HEART ATTACK RISK PREDICTION

by

Group 4

Nathas Sungworawongpana , st124323

Sonu Adhikari, st124409

A Proposal Submitted for Computer Programming for DSAI

Asian Institute of Technology

School of Engineering and Technology

Thailand

October, 2023

Introduction

Cardiovascular diseases, including heart attacks, are a leading cause of mortality worldwide. Understanding the factors contributing to heart attack risk is essential for early intervention and prevention. This project aims to conduct a comprehensive analysis of a heart attack risk dataset to gain insights into the factors that influence heart attack risk. The primary goal of this project is to create a predictive model to identify individuals at a higher risk of heart attacks.

The main problem we aim to address is the prediction of heart attack risk. We want to develop a model that can predict an individual's likelihood of experiencing a heart attack based on a combination of medical and lifestyle factors, such as age, blood pressure, cholesterol levels, smoking habits, and more. By identifying high-risk individuals, healthcare providers can offer preventative measures, lifestyle advice, or medical treatment. This problem statement is significant as it directly impacts public health and healthcare resources.

Who are our target audience?

- Healthcare Practitioners
- Patients

What are the values we are adding?

- **Early Prevention:** Identifying heart attack risk factors and enabling early intervention.
- **Personalized Care:** Allowing healthcare professionals to tailor treatment and lifestyle recommendations.
- **Reduced Healthcare Costs:** Lowering the financial burden of treating heart-related conditions.
- **Data-Driven Insights:** Providing valuable insights for research and public health initiatives.

Business Understanding

The project addresses a critical public health concern and has significant implications for healthcare providers, insurance companies, and individuals. Early detection of heart attack risk can lead to better patient outcomes and reduce healthcare costs.

Possible Impacts

The project's results can lead to the development of risk assessment tools, which can be used in clinical settings to identify individuals at high risk of heart attacks. This can contribute to more personalized healthcare and interventions.

Problem Statement

The main problem we aim to address is the prediction of heart attack risk. We want to develop a model that can predict an individual's likelihood of experiencing a heart attack based on a combination of medical and lifestyle factors, such as age, blood pressure, cholesterol levels, smoking habits, and more. By identifying high-risk individuals, healthcare providers can offer preventative measures, lifestyle advice, or medical treatment. This problem statement is significant as it directly impacts public health and healthcare resources.

Specifically, we aim to address the following questions:

- What are the key risk factors associated with heart attacks?
- How can we build an accurate predictive model for heart attack risk?
- How can healthcare providers use this model to improve patient care?

Related Works

We reviewed and researched several datasets, and models related to heart attack risk analysis. Some relevant sources include medical journals, research papers, and existing predictive models for heart attack risk assessment.

Relevant Research Papers:

Heart disease prediction using machine learning algorithms

Here is a part of the abstract from the above paper:

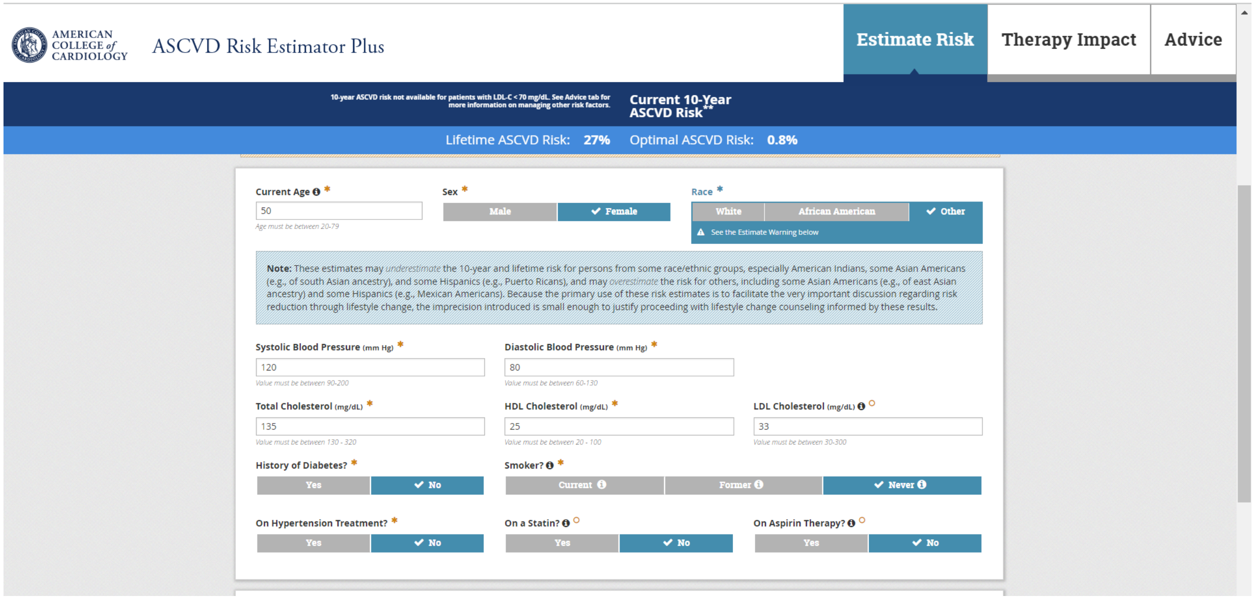
"The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with a heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as logistic regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attack in any individual. The strength of the proposed model was quite satisfying and was able to predict evidence of having a heart disease in a particular individual by using KNN and Logistic Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc."

We have taken inspiration from several verified sources on the internet for the Heart Attack Risk Prediction. This article from the Cleveland Clinic in the USA provides some insight on the Heart Disease Risks.

https://my.clevelandclinic.org/health/articles/17085-heart-risk-factor-calculators

Applications:

We have also researched the applications based on Cardiac disease which has served as a valuable



10-year ASCVD risk not available for patients with LDL-C < 70 mg/dL. See Advice tab for more information on managing other risk factors.

Current 10-Year ASCVD Risk**

Lifetime ASCVD Risk: 27% Optimal ASCVD Risk: 0.8%

Current Age * : 50
Age must be between 20-79

Sex * : Male [✓ Female]

Race * : White African American [✓ Other]
See the Estimate Warning below

Note: These estimates may *underestimate* the 10-year and lifetime risk for persons from some race/ethnic groups, especially American Indians, some Asian Americans (e.g., of south Asian ancestry), and some Hispanics (e.g., Puerto Ricans), and may *overestimate* the risk for others, including some Asian Americans (e.g., of east Asian ancestry) and some Hispanics (e.g., Mexican Americans). Because the primary use of these risk estimates is to facilitate the very important discussion regarding risk reduction through lifestyle change, the imprecision introduced is small enough to justify proceeding with lifestyle change counseling informed by these results.

Systolic Blood Pressure (mm Hg) * : 120
Value must be between 90-200

Diastolic Blood Pressure (mm Hg) * : 80
Value must be between 60-130

Total Cholesterol (mg/dL) * : 135
Value must be between 130 - 320

HDL Cholesterol (mg/dL) * : 25
Value must be between 20 - 100

LDL Cholesterol (mg/dL) ○ : 33
Value must be between 30-300

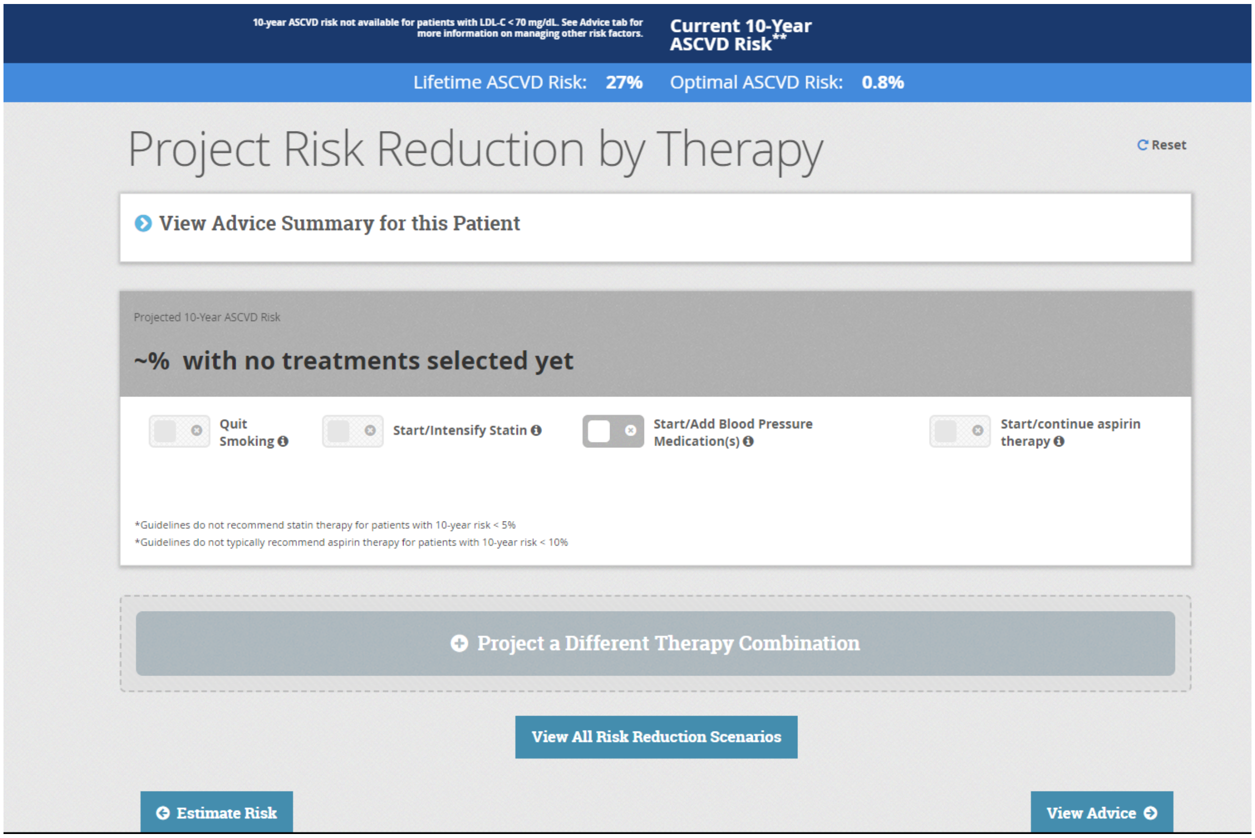
History of Diabetes? * : Yes [✓ No]

Smoker? * : Current Former [✓ Never]

On Hypertension Treatment? * : Yes [✓ No]

On a Statin? ○ : Yes [✓ No]

On Aspirin Therapy? ○ : Yes [✓ No]



10-year ASCVD risk not available for patients with LDL-C < 70 mg/dL. See Advice tab for more information on managing other risk factors.

Current 10-Year ASCVD Risk**

Lifetime ASCVD Risk: **27%** Optimal ASCVD Risk: **0.8%**

Project Risk Reduction by Therapy

Projected 10-Year ASCVD Risk

~% with no treatments selected yet

☐ Quit Smoking ☐ Start/Intensify Statin ☐ Start/Add Blood Pressure Medication(s) ☐ Start/continue aspirin therapy

*Guidelines do not recommend statin therapy for patients with 10-year risk < 5%

*Guidelines do not typically recommend aspirin therapy for patients with 10-year risk < 10%

Project a Different Therapy Combination
View All Risk Reduction Scenarios
Estimate Risk
View Advice

Datasets:

For addressing our problem statement, we have selected the following datasets. The first dataset is our main dataset that we will work on. However, in order to ensure model stability and higher accuracy, we plan on integrating other datasets that will be provided below:

Primary Dataset:

The primary dataset we plan to use is the "**Heart Attack Risk Analysis Dataset**." This dataset contains information on various patient attributes, including age, sex, blood pressure, cholesterol levels, and other lifestyle and health-related factors. Proper attribution and citation will be provided for the dataset.

The dataset, consisting of **8763 records** from patients around the globe, culminates in a crucial binary classification feature denoting the presence or absence of a heart attack risk, providing a comprehensive resource for predictive analysis and research in cardiovascular health. There are **24 features** excluding the PatientID(which is irrelevant and one will be the Target(Heart Attack Risk).

Features:

Age - Age of the patient
Sex - Gender of the patient (Male/Female)
Cholesterol - Cholesterol levels of the patient
Blood Pressure - Blood pressure of the patient (systolic/diastolic)
Heart Rate - Heart rate of the patient
Diabetes - Whether the patient has diabetes (Yes/No)
Family History - Family history of heart-related problems (1: Yes, 0: No)
Smoking - Smoking status of the patient (1: Smoker, 0: Non-smoker)
Obesity - Obesity status of the patient (1: Obese, 0: Not obese)
Alcohol Consumption - Level of alcohol consumption by the patient (None/Light/Moderate/Heavy)
Exercise Hours Per Week - Number of exercise hours per week
Diet - Dietary habits of the patient (Healthy/Average/Unhealthy)
Previous Heart Problems - Previous heart problems of the patient (1: Yes, 0: No)
Medication Use - Medication usage by the patient (1: Yes, 0: No)
Stress Level - Stress level reported by the patient (1-10)
Sedentary Hours Per Day - Hours of sedentary activity per day
Income - Income level of the patient
BMI - Body Mass Index (BMI) of the patient

Triglycerides - Triglyceride levels of the patient
Physical Activity Days Per Week - Days of physical activity per week
Sleep Hours Per Day - Hours of sleep per day
Country - Country of the patient
Continent - Continent where the patient resides
Hemisphere - Hemisphere where the patient resides

Target:
Heart Attack Risk - Presence of heart attack risk (1: Yes, 0: No)

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

Note: This dataset is a synthetic creation generated using ChatGPT to simulate a realistic experience. Its purpose is to provide a platform for beginners and data enthusiasts, allowing them to create, enjoy, practice, and learn from a dataset that mirrors real-world scenarios. The aim is to foster learning and experimentation in a simulated environment, encouraging a deeper understanding of data analysis and interpretation.

Other Datasets:

https://worldpopulationreview.com/country-rankings/hdi-by-country

https://www.kaggle.com/datasets/nitishabharathi/gdp-per-capita-all-countries

https://ourworldindata.org/grapher/cardiovascular-disease-death-rates?tab=table

https://www.prosperity.com/rankings

Currently, we are trying to tie up the country's development indexes based on several factors such as GDP, HDI and other rankings so that we can integrate it with the primary dataset. This, we believe, will ensure a better predictive model and a better result.

Since some of the datasets aren't present in csv format, we are also trying to make use of web scraping in order to fetch the data directly from the webpage.

Methodology

Step 1: Data Collection

We have gathered a comprehensive dataset that includes relevant health information for individuals. This dataset contains a wide range of features, including demographic information (e.g., age, sex), medical history (e.g., diabetes, hypertension), clinical measurements (e.g., blood pressure, cholesterol levels), lifestyle factors (e.g., smoking, physical activity), among others.

Step 2: Exploratory Data Analysis (EDA)

Before we develop a predictive model, we will conduct exploratory data analysis to understand the dataset's characteristics.

Data Visualization: We will create histograms, box plots, and scatter plots to visualize the distribution of variables and identify outliers.
Correlation Analysis: We will calculate and visualize correlations between features to identify potential relationships with heart attack risk.
Impute Missing Data: We will handle missing values through imputation methods suitable for the type of data (e.g., mean imputation, interpolation, or machine learning-based imputation).

Step 3: Data Pre-processing

We will ensure the data is clean and ready for modeling by performing the following tasks:

- **Feature Scaling:** We will normalize or standardize numerical features to bring them to a common scale, reducing the impact of features with larger values.
- **Feature Encoding:** We will convert categorical variables (if any) into numerical format using techniques like one-hot encoding.
- **Feature Selection:** We will identify the most relevant features for heart attack risk prediction through methods like feature importance scores or domain expertise.
- **Data Split:** We will split the dataset into training and testing sets (e.g., 70-30 or 80-20 split) to evaluate the model's performance.

Step 4 : Model Selection and Training

We will experiment with various machine learning models and algorithms to identify the one that provides the best heart attack risk prediction. We will make use of several classification algorithms in order to ensure the best model stability and prediction accuracy. Some common algorithms to consider include:

Logistic Regression: A linear model used for probabilistic binary classification.
Decision Trees: Non-linear models that can capture complex decision boundaries.
Random Forest: An ensemble model consisting of multiple decision trees.
Support Vector Machine (SVM): Effective for finding optimal hyperplanes that separate classes.
Naive Bayes: A probabilistic model based on Bayes' theorem.

We will perform several techniques such as cross-validation, Grid search to ensure the best model selection.

Step 5: Model Evaluation:

We will assess the model's performance using the reserved testing dataset. The model's performance will be evaluated using appropriate metrics like accuracy, precision, recall, and F1-score.

Sensitivity (True Positive Rate): The ability to correctly identify individuals at risk of a heart attack.
Specificity (True Negative Rate): The ability to correctly identify individuals not at risk of a heart attack.
ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): A measure of the model's ability to discriminate between positive and negative cases.
Precision-Recall Curve: A visualization of the trade-off between precision and recall.

Step 6: Model Deployment:

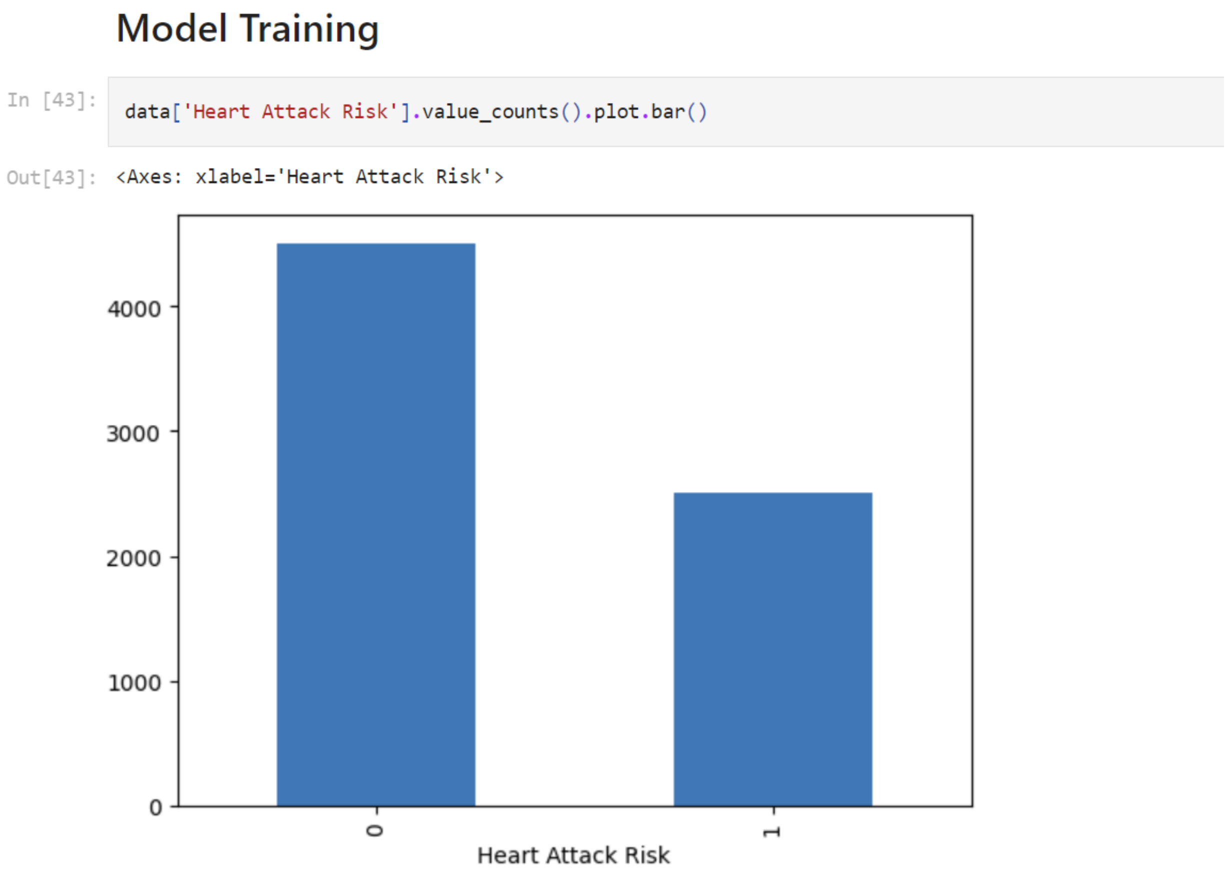
Once a reliable model is developed and evaluated, we will consider options for deployment. We will use either Flask, Streamlit, Dash or some other program for web development.

Preliminary Results

At this stage, we have provided an initial analysis of the dataset, including any significant findings from the exploratory data analysis. Based on the findings, we will work towards improving the Modeling process.

Here are some of the screenshots of the results:

Screenshot 1:

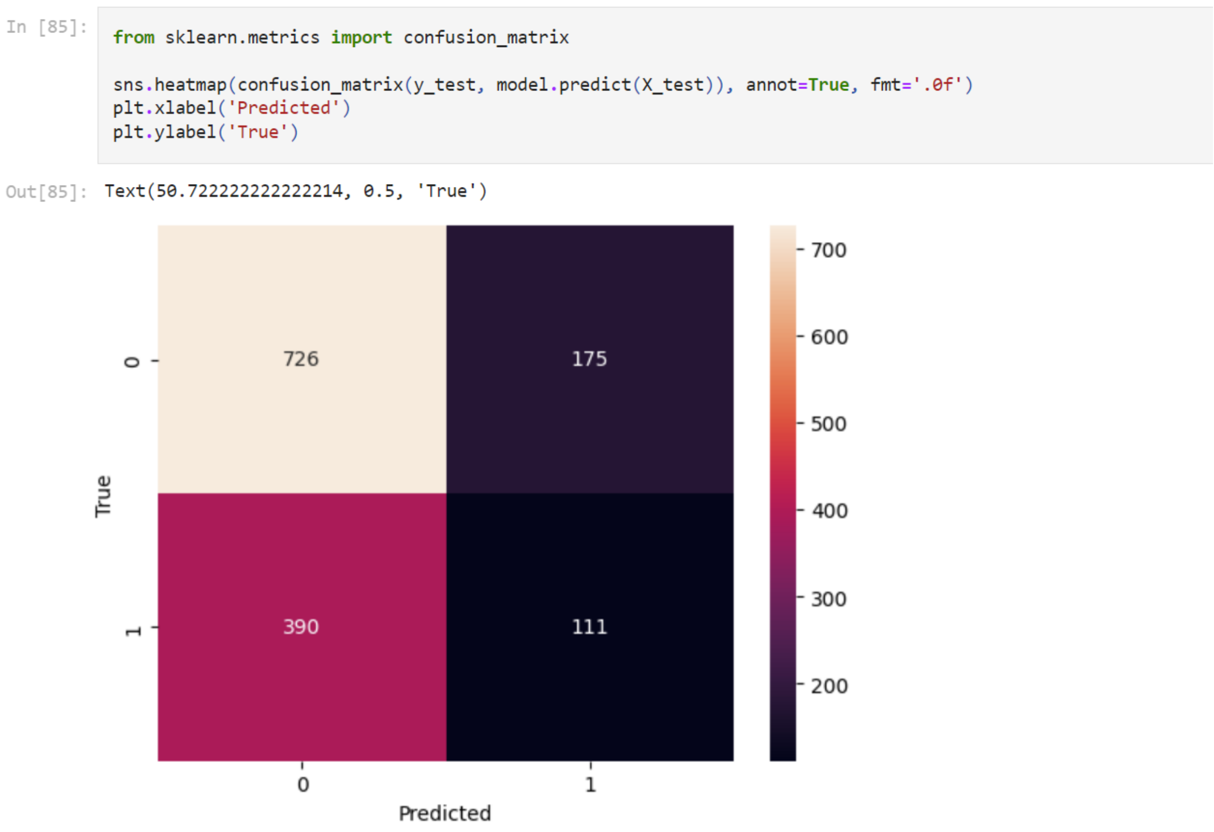


Model Training

```
In [43]: data['Heart Attack Risk'].value_counts().plot.bar()
```

```
Out[43]: <Axes: xlabel='Heart Attack Risk'>
```

Screenshot 2:



```
In [85]:  from sklearn.metrics import confusion_matrix

          sns.heatmap(confusion_matrix(y_test, model.predict(X_test)), annot=True, fmt='.0f')
          plt.xlabel('Predicted')
          plt.ylabel('True')

Out[85]:  Text(50.722222222222214, 0.5, 'True')
```

Github Link: https://github.com/thassung/AIT_cp2023/tree/main

Preliminary Findings Analysis:

Data Overview: The dataset consists of 26 features(24 excluding target variable and patient Id), including demographic information, health-related metrics, lifestyle choices, and geographical details. The target variable is "Heart Attack Risk," which is binary (1 for Yes and 0 for No).

Data Preprocessing: Before modeling, the data underwent preprocessing steps. Categorical variables like "Sex," "Diet," and "Country" were encoded into numerical values for modeling. The "Blood Pressure" column was split into systolic and diastolic values. Redundant columns like "Continent" and "Hemisphere" were dropped since "Country" information already contains geographical details.

Imbalanced Data: The bar chart of the "Heart Attack Risk" column shows that the dataset is imbalanced, with more records labeled as "No" (0) than "Yes" (1) for heart attack risk.

Initial Model Training: Two classification models were tested without hyperparameter tuning: Random Forest and XGBoost. The initial results showed moderate F1 scores in the range of 0.55 to 0.58, indicating some predictive power. However, there is room for improvement.

Class Imbalance Handling: To address the class imbalance issue, a random undersampling technique was applied to balance the dataset. After resampling, the models were retrained, but the F1 scores remained in the similar range (around 0.48 to 0.50).

Confusion Matrix: The confusion matrix was plotted for the models, showing the number of true positives, true negatives, false positives, and false negatives. While random undersampling balanced the dataset, it did not significantly improve model performance, as indicated by the confusion matrices.

Next Steps: The preliminary findings suggest that there is room for model improvement. Further steps may include hyperparameter tuning for the models, exploring different feature engineering techniques, and considering additional strategies for handling class imbalance, such as oversampling or using different evaluation metrics. Additionally, feature importance analysis can help identify the most influential factors in predicting heart attack risk.

Expected Outcomes:

We anticipate the following outcomes:

- A predictive model that offers accurate heart attack risk assessments.
- Insights into the most influential risk factors for heart attacks.
- A practical tool for healthcare providers to enhance patient care and preventive measures.

Conclusion

This project aims to provide valuable insights into heart attack risk factors and the development of an accurate predictive model. The results can be used in clinical practice to identify high-risk individuals and improve patient care. By making this project publicly available on platforms like LinkedIn or GitHub, we hope to contribute to the field of cardiovascular health and data science.