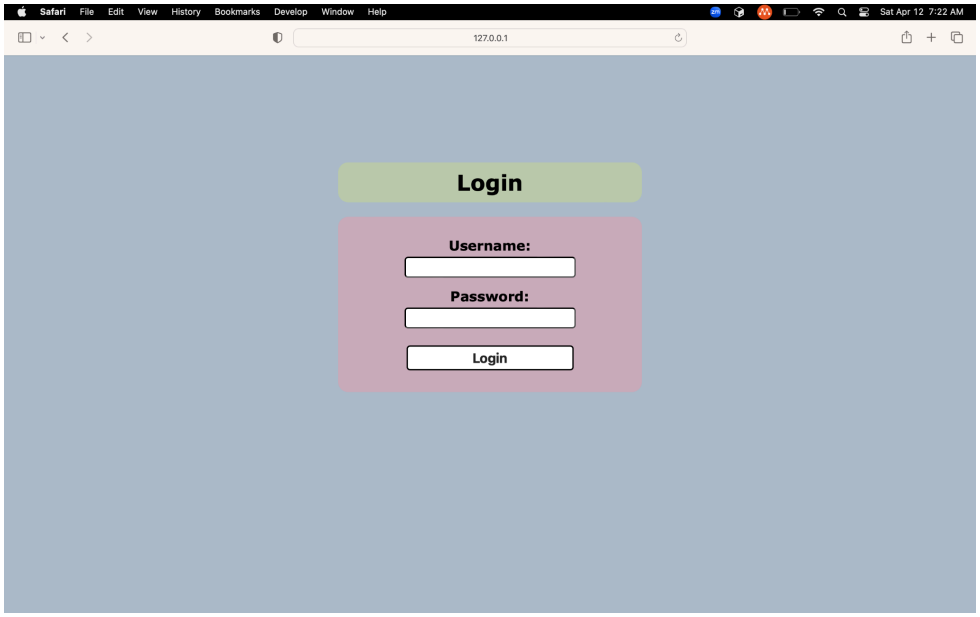
COMPONENT 1: HOME PAGE



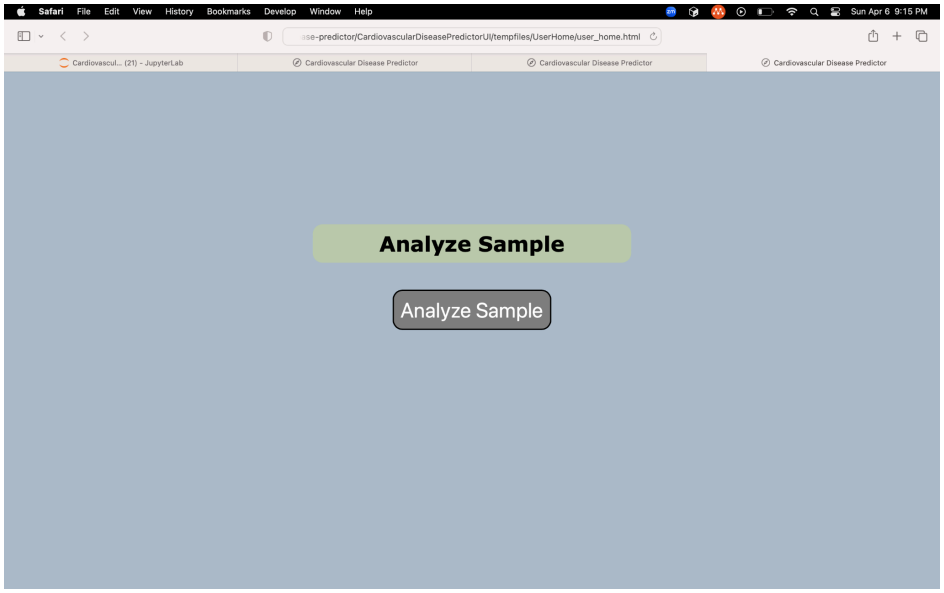
COMPONENT 2: USER SIGNUP



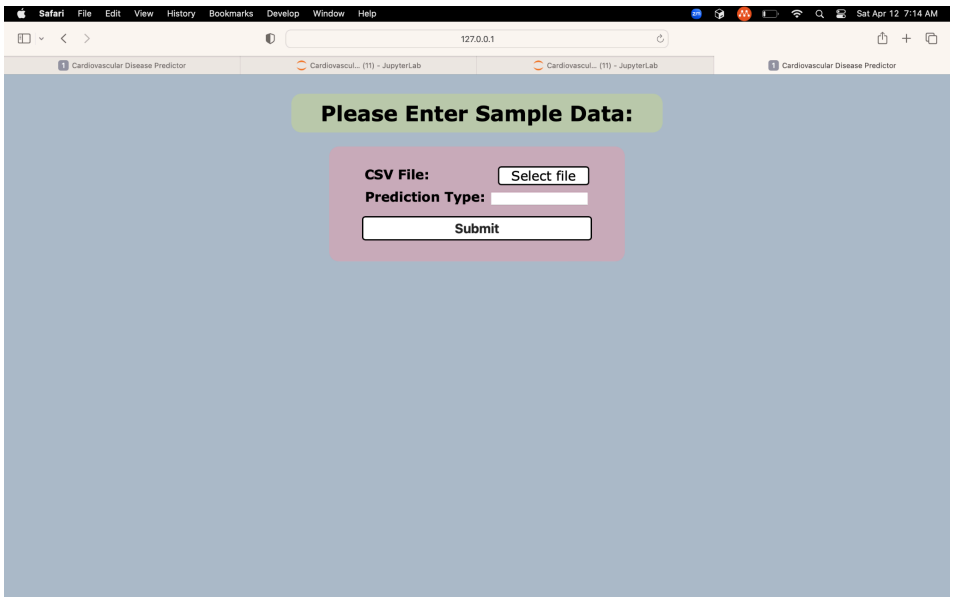
COMPONENT 3: USER LOGIN



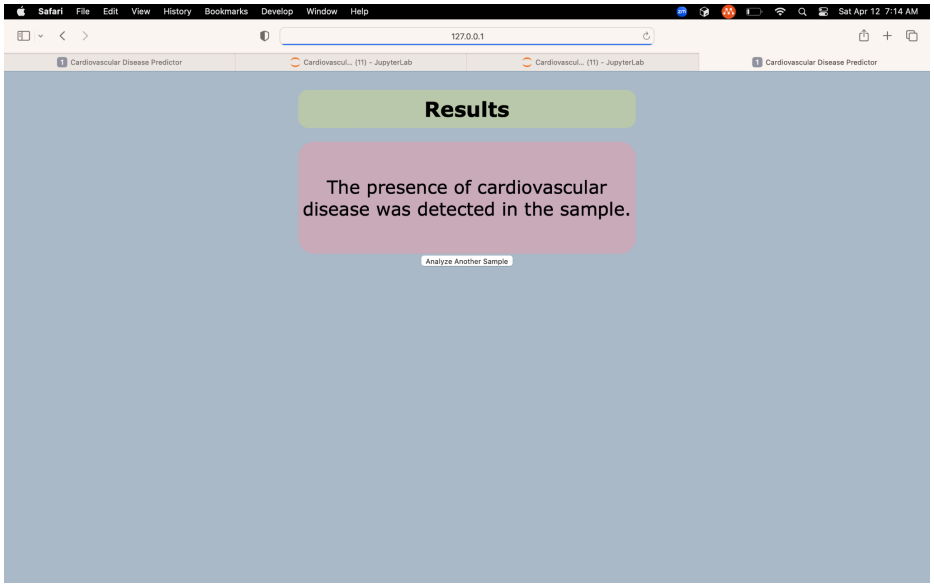
COMPONENT 4: SAMPLE ANALYSIS PAGE



COMPONENT 5: SAMPLE ENTRY PAGE



COMPONENT 5: RESULT PAGE



Backend exploration

Developer: Mariama Njie, Ethan Nolet

Language/Tools to use for the project

Python,

SQlite

Implementation