CS545 Machine Learning
Final Project
August 20, 2023

Applying Machine Learning Models to Heart Disease Prediction

Contributors
Cristian Ion - MLP implementation + analysis, preprocessing, data methods
Leo Lu - Logistic Regression implementation + analysis, intro, discussion
Madeleine Fenner - SVM implementation + analysis, box plots, intro, data methods, discussion

Intro

Heart disease is a leading cause of death around the world, and prevention through early prediction and diagnosis is an important task. To aid in this effort, machine learning models can be applied to large datasets curated with relevant patient records. However the choice of model is not a straightforward process, and previous research articles have aimed to compare the heart disease prediction performance for a variety of models such as Support Vector Machines (SVMs), Multilayer Perceptrons (MLPs), Logistic Regression, Naive Bayes, and XGBoost (Bhatt et al., 2023; Nandal et al., 2022).

Similar to previous research, the aim of this project was to investigate the most effective heart disease prediction model by comparing the performance of three algorithms, MLP, SVM, and Logistic Regression, utilizing the Cardiovascular Disease Dataset sourced from Kaggle. The dataset contains a comprehensive collection of health attributes and cardiovascular disease outcomes, which enabled us to build and evaluate the predictive capacity of the selected models. Using the results, we compared measurements of accuracy, precision, and recall, to determine the most effective algorithm for heart disease prediction.

Methods

The Cardiovascular Disease Dataset contains records for 70000 patients, which each include the status of cardiovascular disease diagnosis and a set of 11 features relevant to cardiac health (Table 1). These features include a mix of objective features involving demographic information, examination features taken in a clinical setting, and subjective features that relate to lifestyle. The proportion of entries for patients with and without cardiovascular disease is roughly equal, with 49.97% of the dataset having the disease.

Table 1. Features and corresponding data types for the Cardiovascular Disease Dataset

feature	data type
age	int
height	int
weight	int
gender	categorical
systolic blood pressure	int
diastolic blood pressure	int
cholesterol	1: normal, 2: above normal, 3: well above normal
glucose	1: normal, 2: above normal, 3: well above normal
smoking	binary
alcohol intake	binary
physical activity	binary

When plotting the data for each feature as box plots to view the distributions, there are some clear outliers which may be related to data input errors (Figure 1). Specifically, value ranges for height (55 - 250 cm), weight (10 - 200 kg), systolic blood pressure (-150 - 16020 mmHg), and diastolic blood pressure (-70 - 11000 mmHg) were biologically impossible. Given that the age ranges indicated the data is only from adults, the height was thresholded to a more reasonable range of 140 cm to 220 cm, and all entries outside of this range were removed. Similarly, all entries with weights below 35 kg were removed. The highest ever recorded blood pressure in a human was 370/360 (Narloch et al., 1995). Because there were values way above this range in the dataset, the data was thresholded to not go above 370 mmHg for systolic blood pressure and 360 mmHg for diastolic blood pressure. Entries with systolic and diastolic blood pressures that were less than 1 mmHg or even negative were also removed, since these values are not biologically relevant. These changes removed 1,182 patients, making 68,818 the new total. Box plots visualizing the distributions after thresholding these features are shown in Figure 2.

Additionally, some steps were taken to normalize the data values across features. But, when comparing the results of model performance on the original data values and the normalized values, we found that the normalized data drastically reduced our accuracies. Therefore original values were kept.

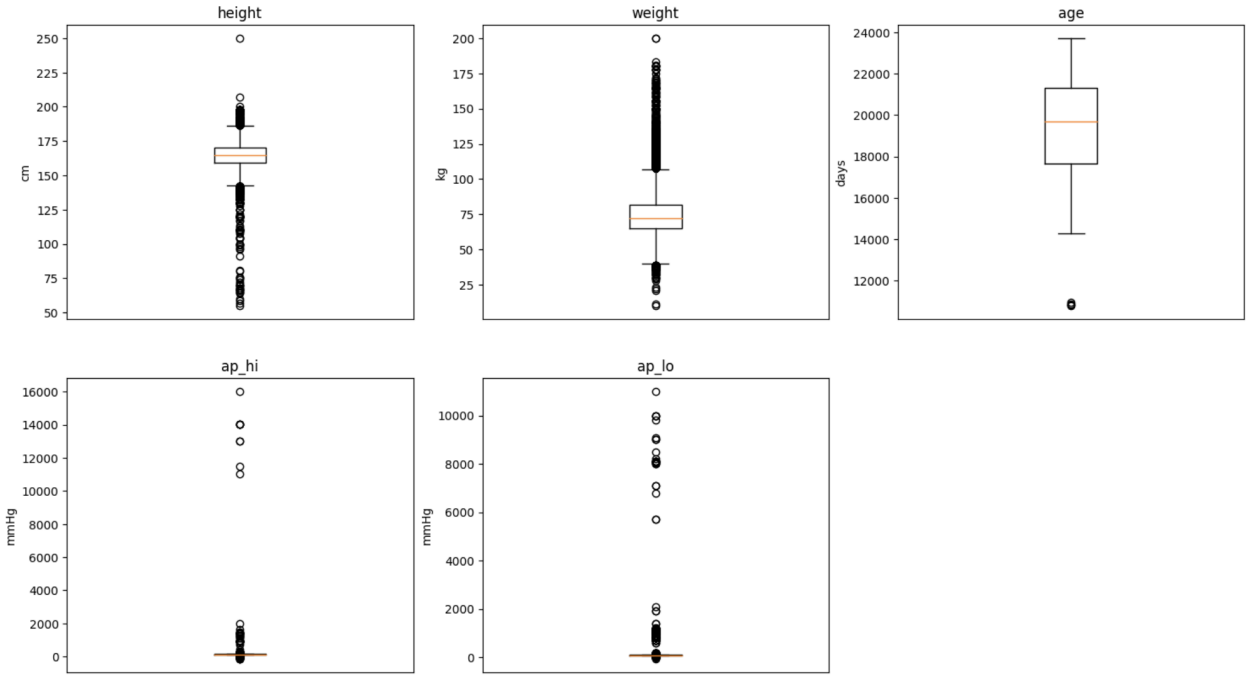


Figure 1. Box plots for all continuous value features of the raw dataset prior to preprocessing (height, weight, age, systolic blood pressure/ap_hi, diastolic blood pressure/ap_lo).

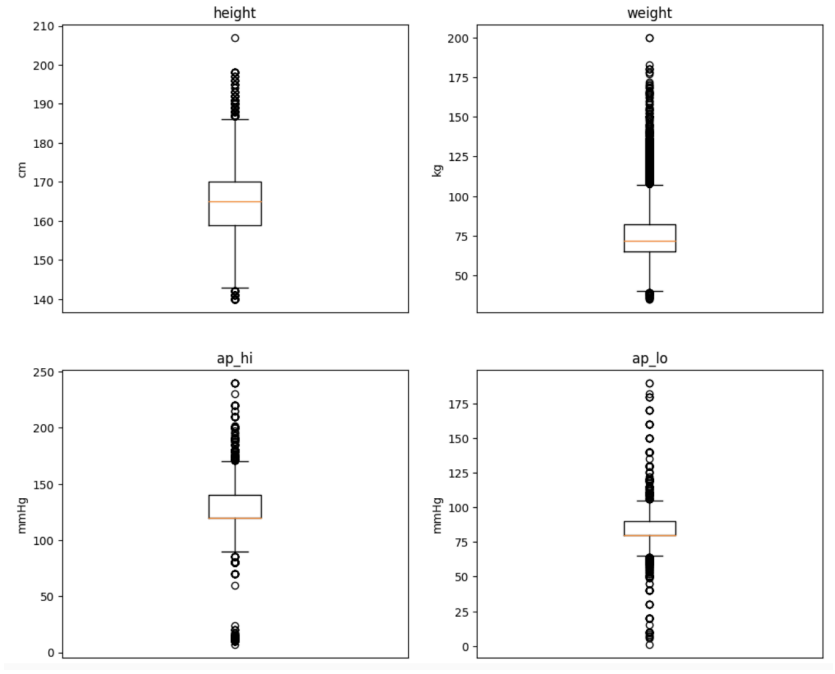


Figure 2. Box plots for features of the dataset after preprocessing by thresholding the ranges (height, weight, systolic blood pressure/ap_hi, diastolic blood pressure/ap_lo).

To perform classification on the dataset, MLP, SVM, and Logistic Regression models were applied. All models were implemented through the scikit-learn package, and default parameters were used unless specified otherwise.

MLP

Fully connected neural networks with multiple layers are MLPs. This means that every input node and the bias have an effect on each of the next layer's hidden nodes. If an MLP has multiple hidden layers, each hidden layer and its corresponding bias have an effect on the next layer's hidden nodes. The last hidden layer and its bias have an effect on each of the output layer nodes. This was implemented on the Cardiovascular Disease dataset using MLPClassifier from sklearn.neural_network.

SVM

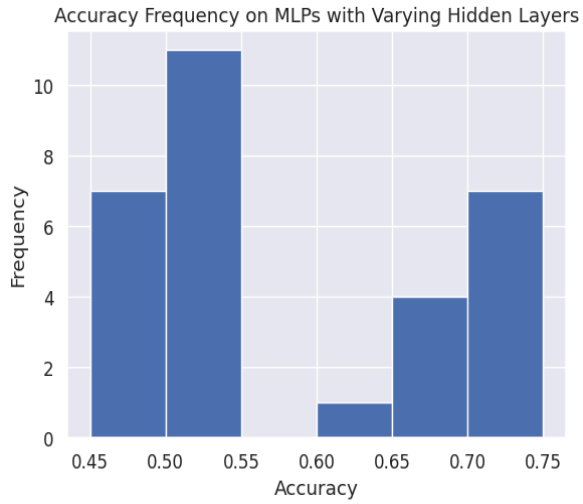
SVMs are supervised learning models that find the best separating plane to classify data by maximizing the margin with respect to the support vectors (the points in each class that lie closest to the decision boundary). Given non-linearly separable data, kernel functions are used to map the data to a higher-dimensional feature space where it can be better separated to achieve classification. An SVM was implemented on the cardiovascular dataset with the sklearn.svm.SVC class. To optimize the kernel parameter, accuracy was compared using a linear, polynomial, radial basis function, and sigmoid kernel with default parameters and the same random state. Then, using the best kernel, the highest performance over 5 random initializations was found.

Logistic Regression

Logistic Regression is a classification algorithm that employs a statistical technique used for binary classification tasks, where the goal is to predict one of two possible outcomes based on input features. It transforms the linear combination of input features using logistic (sigmoid) function that maps the output to a value between 0 and 1. This output represents the estimated probability of the instance belonging to the positive class. A threshold is then applied to classify instances into their respective classes. For this experiment, a Logistic Regression model was implemented using the sklearn.linear_model class. The various hyperparameters of C (Inverse of Regularization Strength), and solver (lbfgs, liblinear, 'newton-cg), were tested with the same random state to evaluate which combination produced the best accuracy. The best combination is then used to perform 5 rounds of modeling to obtain the highest performance.

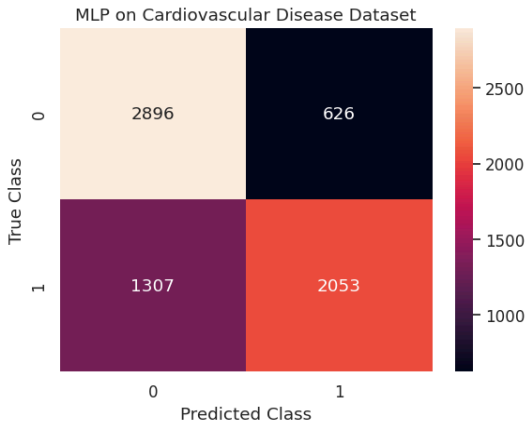
Results

MLP



Bin	Frequency
(0.4, 0.45]	0
(0.45, 0.5]	7
(0.5, 0.55]	11
(0.55, 0.6]	0
(0.6, 0.65]	4
(0.65, 0.7]	7
(0.7, 0.75]	0

Figure 3. Histogram of frequencies of MLP accuracies using different hidden layer sizes.



Accuracy	0.719
Precision	0.766
Recall	0.611

Figure 4. Confusion matrix of the MLP with two hidden layers each with 7 hidden units, and its accuracy, precision, and recall.

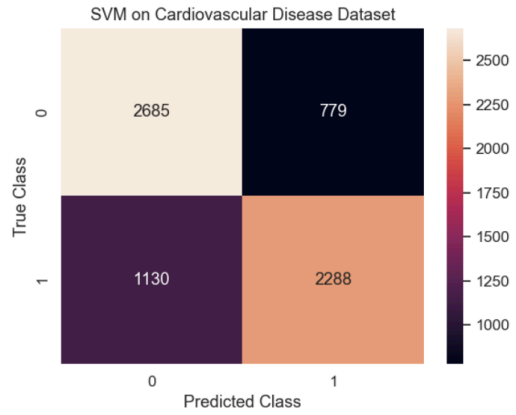
SVM

After applying SVM with a linear, polynomial, rbf, and sigmoid kernel, the linear kernel was found to have the best performance based on accuracy (Table 2).

Table 2. Classification accuracies for SVM with different kernels

kernel	accuracy
linear	0.728
polynomial	0.6136
rbf	0.604
sigmoid	0.4064

When running the SVM using a linear kernel over 5 random state initializations, the best accuracy was found to be 72.26% (Figure 3). The model also has higher precision than recall, meaning it has higher sensitivity and does not predict as many positives when the entry is negative, but it also misses a lot of positives.



Accuracy	0.7226
Precision	0.7460
Recall	0.6694

Figure 5. Confusion matrix for the best performance of SVM over 5 runs (0 = no disease, 1 = disease present). Accuracy, precision, and recall is also reported.

Logistic Regression

During the experiment with various combinations of C-values and solvers, it was found that the different solvers only had insignificant differences in accuracy (Table 3). C-value was the determining factor (Figure 6). Ultimately, C=100 and solver = newton-cg had the best performance based on accuracy.

C	Solver	Accuracy
100	newton-cg	0.720243
100	lbfgs	0.720233
100	liblinear	0.720238
10	newton-cg	0.720222
10	lbfgs	0.720222
10	liblinear	0.720233
1.0	newton-cg	0.720005
1.0	lbfgs	0.720005
1.0	liblinear	0.720005
0.1	newton-cg	0.717550
0.1	lbfgs	0.717561
0.1	liblinear	0.717556
0.01	newton-cg	0.699730
0.01	lbfgs	0.699730
0.01	liblinear	0.699704

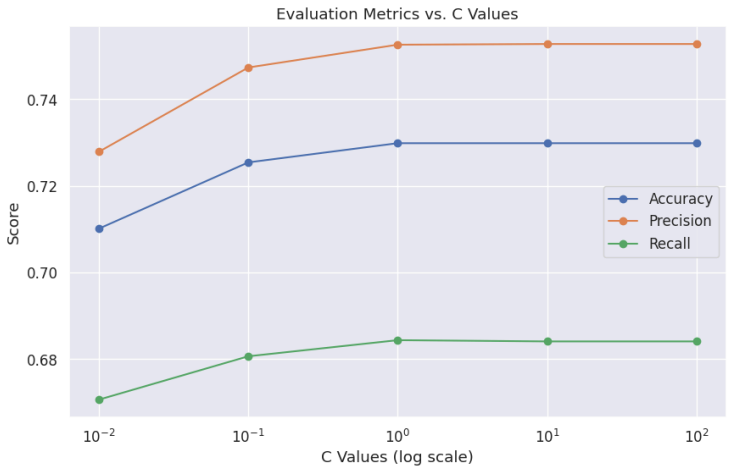
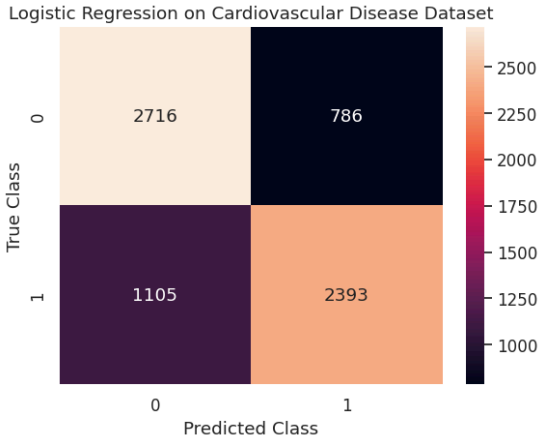


Figure 6. ↑ Plot of Accuracy, Precision, and Recall for C values ranging from 10^{-2} to 10^{2} by the power of 10.

←**Table 3.** Classification accuracies for various solver and C-value combinations



Accuracy	0.7299
Precision	0.7528
Recall	0.6841

Figure 7. Confusion matrix for the best performance of Logistic Regression over 5 runs (0 = no disease, 1 = disease present)

The conducted experiment involved running a Logistic Regression model 5 times, selecting the best-performing result based on accuracy. The best result achieved an accuracy of approximately 72.99%. Precision was higher at around 75.28%, indicating that the model is cautious and accurate in identifying true positive cases. However, the recall was slightly lower at approximately 68.41%, suggesting that while the model is proficient at minimizing false positives, it may miss some actual positive cases.

Discussion

Among the three models, Logistic Regression achieved the highest accuracy of 72.99%, followed closely by SVM with 72.26%, and MLP with 71.9%. Logistic Regression's slightly higher accuracy suggests that it made more correct predictions overall compared to the other models. MLP had the highest precision of 76.6%, followed by Logistic Regression with 75.28%, and SVM with 74.6%. Both MLP and Logistic Regression performed well in minimizing false positives, indicating their strong ability to accurately identify true positive cases. Logistic Regression achieved the highest recall of 68.41%, followed by SVM with 66.94%, and MLP with 61.1%. Logistic Regression's higher recall signifies its ability to identify a larger proportion of actual positive cases, making it suitable for scenarios where capturing all cases is crucial.

Logistic Regression and SVM achieved more of a balance between precision and recall, while MLP prioritized precision, achieving the highest precision but at the cost of lower recall. For a medical diagnostic tool used as an initial screening process, recall would likely be the most important attribute. Follow up tests can be used to identify any patients who in reality do not have the disease, however missing a positive patient may have more serious health consequences if their disease goes untreated. With this in mind, Logistic Regression is the most promising model for heart disease prediction because it achieves the highest recall, highest accuracy, and still relatively high precision.

References

Bhatt CM, Patel P, Ghetia T, Mazzeo PL. Effective Heart Disease Prediction Using Machine Learning Techniques. Algorithms. 2023; 16(2):88. https://doi.org/10.3390/a16020088

Nandal N, Goel L and TANWAR R. Machine learning-based heart attack prediction: A symptomatic heart attack prediction method and exploratory analysis [version 1; peer review: awaiting peer review]. F1000Research 2022, 11:1126. https://doi.org/10.12688/f1000research.123776.1

Narloch, J. A., & Brandstater, M. E. (1995). Influence of breathing technique on arterial blood pressure during heavy weight lifting. Archives of Physical Medicine and Rehabilitation, 76(5), 457–462. https://doi.org/10.1016/s0003-9993(95)80578-8

Navlani, Avinash. “Multi-Layer Perceptron Neural Network using Python” *Machine Learning Geek,* 23 Apr. 2021,https://machinelearninggeek.com/multi-layer-perceptron-neural-network-using-python/