

Heart Attack Prediction Using Machine Learning

By Natalie Miner

AUTHOR BIO

Natalie Miner is currently a senior at Mound Westonka High School. She is particularly interested in applied math, machine learning, and business. Natalie has been part of her school's DECA program since 9th grade, which is a business club aimed at career development. She is a Minnesota first place State Recipient and International Finalist in DECA. Additionally, she is the founder and president of Like A Girl, an empowerment-focused project for middle and high school girls to help them build confidence and leadership skills. Natalie is also her school's Student Body President and is involved in Soccer, Hockey, and Track.

ABSTRACT

Heart disease is one of the leading causes of mortality due to heart attacks and other complications. When a person has a heart attack, their heart muscle begins to die because of a lack of blood flow due to a blocked or reduced blood supply. The most severe effects of a heart attack can be reduced with prompt identification and treatment. Machine learning can help medical professionals make an earlier diagnosis of heart disease and start treatment, reducing complications and saving lives. Using various attributes relevant to detecting heart disease, a machine-learning model was created that predicts a person's risk of heart disease. The logistic regression model created in the study has an overall accuracy of 85%. Ultimately, this model could play an important role in predicting heart disease and preventing the severe effects of a heart attack.

INTRODUCTION AND BACKGROUND

A heart attack is a life-threatening medical emergency that requires immediate treatment. When a person has a heart attack, they experience a sudden blockage or reduction of the flow of oxygen-bearing blood to the heart. Heart attacks are often a direct result of heart disease. This occurs when coronary arteries that supply the heart muscle with blood flow become narrowed by a buildup of fat, cholesterol, and other substances that are called plaque (Mayo Clinic). The buildup of plaque, called Atherosclerosis, slowly progresses to narrow the artery, but if the plaque ruptures a blood clot can form on it suddenly, blocking all blood flow in that artery, and causing a heart attack. According to the Centers for Disease Prevention and Control, more than 800,000 people in the United States have a heart attack yearly. Furthermore, about every 40 seconds, someone in the United States has a heart attack. Factors including a person's age, gender, lifestyle habits, and other medical conditions can raise a person's risk of heart disease and heart attacks.

A heart needs oxygen to survive; if blood flow is not restored quickly, the heart muscle will begin to die. The amount of damage to the heart muscle depends on the time between the loss of blood flow and treatment. Therefore, prompt treatment is needed to prevent death or severe damage. Half the deaths from a heart attack occur in the first 3 or 4 hours after symptoms begin (Cedars Sinai), so receiving treatment quickly and efficiently for a heart attack is a crucial part of limiting severe damage to the heart and death. Machine learning can help position professionals to make an earlier diagnosis and prevent severe effects.

Machine learning is a data analysis method that focuses on building systems that are able to learn and adapt without following

explicit instructions. It uses algorithms and statistical models to analyze and predict patterns in a set of data. In a healthcare setting, machine learning can help medical professionals make quicker, more accurate diagnoses leading to improved patient outcomes. The primary purpose of this study is to use machine learning to create an efficient logistic regression model that predicts a person's risk of a heart attack.

METHODS

Exploration of the dataset

The dataset is from the UCI Machine Learning Repository. The dataset includes 13 features, 7 categorical and 6 numerical, and one label. There is information from 303 different patients in the dataset and there are no missing values. Most of the features are normally distributed with no significant skewing as shown in Table 1 below. A small number of features did not distribute normally.

	count	mean	std	min	25%	50%	75%	max
age	303.0	54.399337	5.016719	29.0	47.0	55.0	61.0	71.0
sex	303.0	0.663168	0.469311	0.0	0.0	1.0	1.0	1.0
cp	303.0	0.969967	1.030052	0.0	0.0	1.0	2.0	3.0
trestbps	303.0	137.822762	17.058743	94.0	120.0	130.0	140.0	200.0
chol	303.0	246.264028	91.830791	126.0	210.0	240.0	274.0	584.0
fbs	303.0	0.148875	0.368788	0.0	0.0	0.0	0.0	1.0
restecg	303.0	0.526653	0.505860	0.0	0.0	1.0	1.0	2.0
thalach	303.0	159.648866	22.804787	71.0	133.0	153.0	166.0	202.0
exersecg	303.0	0.326773	0.469784	0.0	0.0	0.0	1.0	1.0
oldpeak	303.0	1.028654	1.161075	0.0	0.0	0.8	1.6	6.2
slo	303.0	1.386940	0.016226	0.0	1.0	1.0	2.0	2.0
caa	303.0	0.729073	1.022606	0.0	0.0	0.0	1.0	4.0
thal	303.0	2.373831	0.872277	0.0	2.0	2.0	3.0	3.0
output	303.0	0.544854	0.498803	0.0	0.0	1.0	1.0	1.0

Table 1: Distribution of each feature in the dataset

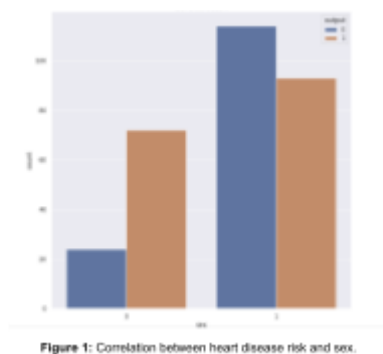
The understanding and interpretation of the features in this dataset are crucial to interpreting patterns within the data and correlations between different features. Each feature is shown to have an important role in the development of heart disease.

1. Age (age) - numerical

As a person's age increases, their risk of heart attack also increases. Aging causes changes in the heart and blood vessels that may increase a person's risk of heart disease.

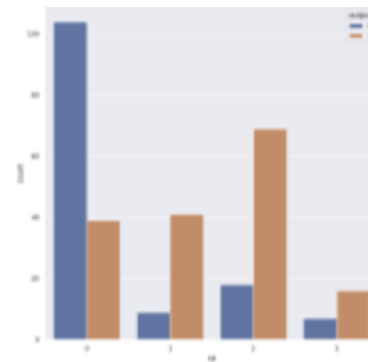
2. Sex (sex) - categorical

Heart attacks are twice as common in men than in women. Women's naturally occurring hormone levels may protect against heart disease until menopause. After menopause, a woman's risk increases to match that of men (National Institutes of Health). There is a strong correlation between heart disease and sex as shown in Figure 1, where the orange bars signify a high risk of heart disease. A value of 0 indicates female and a value of 1 indicates male.



3. Chest pain (cp) - categorical

Most heart attacks include symptoms of chest pain or discomfort. This discomfort can feel like uncomfortable pressure, squeezing, fullness, or pain. This can be seen in four different forms: typical angina (0), atypical angina (1), non-anginal pain (2), or some may be asymptomatic (3). There is a correlation between chest pain and a risk of heart attack as shown in Figure 2. The orange bars signify a high risk of heart disease.



4. Blood pressure (trtbps) - numerical

High blood pressure can cause plaque to build up in the arteries. Therefore, the flow of blood through the heart muscle is interrupted, resulting in a heart attack (National Heart Association). A normal, healthy adult has an average blood pressure of approximately 120/80 mm Hg. Blood pressure above this can signify a risk of heart attack. The blood pressure is measured in millimeters of mercury (mm Hg).

5. Cholesterol (chol) - numerical

High levels of cholesterol can result in fatty deposits in the blood vessels. These deposits can make it difficult for enough blood to flow through the arteries. Cholesterol is measured in milligrams per deciliter and is fetched by a BMI sensor.

6. Fasting blood sugar (fbs) - categorical

High fasting blood sugar levels can lead to the buildup of plaque in the arteries and interrupt blood flow. A fasting blood sugar above 120 mg/dl is an indicator of diabetes and contributes to a higher risk of heart disease (National Institutes of Health). A value of 0 indicates blood sugar above 120 mg/dl and a value of 1 indicates a value lower.

7. *Resting electrocardiographic results (restecg) - categorical*

ECG is a test that records the electrical activity of the heart while a person is at rest. Abnormal tests can be used as evidence of coronary heart disease. A value of 0 indicates normal results, a value of 1 indicates ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV, and a value of 2 indicates probable or definite left ventricular hypertrophy by Estes' criteria. A correlation can be seen between these results and heart disease in Figure 3, where the orange bars signify a high risk of heart disease.

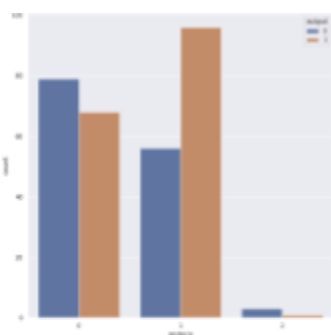


Figure 3: Correlation between heart disease risk and resting electrocardiographic results.

8. *Maximum heart rate achieved (thalach) - numerical*

The maximum heart rate achieved through exercise or stress can be an indicator of the heart's strength and ability to handle exertion.

9. *Exercise-induced angina (exng) - categorical*

Exercise-induced angina is chest pain that is a result of exercise or extreme stress. The heart muscle does not get enough blood or oxygen that it needs at a high activity level. A value of 0 indicates no and a value of 1 indicates yes. Figure 4 shows the correlation between

exercise-induced angina and the risk of a heart attack. The orange bars signify a high risk of heart disease.

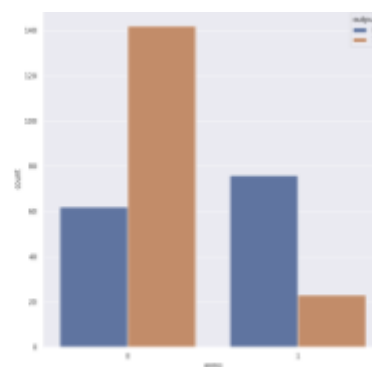


Figure 4: Correlation between heart disease risk and exercise-induced angina.

10. *ST depression induced by exercise relative to rest (oldpeak) - numerical*

ST depression occurs when the ST segment is abnormally low and appears below the baseline in a person's results. ST depression can indicate a lack of sufficient blood flow to the heart muscle (National Institutes of Health).

11. *The slope of the peak exercise ST segment (slp) - categorical*

The slope of the peak exercise ST segment can indicate the relative oxygen demand by the heart during exercise. This can reveal the overall health and condition of the heart. A value of 0 indicates upsloping, a value of 1 indicates a flat slope, and a value of 2 indicates downsloping.

12. *Number of major vessels colored by Fluoroscopy (caa) - numerical*

The number of major vessels colored by fluoroscopy reveals the measure of the presence of disease in the major blood vessels to the heart. The higher the number, the higher risk of severe disease.

13. Thallium stress test (thall) - categorical

A thallium stress test is an imaging test that indicates how well blood flows into the heart while exercising or at rest. This test can show areas of the heart muscle that aren't receiving enough blood, which is a sign of heart disease according to UCSF Health. A value of 1 indicates normal flow, a value of 2 indicates a fixed defect and a value of 3 indicates a reversible defect. A fixed defect poses a more significant problem. The orange bars signify a high risk of heart disease.

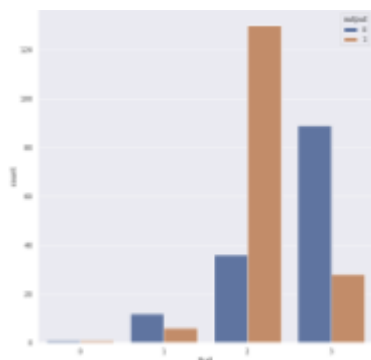


Figure 5: Correlation between heart disease risk and thallium stress test.

14. Target (diagnosis of heart disease) - categorical

This is the outcome of the prediction. A value of 0 indicates no significant heart disease and a value of 1 indicates a significant risk of heart disease.

PREPARATION OF DATASET FOR MACHINE LEARNING

Some features in this dataset are multicategorical. Because of this, the data must be one hot encoded so that it can be used in the model. One hot encoding will create new columns as much as the number of unique categories for the feature. The one hot encoded dataset can be seen in Table 2. This creates a

form that can be provided to machine learning algorithms to run a prediction model.

	age	trtbps	chol	skatrbks	sktpeak	caa	exer	exr_1	exr_2	exr_3	fta_1	resting_1	resting_2	resting_3	sig_1	sig_2	sig_3
0	63	160	260	100	2.0	0	1	1	0	0	1	1	0	0	0	0	0
1	37	130	250	107	3.0	0	1	1	0	1	0	0	1	0	0	0	0
2	41	150	204	172	1.4	0	1	0	1	0	0	0	0	0	0	0	0
3	58	150	238	119	3.8	0	1	1	1	0	0	0	1	0	0	0	0
4	57	120	264	103	3.0	0	1	0	0	0	0	0	1	0	0	1	0

Table 2: One hot encoded dataset.

There are some significant numerical outliers in the data. These values can be deleted to create a more accurate model. Figure 6 shows the code below where these outliers are eliminated by calculating the upper and lower boundary and deleting values outside of this range.

```
[11]:
for i in numeric_list:
    Q1 = np.percentile(df.loc[:, i], 25)
    Q3 = np.percentile(df.loc[:, i], 75)
    IQR = Q3 - Q1
    upper = np.where(df.loc[:, i] == (Q3 + 2.5*IQR))
    lower = np.where(df.loc[:, i] == (Q1 - 2.5*IQR))
    df.drop(upper[0], inplace = True)
    df.drop(lower[0], inplace = True)
```

Figure 6: Eliminate outliers

As a result, there are 298 data points without changing the number of features in the dataset.

MACHINE LEARNING MODEL

To start the prediction model, the data set was split into a training set and a validation set, with the training set being 75% and the validation set being 25%. The training set is used to train the model for the prediction of heart disease risk. The validation is then used to validate that the model accurately represents the data. This helps to check for overfitting later in the project, where the model gives accurate predictions for the training data, but not for any new data. The code for this step is seen in Figure 7.

```

1071: print('had a heart attack', (y == 1).sum())
1072: print('did not have a heart attack', (y == 0).sum())
1073:
1074: had a heart attack 104
1075: did not have a heart attack 104
1076:
1077:
1078: X_train, X_val, y_train, y_val = train_test_split(X, y, test_size=0.25, random_state=42)

```

Figure 7: Check that the dataset is balanced and create training and validation sets

The data was scaled using StandardScaler as seen in Figure 8. This step was important because it helps the machine learning model interpret features that have different ranges/magnitudes and interpret values on the same scale.

```

1081: scaler = preprocessing.StandardScaler()
1082: scaler.fit(X_train)
1083: X_train_scaled = scaler.transform(X_train)
1084: X_val_scaled = scaler.transform(X_val)

```

Figure 8: Scale the dataset

A logistic regression model was then created as seen in Figure 9. Logistic regression models are helpful in solving binary classification problems. The model can take into consideration multiple input criteria. An existing dataset is used to train a model to classify new data as either at high or low risk of heart disease. To do this, the model will estimate the probability of heart disease occurring (bounded between 0 and 1). For binary classification, a probability less than 0.5 will predict 0 (low heart disease risk) while a probability greater than 0.5 will predict 1 (high heart disease risk). Since there is a relatively small amount of data in the dataset, the model will be kept simple to avoid overfitting.

```

1085: model = 0
1086: model = Sequential()
1087: model.add(Dense(1, activation='sigmoid'))
1088: model.compile(loss='binary_crossentropy')
1089: model.fit(X_train_scaled, y_train, epochs=1000, verbose=0)
1090: J_list = model.history.history['loss']
1091: plt.plot(J_list)

```

Figure 9: Logistic regression model

RESULTS

When looking at the results of the logistic regression model, it is important to

compare the accuracy of the training and validation sets. We can check for overfitting using the validation and training errors seen in Figure 10 below. The model was able to avoid overfitting by confirming that the errors of the validation and training sets were not significantly different. We found that the validation error is approximately 0.379 and the training error is approximately 0.332. This shows a small but insignificant amount of overfitting.

```

1092: y_val_hat = model.predict(X_val_scaled)
1093: print('validation error =', loss(y_val_hat, y_val).numpy())
1094:
1095: y_train_hat = model.predict(X_train_scaled)
1096: print('training error =', loss(y_train_hat, y_train).numpy())
1097:
1098: 3/3 [=====] - 8s 28s/step
1099: validation error = 0.3792874
1100: 1/1 [=====] - 8s 28s/step
1101: training error = 0.3321883

```

Figure 10: Comparison of validation error and training error of logistic regression model

The logistic regression model created in this project to predict a person's risk of a heart attack is overall significantly accurate. We can look at the overall accuracy of the trained model using the classification report shown in Figure 11 below. This report provides important values such as accuracy, precision, and recall. The overall accuracy of this model was 0.85, indicating that it correctly classified 85% of the patients as either at a high or low risk of developing a heart attack. This accuracy score was relatively high given the size and dimensions of the dataset used in the project.

```

1102: y_hat_cat = 1*(model.predict(X_train_scaled) > 0.5)
1103: print(classification_report(y_train, y_hat_cat))
1104:
1105: 1/1 [=====] - 8s 28s/step
1106:
1107: precision    recall  f1-score   support
1108:
1109: 0       0.88       0.81       0.85       104
1110: 1       0.84       0.89       0.86       104
1111:
1112: accuracy       0.85       0.85       0.85       208
1113: macro avg     0.85       0.85       0.85       208
1114: weighted avg   0.85       0.85       0.85       208

```

Figure 11: Classification report of logistic regression model

Importantly, by creating a prediction model that has a high performance level, this model has the potential to save lives by identifying individuals at a high risk of heart disease at an early stage. Early identification of at-risk individuals allows for timely

interventions such as lifestyle modifications, medication management, and targeted medical interventions. Additionally, we can look at the features that are most impactful in the prediction of heart disease. Examples of these include age, blood pressure, cholesterol, and blood sugar. By intervening at an early stage of heart disease, medical professionals can reduce the number of heart attacks and their severe effects.

CONCLUSION

The machine learning model demonstrates the significant potential of machine learning algorithms to predict heart disease risk. By utilizing a dataset with relevant patient information, a model was created that could potentially save lives. Furthermore, each feature in the dataset correlates to heart disease risk, which was important to the machine learning model. Implementing a larger dataset to the logistic regression model created in this project could increase the overall accuracy of this model and help decrease overfitting. This study emphasizes the potential of early detection and intervention in preventing severe heart attacks using machine learning. This early detection may represent a useful tool to implement preventive measures in high-risk patients detected by this tool.

REFERENCE

Mayo Clinic Staff. (2022, May 25). Coronary Artery Disease. *Mayo Clinic*. <https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613#:~:text=CAD%20happens%20when%20coronary%20arteries,or%20even%20a%20heart%20attack.>

Tsao, C. (2023, May 15). Heart Disease Facts. *Centers for Disease Prevention and Control*. <https://www.cdc.gov/heartdisease/facts.htm#:~:text=In%20the%20United%20States%2C%20so>

meone,heart%20attack%20every%2040%20seconds.&text=Every%20year%2C%20about%20805%2C000%20people,States%20have%20a%20heart%20attack.&text=Of%20these%2C,are%20a%20first%20heart%20attack

Cedars Sinai Staff. (2022). Heart Attack (Myocardial Infarction) *Cedars Sinai*. <https://www.cedars-sinai.org/health-library/diseases-and-conditions/h/heart-attack-myocardial-infarction.html#:~:text=Half%20the%20deaths%20from%20a,an%20ambulance%20with%20trained%20personnel.>

Harvard Health Staff. (2017, March 25). Gender Matters: Heart Disease Risk in Women. *Harvard Health Publishing*. <https://www.health.harvard.edu/heart-health/gender-matters-heart-disease-risk-in-women>

Ryckowska, K. (2022, December 10). Menopause and Womens' Cardiovascular Health: Is It Really an Obvious Relationship? *National Library of Medicine*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10074318/>

American Heart Association Editorial Staff. (2022, December 2). What is a Heart Attack? *American Heart Association*. <https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks>

Park, C. (2013, February 12). Fasting Glucose Level and Risk of Incident Atherosclerotic Cardiovascular Diseases. *National Library of Medicine*. <https://pubmed.ncbi.nlm.nih.gov/23404299/>

Boyette, L. (2022, July 11). Physiology, Myocardial Oxygen Demand. *National Library of Medicine*. <https://www.ncbi.nlm.nih.gov/books/NBK49989/>

UCSF Health Staff. (2018, June 18). Nuclear Stress Test. *UCSF Health*.
<https://www.ucsfhealth.org/medical-tests/nuclear-stress-test>