Project Title : Heart Attack Risk Prediction

Humayera Tabassum Tunan

Table of Contents

Introduction

Globally, heart disease is one of the main causes of death. The therapy and prevention of cardiovascular illnesses greatly depend on the capacity to forecast the risk of heart attacks. In this study, we forecast the risk of heart attack based on multiple lifestyle and health characteristics by using machine learning algorithms.

Dataset Description

Source:

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

Acknowledgement:

This dataset is a synthetic creation generated using ChatGPT.

Thumbnail by: vectorjuice on Freepik

Dataset Description:

We have 26 features in our dataset.

1. Patient ID - Unique identifier for each patient
2. Age - Age of the patient
3. Sex - Gender of the patient (Male/Female)
4. Cholesterol - Cholesterol levels of the patient
5. Blood Pressure - Blood pressure of the patient (systolic/diastolic)
6. Heart Rate - Heart rate of the patient
7. Diabetes - Whether the patient has diabetes (Yes/No)
8. Family History - Family history of heart-related problems (1: Yes, 0: No)
9. Smoking - Smoking status of the patient (1: Smoker, 0: Non-smoker)
10. Obesity - Obesity status of the patient (1: Obese, 0: Not obese)

11. Alcohol Consumption - Level of alcohol consumption by the patient (None/Light/Moderate/Heavy)
12. Exercise Hours Per Week - Number of exercise hours per week
13. Diet - Dietary habits of the patient (Healthy/Average/Unhealthy)
14. Previous Heart Problems - Previous heart problems of the patient (1: Yes, 0: No)
15. Medication Use - Medication usage by the patient (1: Yes, 0: No)
16. Stress Level - Stress level reported by the patient (1-10)
17. Sedentary Hours Per Day - Hours of sedentary activity per day
18. Income - Income level of the patient
19. BMI - Body Mass Index (BMI) of the patient
20. Triglycerides - Triglyceride levels of the patient
21. Physical Activity Days Per Week - Days of physical activity per week
22. Sleep Hours Per Day - Hours of sleep per day
23. Country - Country of the patient
24. Continent - Continent where the patient resides
25. Hemisphere - Hemisphere where the patient resides
26. Heart Attack Risk - Presence of heart attack risk (1: Yes, 0: No)

How many features?

There are 25 features in the dataset.

Classification or regression problem?

It is a classification problem because the aim is to predict whether a person is at risk of a heart attack based on various health-related factors. The target variable "Heart Attack Risk" is categorical.

How many data points?

There are 8763 data points (patient records) in the dataset.

What kind of features are in your dataset?

Numerical Features:

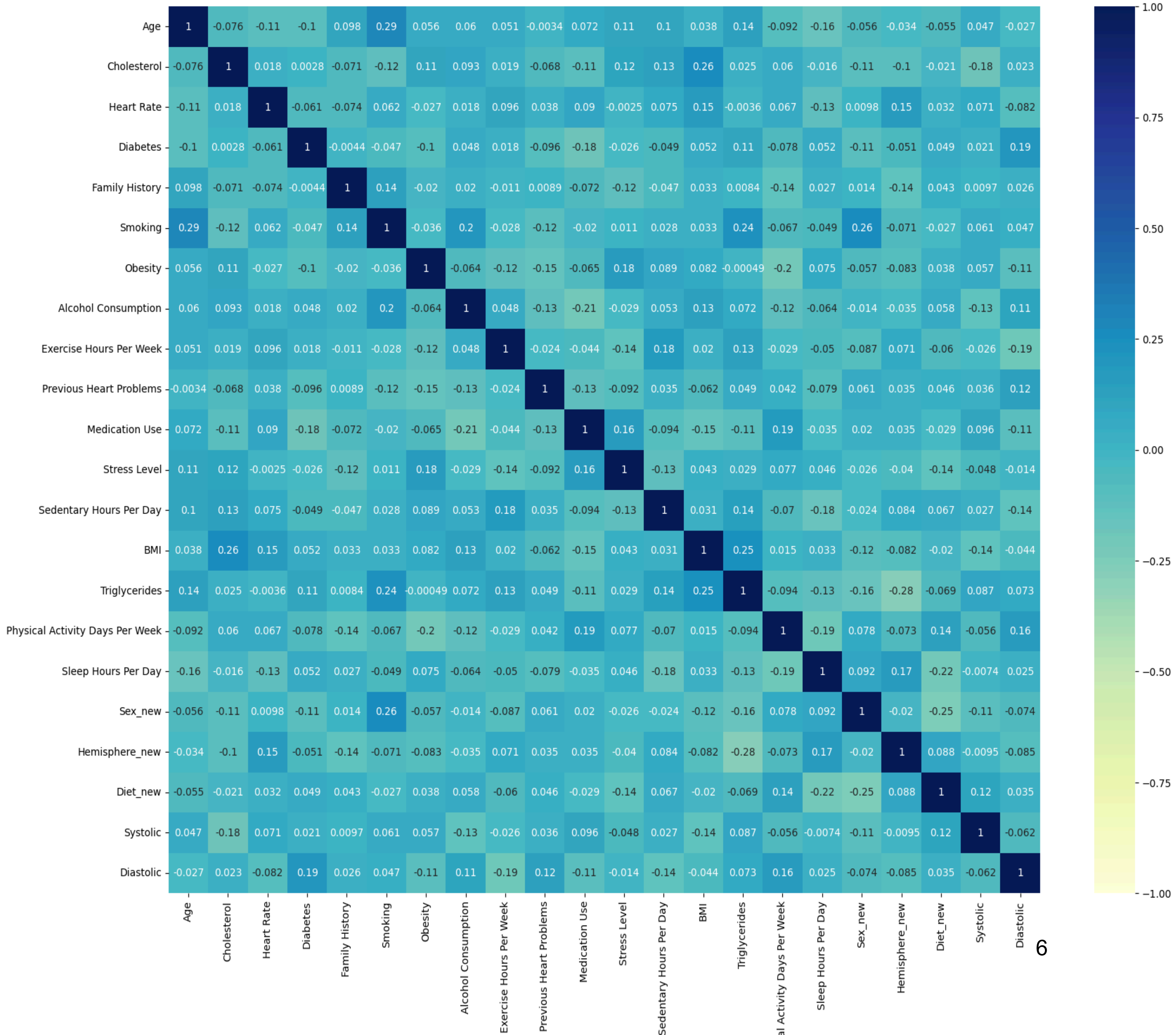
1. Age
2. Cholesterol
3. Blood Pressure
4. Heart Rate
5. Exercise Hours Per Week
6. Previous Heart Problems
7. Stress Level
8. Sedentary Hours Per Day
9. Income
10. BMI
11. Triglycerides
12. Physical Activity Days Per Week
13. Sleep Hours Per Day

Categorical Features:

1. Sex
2. Diabetes
3. Family History
4. Smoking
5. Obesity

6. Alcohol Consumption
7. Diet
8. Medication Use
9. Country
10. Continent
11. Hemisphere

Correlation of all features:



	Age	Cholesterol	Heart Rate	Diabetes	Family History	Smoking	Obesity	Alcohol Consumption	Exercise Hours Per Week	Previous Heart Problems	Medication Use	Stress Level	Sedentary Hours Per Day	BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day	Sex_new	Hemisphere_new	Diet_new	Systolic	Diastolic
Age	1	-0.076	-0.11	-0.1	0.098	0.29	0.056	0.06	0.051	-0.0034	0.072	0.11	0.1	0.038	0.14	-0.092	-0.16	-0.056	-0.034	-0.055	0.047	-0.027
Cholesterol	-0.076	1	0.018	0.0028	-0.071	-0.12	0.11	0.093	0.019	-0.068	-0.11	0.12	0.13	0.26	0.025	0.06	-0.016	-0.11	-0.1	-0.021	-0.18	0.023
Heart Rate	-0.11	0.018	1	-0.061	-0.074	0.062	-0.027	0.018	0.096	0.038	0.09	-0.0025	0.075	0.15	-0.0036	0.067	-0.13	0.0098	0.15	0.032	0.071	-0.082
Diabetes	-0.1	0.0028	-0.061	1	-0.0044	-0.047	-0.1	0.048	0.018	-0.096	-0.18	-0.026	-0.049	0.052	0.11	-0.078	0.052	-0.11	-0.051	0.049	0.021	0.19
Family History	0.098	-0.071	-0.074	-0.0044	1	0.14	-0.02	0.02	-0.011	0.0089	-0.072	-0.12	-0.047	0.033	0.0084	-0.14	0.027	0.014	-0.14	0.043	0.0097	0.026
Smoking	0.29	-0.12	0.062	-0.047	0.14	1	-0.036	0.2	-0.028	-0.12	-0.02	0.011	0.028	0.033	0.24	-0.067	-0.049	0.26	-0.071	-0.027	0.061	0.047
Obesity	0.056	0.11	-0.027	-0.1	-0.02	-0.036	1	-0.064	-0.12	-0.15	-0.065	0.18	0.089	0.082	-0.00049	-0.2	0.075	-0.057	-0.083	0.038	0.057	-0.11
Alcohol Consumption	0.06	0.093	0.018	0.048	0.02	0.2	-0.064	1	0.048	-0.13	-0.21	-0.029	0.053	0.13	0.072	-0.12	-0.064	-0.014	-0.035	0.058	-0.13	0.11
Exercise Hours Per Week	0.051	0.019	0.096	0.018	-0.011	-0.028	-0.12	0.048	1	-0.024	-0.044	-0.14	0.18	0.02	0.13	-0.029	-0.05	-0.087	0.071	-0.06	-0.026	-0.19
Previous Heart Problems	-0.0034	-0.068	0.038	-0.096	0.0089	-0.12	-0.15	-0.13	-0.024	1	-0.13	-0.092	0.035	-0.062	0.049	0.042	-0.079	0.061	0.035	0.046	0.036	0.12
Medication Use	0.072	-0.11	0.09	-0.18	-0.072	-0.02	-0.065	-0.21	-0.044	-0.13	1	0.16	-0.094	-0.15	-0.11	0.19	-0.035	0.02	0.035	-0.029	0.096	-0.11
Stress Level	0.11	0.12	-0.0025	-0.026	-0.12	0.011	0.18	-0.029	-0.14	-0.092	0.16	1	-0.13	0.043	0.029	0.077	0.046	-0.026	-0.04	-0.14	-0.048	-0.014
Sedentary Hours Per Day	0.1	0.13	0.075	-0.049	-0.047	0.028	0.089	0.053	0.18	0.035	-0.094	-0.13	1	0.031	0.14	-0.07	-0.18	-0.024	0.084	0.067	0.027	-0.14
BMI	0.038	0.26	0.15	0.052	0.033	0.033	0.082	0.13	0.02	-0.062	-0.15	0.043	0.031	1	0.25	0.015	0.033	-0.12	-0.082	-0.02	-0.14	-0.044
Triglycerides	0.14	0.025	-0.0036	0.11	0.0084	0.24	-0.00049	0.072	0.13	0.049	-0.11	0.029	0.14	0.25	1	-0.094	-0.13	-0.16	-0.28	-0.069	0.087	0.073
Physical Activity Days Per Week	-0.092	0.06	0.067	-0.078	-0.14	-0.067	-0.2	-0.12	-0.029	0.042	0.19	0.077	-0.07	0.015	-0.094	1	-0.19	0.078	-0.073	0.14	-0.056	0.16
Sleep Hours Per Day	-0.16	-0.016	-0.13	0.052	0.027	-0.049	0.075	-0.064	-0.05	-0.079	-0.035	0.046	-0.18	0.033	-0.13	-0.19	1	0.092	0.17	-0.22	-0.0074	0.025
Sex_new	-0.056	-0.11	0.0098	-0.11	0.014	0.26	-0.057	-0.014	-0.087	0.061	0.02	-0.026	-0.024	-0.12	-0.16	0.078	0.092	1	-0.02	-0.25	-0.11	-0.074
Hemisphere_new	-0.034	-0.1	0.15	-0.051	-0.14	-0.071	-0.083	-0.035	0.071	0.035	0.035	-0.04	0.084	-0.082	-0.28	-0.073	0.17	-0.02	1	0.088	-0.0095	-0.085
Diet_new	-0.055	-0.021	0.032	0.049	0.043	-0.027	0.038	0.058	-0.06	0.046	-0.029	-0.14	0.067	-0.02	-0.069	0.14	-0.22	-0.25	0.088	1	0.12	0.035
Systolic	0.047	-0.18	0.071	0.021	0.0097	0.061	0.057	-0.13	-0.026	0.036	0.096	-0.048	0.027	-0.14	0.087	-0.056	-0.0074	-0.11	-0.0095	0.12	1	-0.062
Diastolic	-0.027	0.023	-0.082	0.19	0.026	0.047	-0.11	0.11	-0.19	0.12	-0.11	-0.014	-0.14	-0.044	0.073	0.16	0.025	-0.074	-0.085	0.035	-0.062	1

Data pre- processing

Problem 1: Null Values.

[The dataset did not have any null values so I inserted 16,000 random null values for demonstration]

Solution: Removing Null Values

In order to remove null values from the dataset, I found some data points where the highly important features had null values. These data points will be of no use to our project, that's why I dropped them using dropna().

Problem 2: Irrelevant features.

Solution: Feature selection

There were some features['Patient ID', 'Continent', 'Country', ‘Income’] that were irrelevant to our project, which will be of no use. I dropped those features using drop().

Problem 3: Missing values.

Solution: Imputing Missing Values

Some of the numerical features['Sedentary Hours Per Day','Physical Activity Days Per Week', 'Sleep Hours Per Day', 'Income', 'Exercise Hours Per Week'] had some missing values. I replaced those missing values with the average of all values in those features by using mean(). For the binary features['Family History', 'Smoking', 'Obesity', 'Alcohol Consumption'] I replaced the missing values with 0 by using fillna(0).

Problem 4: Categorical Features.

Solution: Labeling/One hot Encoding

There were three categorical features: “Diet”, “Sex” and “Hemisphere”. I encoded “Sex” and “Hemisphere” into binary values using LabelEncoder(). For “Diet”, I used map() function and mapped 'Healthy' as 2, 'Unhealthy' as 0, and 'Average' as 1. We can also solve this using One Hot encoding, as shown for demonstration.

Problem 5: Features with multiple values.

Solution: Feature Engineering.

I separated the feature “Blood Pressure” into two different features Systolic and Diastolic.

Feature Scaling

Features that we edited were removed, also unnecessary features were removed. No further Feature scaling was required.

Dataset Splitting

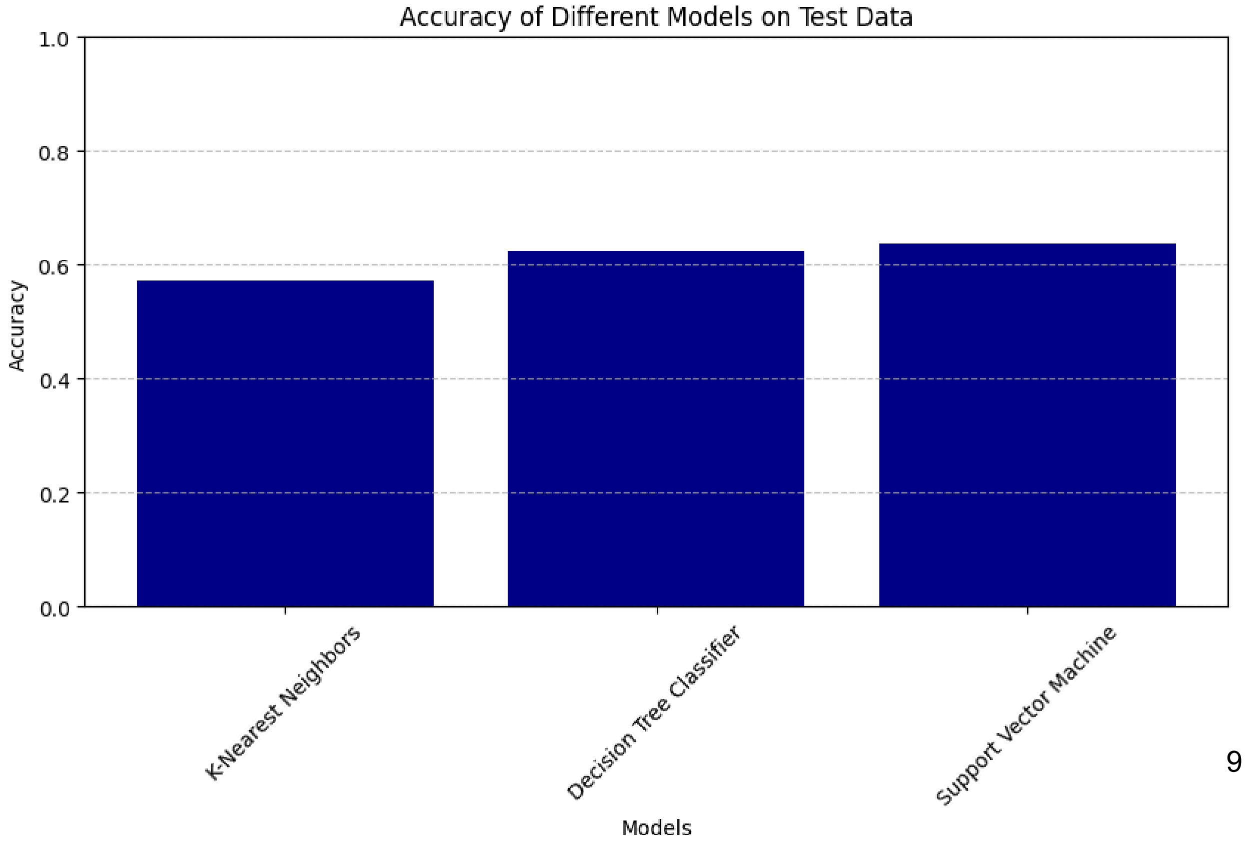
Random Split is done on the dataset using the train_test_split function from SK learn library. Test_size = 0.3 represents the 30% data used for training and the remaining 70% were used for testing .

Model Training and Testing

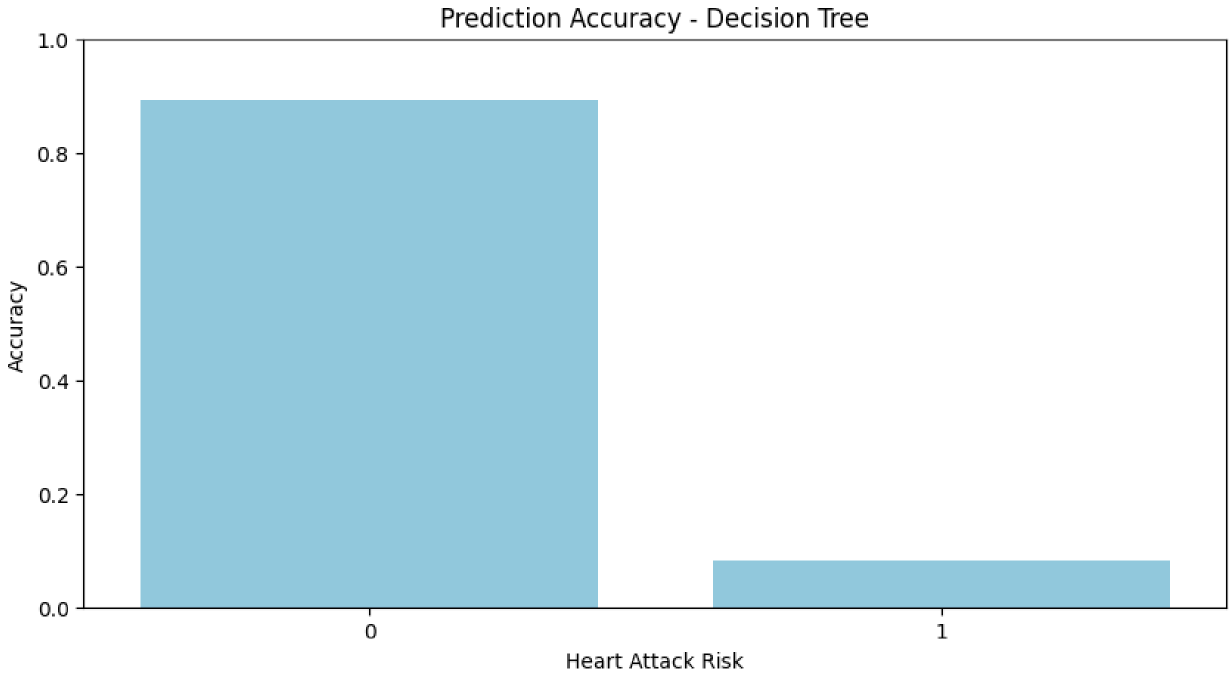
I used the following three Machine Learning mopeds to train and test the dataset.

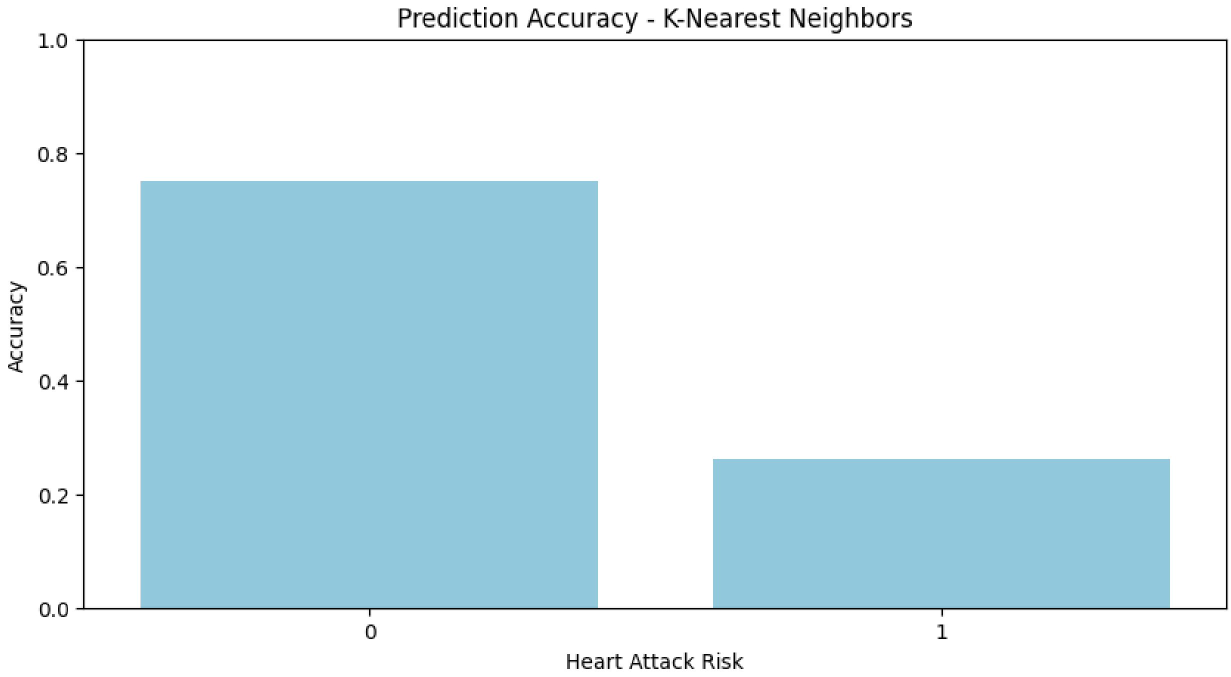
1. Decision Tree (RandomForestClassifier)
2. K-Nearest Neighbors (KNeighborsClassifier)
3. Support Vector Machine (SVC)

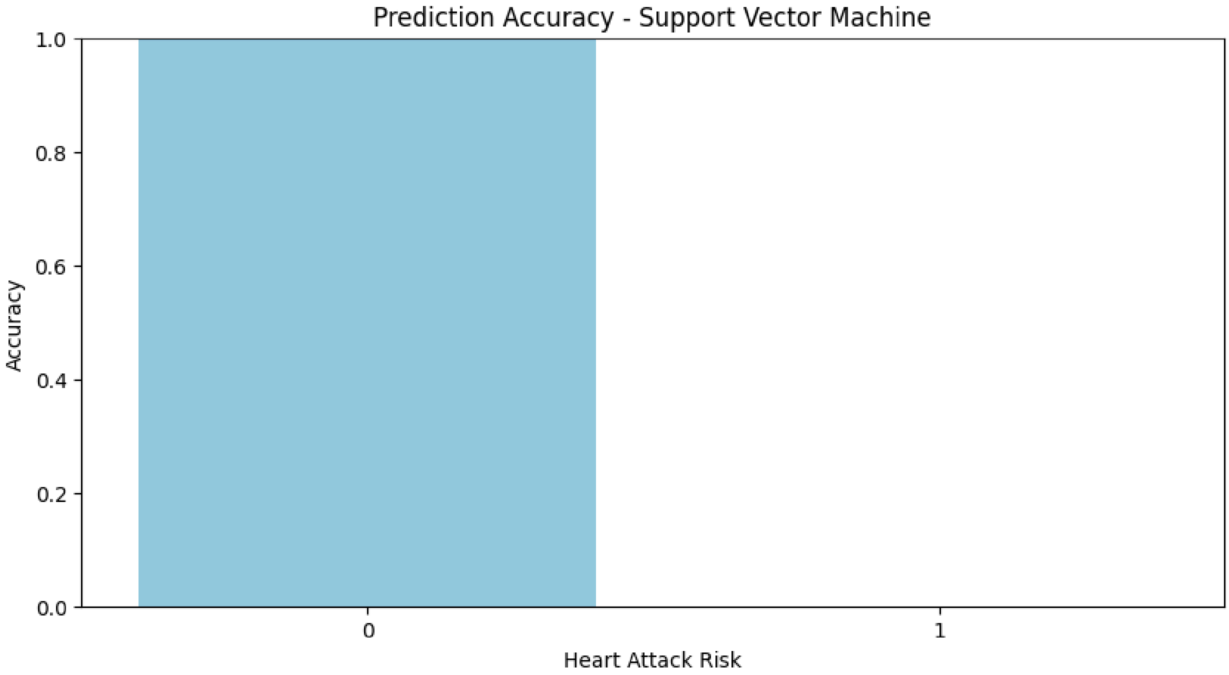
Model Selection / Comparison Analysis



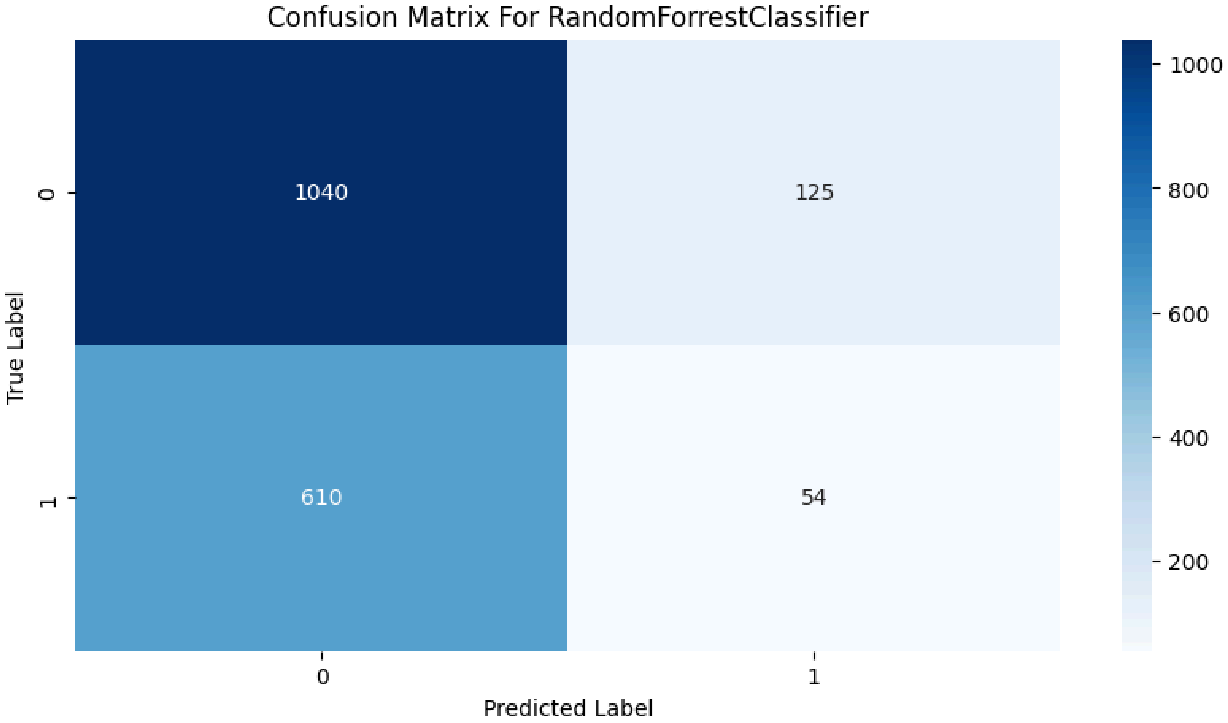
Bar chart showcasing prediction accuracy of all models:

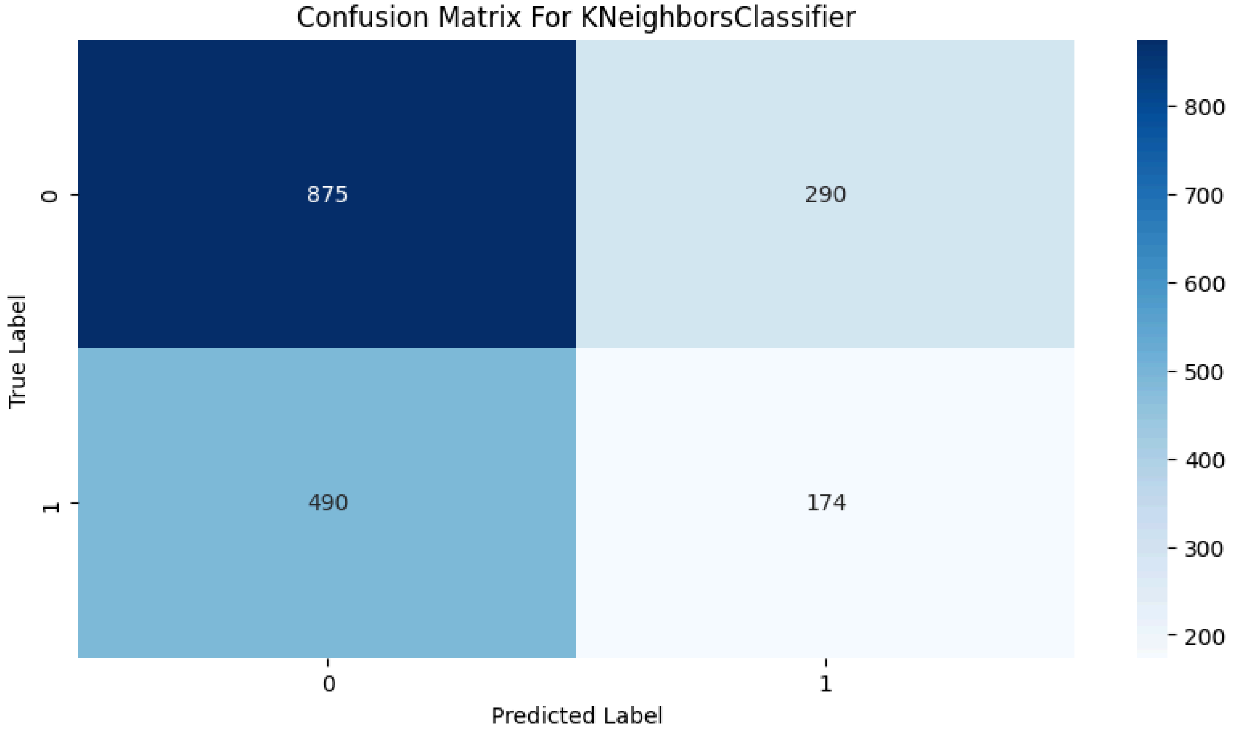

Prediction Accuracy - Decision Tree


Prediction Accuracy - K-Nearest Neighbors

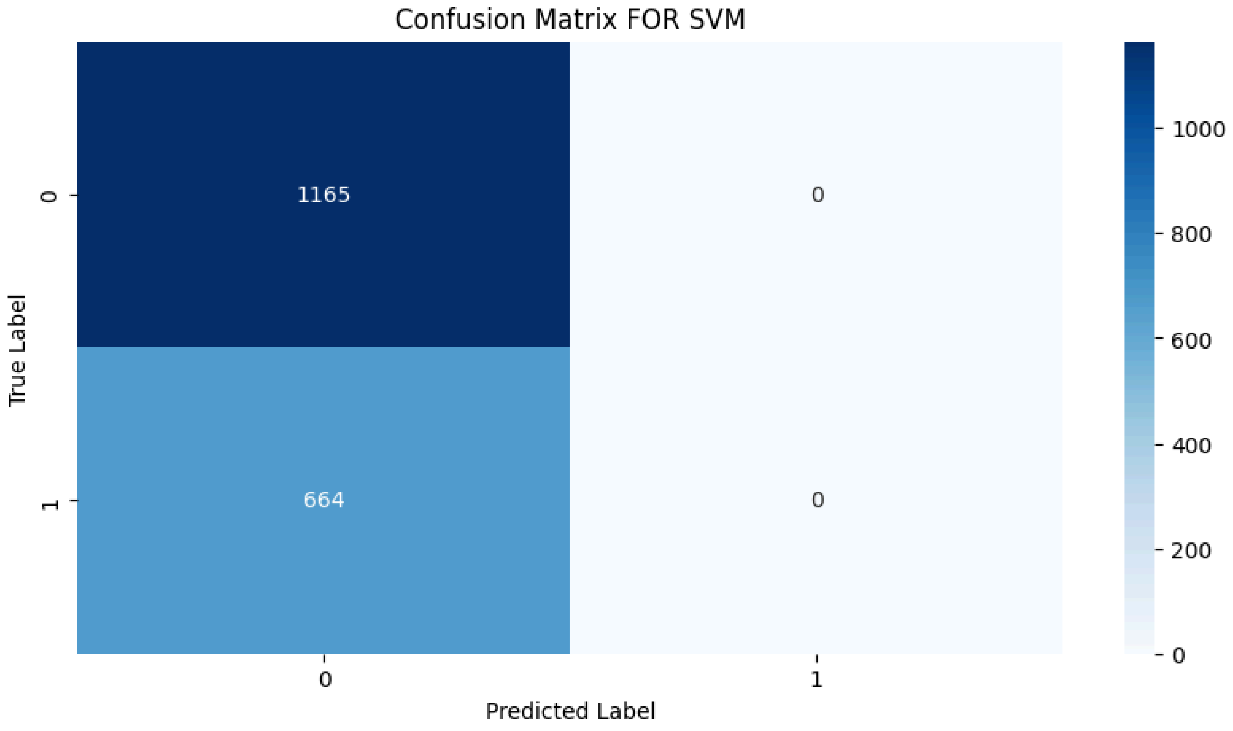


Confusion Matrix:





Confusion Matrix For KNeighborsClassifier



Confusion Matrix FOR SVM

Conclusion

The analysis conducted for the project showed that a number of health indicators significantly affect the chance of having a heart attack. The project's goal is to analyze these data in order to gain important knowledge about possible risk factors that will enable early intervention and prevention. The “Heart Attack Risk Prediction” project is a big step in the right direction toward creating a useful tool for heart attack risk prediction. This initiative intends to help with early diagnosis of heart-related conditions and better healthcare outcomes by utilizing machine learning techniques.