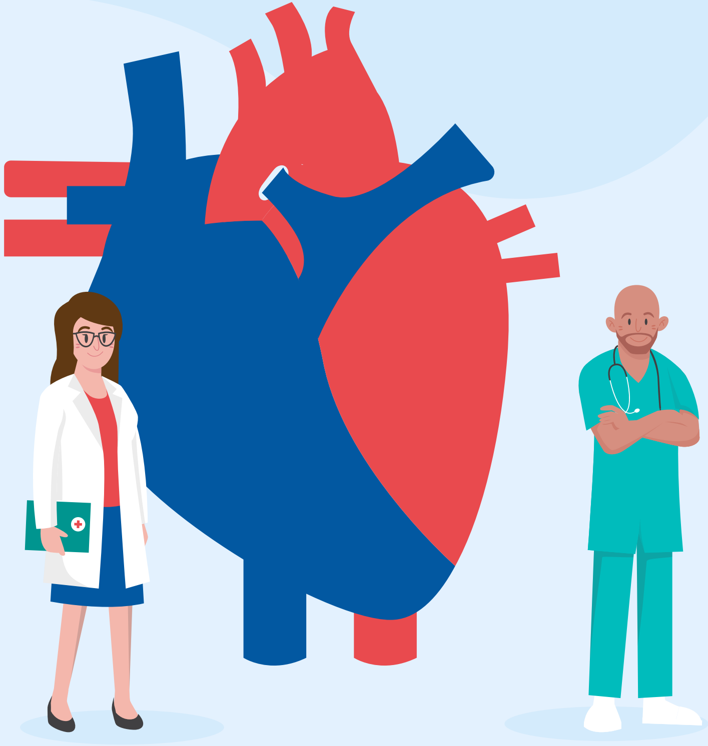
Cloud Project: Heart Attack Risk Prediction

Sydney Li, Zhuoyan Liu, Kira Luo, Jiayue Tian, Judy Zhu



Steps To Follow

01 Planning and Budgeting

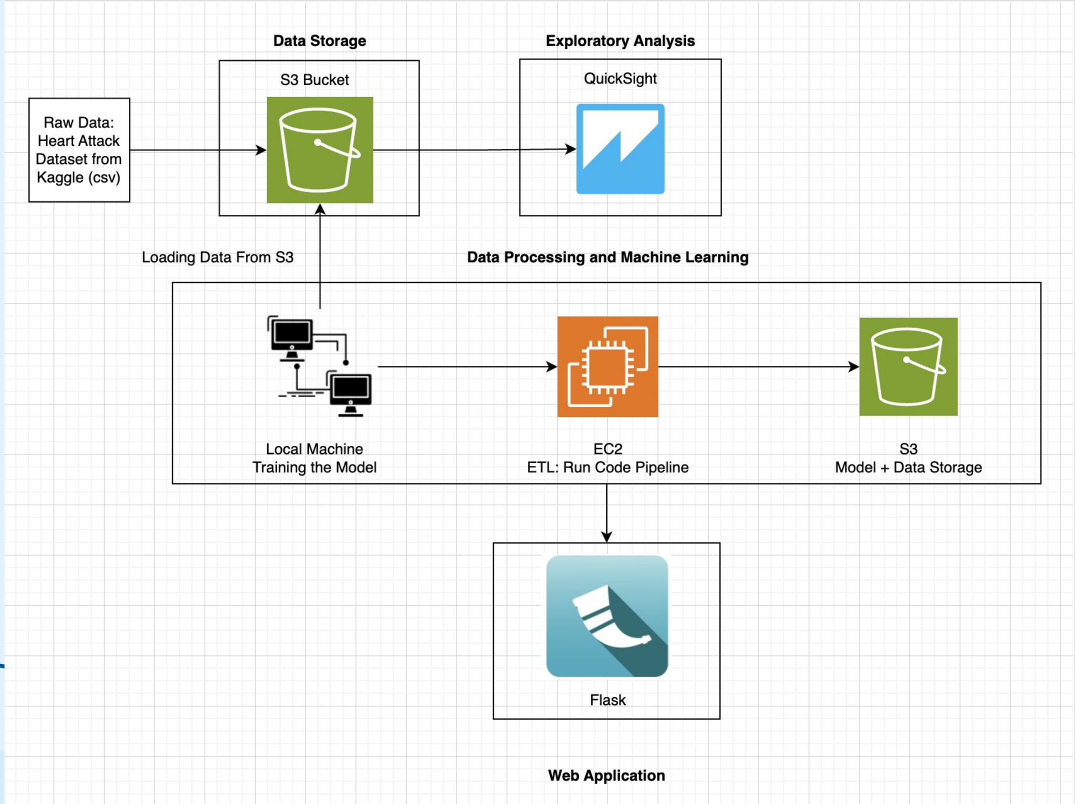
02 Data Collection and Preparation

03 Model Building and Training

04 Model Deployment

05 Configuration Files, Logging, and Monitoring

Planning: Architecture Diagram



Budgeting: Cost Estimate

$22.83 USD

is Monthly Cost

$273.96 USD

Is Total 12 months cost

Detailed Estimate

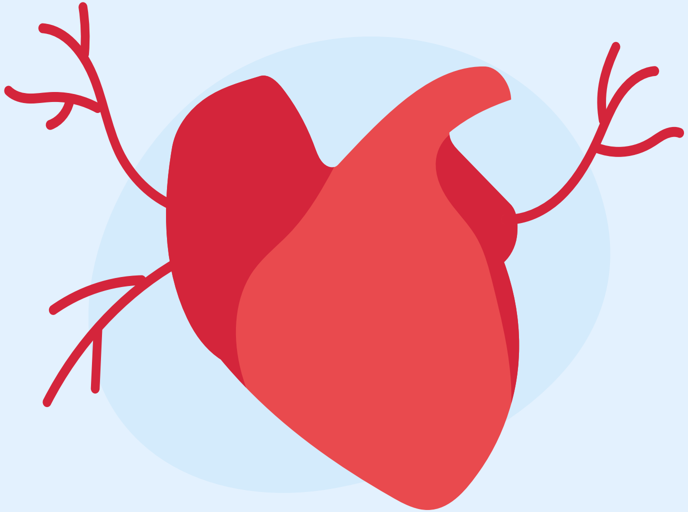
Name	Group	Region	Upfront cost	Monthly cost
Amazon Simple Storage Service (S3)	No group applied	US East (Ohio)	0.00 USD	0.00 USD
Status: - **Description**: **Config summary**: S3 Standard storage (0.01 GB per month), Data returned by S3 Select (0.005 GB per month)				
Amazon QuickSight	No group applied	US East (Ohio)	0.00 USD	18.60 USD
Status: - **Description**: **Config summary**: Number of working days per month (1), SPICE capacity in gigabytes (GB) (10), Number of authors (1), Number of readers (1)				
Amazon EC2	No group applied	US East (Ohio)	0.00 USD	4.23 USD
Status: - **Description**: **Config summary**: Tenancy (Shared Instances), Operating system (Linux), Workload (Consistent, Number of instances: 1), Advance EC2 instance (t2.micro), Pricing strategy (3yr No Upfront), Enable monitoring (disabled), DT Inbound: Not selected (0 TB per month), DT Outbound: Not selected (0 TB per month), DT Intra-Region: (0 TB per month)				

Data Collection

Dataset Source: Kaggle

Dataset Name: Heart Attack Prediction Dataset

Link:
https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset



Data Preparation

Overview: Crucial Factors Predicting Heart Attack Risk

01 Demographics and Geographic

Age, Sex, Income, Country, Continent, Hemisphere

02 Diagnosis

Cholesterol, Blood Pressure, Heart Rate, Diabetes, Family History, Previous Heart Problems, BMI, Triglycerides, Medication Use

03 Lifestyle Choices

Smoking, Obesity, Alcohol Consumption, Diet

04 Activity and Exercise

Exercise Hours Per Week, Physical Activity Days Per Week, Sedentary Hours Per Day

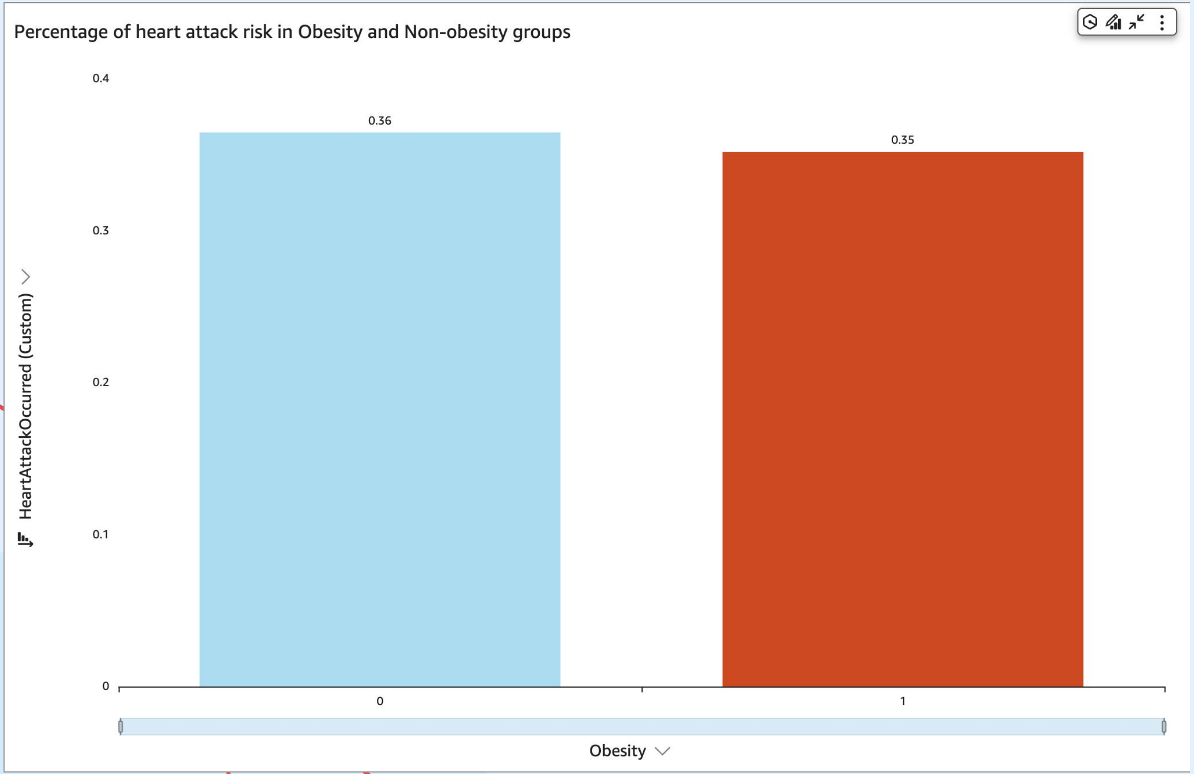
05 Well-being

Stress Level, Sleep Hours Per Day

06 Heart Attack Risk

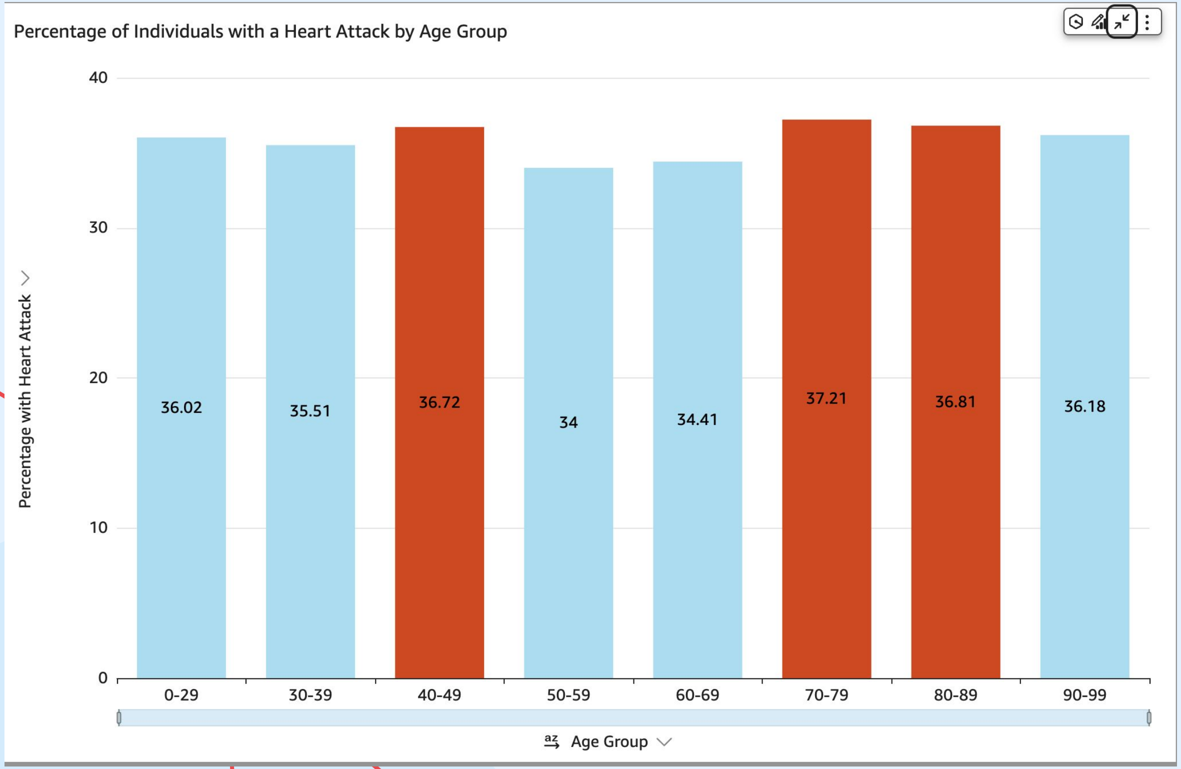
Heart Attack Risk (what we are predicting)

Key Health Metrics Analysis



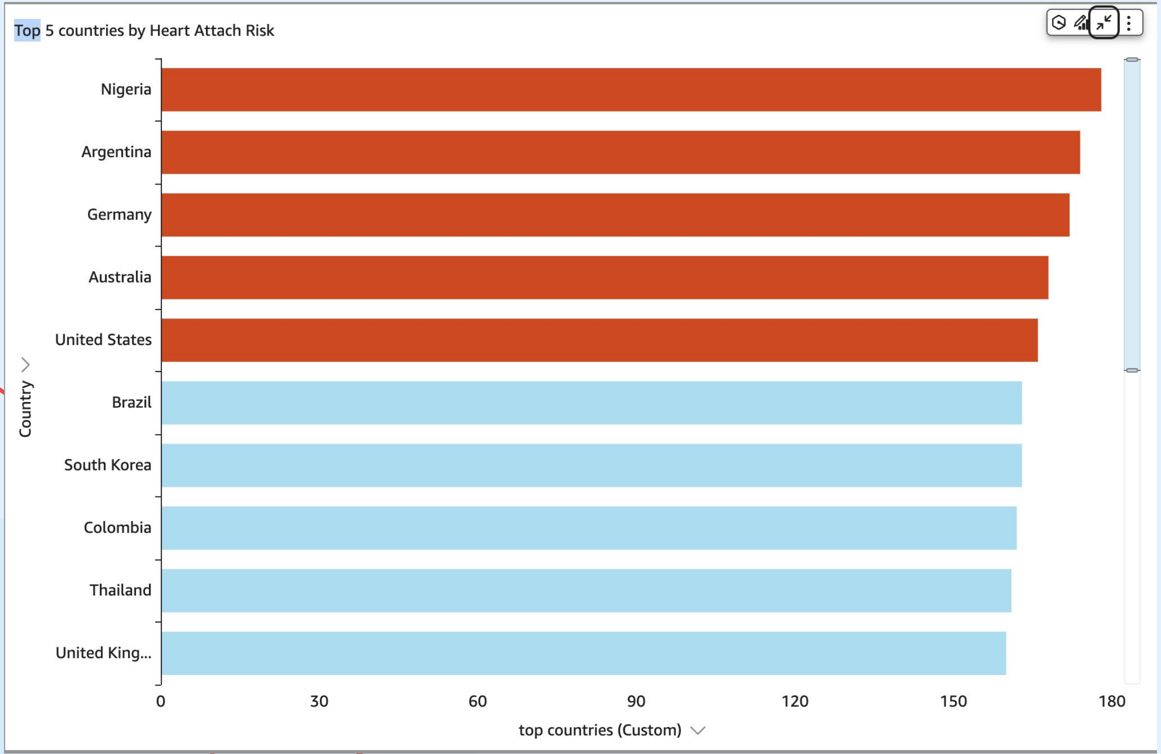
- This bar chart shows the heart attack risk percentage in two groups: individuals with obesity and those without.
- The heart attack risk is marginally higher in the non-obesity group (35%) compared to the obesity group (36%).
- This may indicate other risk factors influencing heart attack incidence beyond obesity.

Demographic and Health Insights



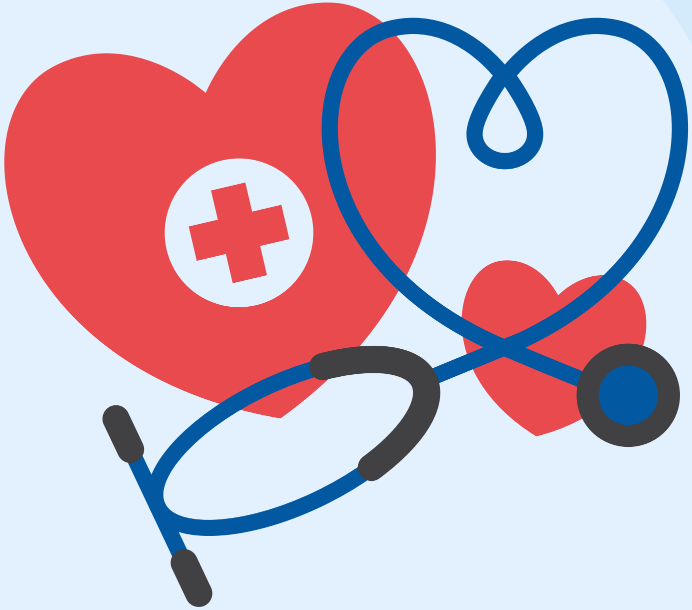
- The heart attack incidence generally increases with age, peaking in the 70-79 age group at 37.21% before slightly decreasing in older age groups.
- Interestingly, the rate is not much lower even in the younger demographic, with the 40-49 age group showing a notable risk at 36.72%.
- This data highlights the substantial risk of heart attacks in middle-aged individuals, alongside the expected higher risk in older adults.

Demographic and Health Insights



- This bar chart ranks countries by the risk of heart attack, with Nigeria having the highest risk and Argentina, Germany, Australia, and the United States following.
- This visualization emphasizes the variation in heart attack risk across different countries, potentially influenced by factors like healthcare access, lifestyle, and population demographics.

Model Building and Training



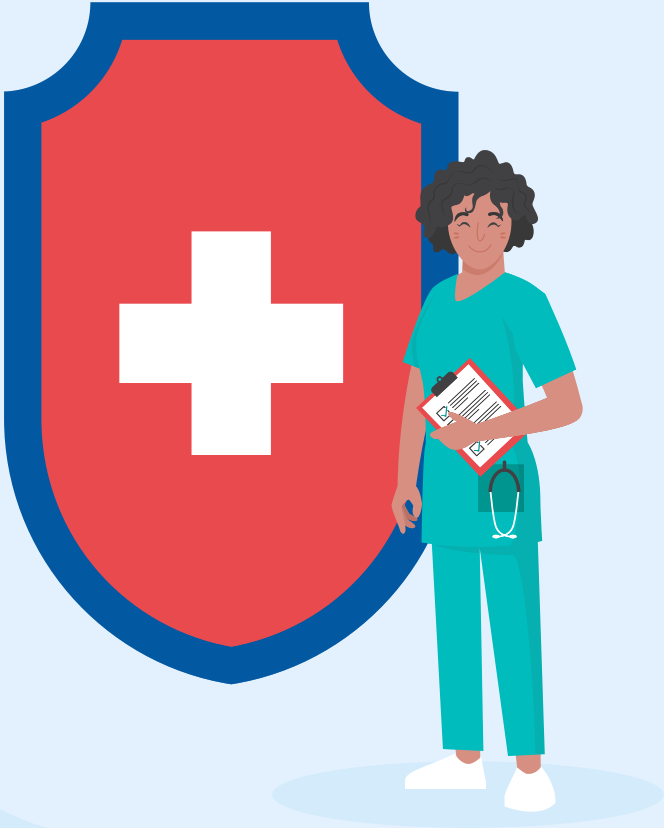
Model Building and Training

- **Data preprocessing:** data cleaning, feature engineering (feature selection, standardize numerical features and one hot encoding for categorical features)
  - Features: age, heart rate, cholesterol, diabetes, family history, smoking, obesity, alcohol consumption, exercise hours pw, diet, blood pressure
- **Deal with unbalanced data using SMOTE**
  - Generate new samples in the minority class (class 1 with higher risk)
- **Model training:**
  - Train/test set split with 0.8 ratio
  - Hyperparameter tuning on random forest model based on grid search cross-validation of 5 folds
  - Best hyperparameters: {'max_depth': 40, 'max_features': 5, 'min_samples_leaf': 4, 'n_estimators': 400}

Configurations And Further...

Configuration Files, Logging, and Monitoring

- **Configuration Files:** pull out all necessary configurations into default-config.yaml
- **Logging:** 3 distinct levels of logging used, using standard naming convention, etc…
- **Enable the reproducible execution of each step of the mode development:**
  - Get raw data from s3 —> modeling —-> save artifacts to S3 in another bucket
  - Split pipeline.py into 8 modular functions in .py files
  - Artifacts are properly saved at each step
- **Unit Testing:**
  - happy path and unhappy path (ensure only numeric values are supplied to StandardScaler)
- **Pylint Evaluation:** 10/10 for all .py files (pylint --rcfile=.pylintrc [files.py])
- **Type Hints, Docstrings, Requirement.txt, Exception Handling, and NO Hard-coding:** they are used appropriately throughout application



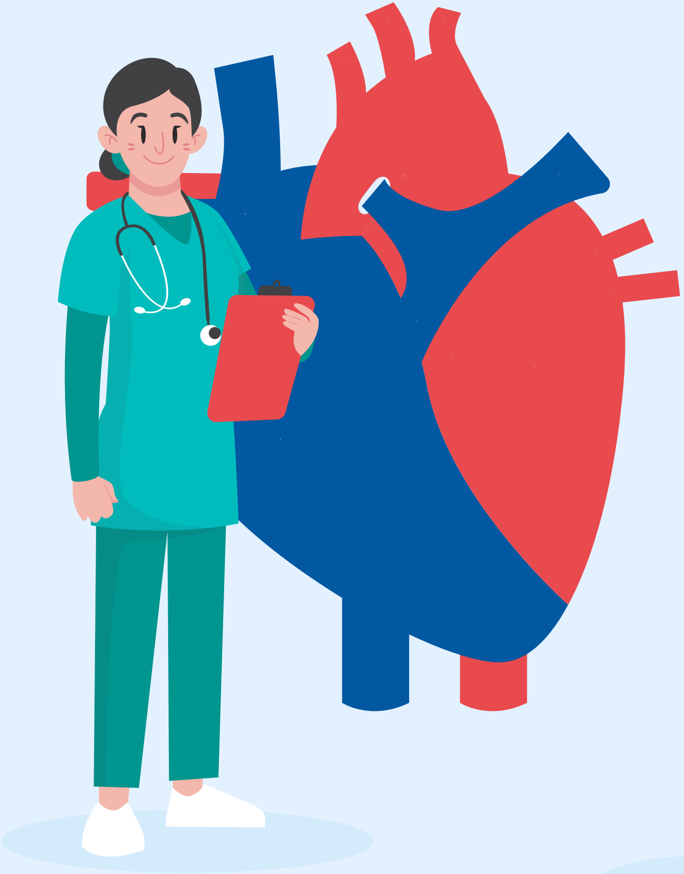
Model Deployment

Model Deployment

- **Approach:** using EC2 to host our model in a scalable and secure manner
- **Steps:**
  1. Launch my own AWS EC2 instance (t2.micro free tier)
  2. Securely connected to AWS EC2 instance in the terminal (project_key.perm)
  3. Clone the existing Github repository
  4. Securely connect the Github repository and the AWS EC2 instance via SSH key

     “sudo yum update -y/ install git -y/ install python3 -y”

     "git clone [our git repo]”
  5. Set Up Python virtual Environment and Manage dependencies (install requirements.txt)
  6. Run the pipeline.py via EC2
  7. All artifacts are saved properly (similar output as running locally)



Web Applications

Thank You