SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Using Machine Learning to Predict Risk of Cardiovascular Disease Within 10 Years

By Austin Fang

AUTHOR BIOGRAPHY

Austin Fang is a high school student at Frisco High School. He has a huge passion for computer science, mathematics, and physics, and plans to study one of these fields in the future. His interest is expressed through competitions, such as math contests, and projects, including conducting Machine Learning research. He is interested in cardiovascular disease because of his experiences and interactions with struggling relatives. Outside of academics, he enjoys basketball and music.

ABSTRACT

Cardiovascular disease (CVD) has emerged as one of the most primary causes of death in South Asia. Therefore, it is extremely important to understand the underlying important factors to prevent disease, as well as acknowledging our current risk in order to potentially treat the disease. Machine Learning is a branch of Artificial Intelligence (AI) which uses computers to analyze and detect patterns in a dataset, using mathematics to create conclusions. This research paper focuses on using Machine Learning to detect CVD risk score, a percentage probability of developing heart-related diseases within 10 years of a person's life. The dataset used was ethically collected in Bangladesh, a country heavily impacted by the negative effects of CVD. Four machine learning models are used - linear regression, decision tree, random forest, and XGBoost regressor. The metric for comparison is MAE, or mean absolute error, and is a good indicator because it is easy to digest, telling us exactly how much we are "off" by. Linear regression and random forest performed the best, each with around an MAE of 0.88. The study also concluded that the features total cholesterol, diabetes, and BMI were the three key contributors to development of CVD.

Keywords: *machine learning, data science, data cleaning, cardiovascular disease, Bangladesh, regression, decision trees, linear regression.*

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

INTRODUCTION

Cardiovascular disease (CVD) is a group of diseases specifically located, or related, to your heart and blood vessels. Examples include coronary heart disease, stroke, heart attack, and peripheral artery disease. In recent years, CVD deaths have shown an intense upward trend, propelling it to the #1 leading cause of death across the globe (World Health Organization, 2021). In 2022, an estimated 19.8 million people died from CVD, contributing to 32% of total deaths across the world. While development of CVD may seem random, there are certain contributing behavioral and environmental risk factors that may increase your chances. These factors, which will later be used to determine an accurate prediction of CVD risk, include high blood pressure, high cholesterol, smoking, diabetes, lack of physical activity, and BMI (World Health Organization, 2021).

Although contributing to around 25% of the world's population (WorldOMeter, 2019), CVD is especially prominent in South Asia. South Asians have higher rates of mortality from heart disease, relating to the increase of premature and intense coronary artery disease (Patel, 2021). Studies from Patel et. al 2021 have shown that genetic factors, including body composition, risk of diabetes, and environmental changes cause them to develop physiological traits linked with heart disease. Additionally, the World Health Organization (2021) mentions that 80% of CVD deaths take place in middle and lower income countries due to the lack of healthcare support and economic resources to prevent CVD. Oftentimes in these circumstances, those who have contracted CVD may not have the resources to properly detect their risk, leaving panic for the last minute. This most affects those who live in poverty, as late detection results in advanced surgeries, creating high out of pocket expenditure (Schultz et al., 2018). Therefore, in this research, we selected a dataset gathered in the South Asian country of Bangladesh. Bangladesh ranked 124 out of 167 in 2023 (Prosperity Institute et al. 2023) on the Legatum Prosperity Index, an annual ranking of countries based on their "prosperity", an overall combination of factors such as social well-being, wealth, education, and health. Since Bangladesh ranks on the lower side of the scale, it is even more important to bring early, accurate prediction and CVD education to the country.

Machine learning emerges as a novel tool to predict CVD (Mohd Faizal et al., 2021). Existing CVD machine learning models are mostly classification models, which predict whether or not to develop CVD (Peng et al. 2023). However, our study is trying to further deep dive the CVD-risky population, and predict their quantitative risk levels using a limited number of biometric inputs. Therefore, we selected the Bangladeshi dataset (Kaggle, 2025) composed of patients with medium/high CVD risk, and used the machine learning regression models to quantitatively predict CVD risk scores. We used multiple features and emphasized on feature and model selection. Another consideration is the population-specific nature of CVD research. Current machine-learning-based CVD predictions mainly focus on populations in developed countries such as the United States (Mezzatesta et al., 2019), Italy (Mezzatesta et al., 2019), United Kingdom (Weng et al., 2017) or Australia (P. Unnikrishnan et al., 2016). Although some Asian populations in China (Peng et al. 2023) or Korea

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

(Roh et al., 2019) have been studied, we aim to help the population in poverty such as those in South Asia as mentioned above. This is the additional reason we used the data from Bangladesh (Kaggle, 2025).

DATASET

The dataset being used (Kaggle, 2025) is a collection of Bangladeshi residents in the Jamalpur Medical College Hospital. The data was collected ethically, regarding patient consent and confidentiality. The goal of the dataset is to build a machine learning model to predict CVD risk score based on previous patient statistics. The data consists of 1529 patients collected in 2024 at the Jamalpur Medical College Hospital in Jamalpur, Bangladesh, containing 21 features (including age, sex, BMI, blood pressure, diabetes/smoking status, ect.) and 1 target variable (CVD risk score). Below is a description for each of the features, according to the original dataset authors on Kaggle.

1. Sex: Male or Female.
2. Age: Patient's age (in years).
3. Weight (kg): Patient's weight in kilograms.
4. Height (m): Patient's height in meters.
5. BMI: Body Mass Index, calculated from weight and height.
6. Abdominal Circumference (cm): Measurement of abdominal girth.
7. BP: Blood pressure readings.
8. Total Cholesterol: Total cholesterol levels in the blood.
9. HDL: High-density lipoprotein levels, also known as the "good" cholesterol.
10. Fasting Blood Sugar: Blood glucose levels after fasting.
11. Smoking Status: Indicates whether the patient is a smoker.
12. Diabetes Status: Indicates whether the patient has diabetes.
13. Physical Activity Level: Level of physical activity.
14. Family History of CVD: Indicates if there is a family history of cardiovascular diseases.
15. CVD Risk Level: Classification of the patient's CVD risk level.
16. Height (cm): Patient's height in centimeters.
17. Waist-to-Height Ratio: Ratio of waist circumference to height.
18. Systolic BP: Systolic blood pressure, this is the number which is more significant.
19. Diastolic BP: Diastolic blood pressure.
20. Blood Pressure Category: Classification of blood pressure levels.
21. Estimated LDL: Estimated low-density lipoprotein levels, also known as the "bad" cholesterol.

In our case, the target variable being used (CVD Risk Score) is Framingham Risk Score (FRS), an estimated percentage chance of someone developing CVD within a ten-year period. FRS is useful when

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

determining proper medication to be prescribed, or convincing patients to adopt healthier lifestyle changes. Normally, a risk score of <10% is healthy, 10%-20% is moderate risk, and >20% is high risk. FRS also tends to overestimate certain regions of the world, including Japanese, Hispanic, and Indo-Carribean (Shaw, 2012). It is important to note that most people visiting this hospital already have severe risks of CVD, due to the high poverty rate and dirty conditions of the city. The following graph shows the distributions of the CVD risk score (Fig. 1). For comparison, a safe CVD risk score for an average person is 0-5%.

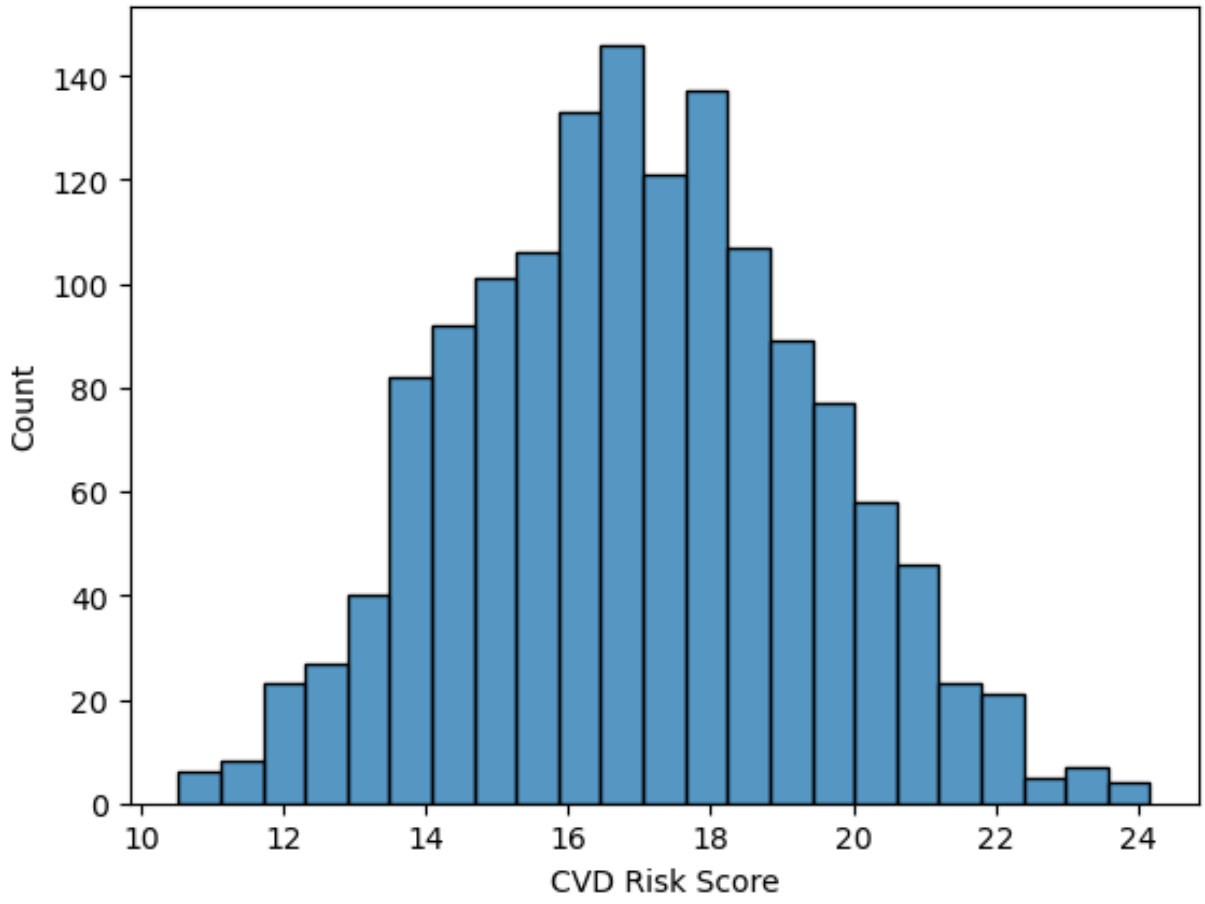

Figure 1: *Histogram describing the distribution of CVD risk scores.*
The majority of the distribution is in the unhealthy range (Dumlao, 2025).

Although it is important to identify general CVD-risky individuals from average persons, it is also imperative to further deep dive those with medium to severe CVD risks, which helps develop CVD prevention

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

and/or diagnosis specific to different risk levels. This dataset is suitable for this purpose because most samples referred to patients with scores between 14-20%, putting them between intermediate and high risk (Fig. 1).

DATA CLEANING

We dropped data with missing values with some exceptions. The first exception is data with missing BMI but available height and weight, since BMI can be directly calculated through height and weight. The second exception is two duplicate features relating to height - one in meters, the other in centimeters. We only removed data missing values in both features. If a data value is missing in one feature, we can use height in meters to find that in centimeters, or vice versa. The exceptions saved 70 data samples, so that after missing values are eliminated, the number of samples becomes 834 from 1529.

	Sex	**Blood Pressure (mmHg)**	**Smoking Status**	**Diabetes Status**	**Physical Activity Level**	**Family History of CVD**	**CVD Risk Level**	**Blood Pressure Category .,**
0	F	125/79	N	Y	Low	N	INTER MEDIA RY	Elevated
1	F	139/70	Y	Y	High	Y	HIGH	Hypertension Stage 1
3	M	140/83	N	N	High	Y	INTER MEDIA RY	Hypertension Stage 1
4	F	142/90	Y	Y	High	N	INTER MEDIA RY	Hypertension Stage 1

Table 1: *Display of String type features with first 4 samples as examples*

We also encoded the features with string types listed in Table 1. First, we can drop the Blood Pressure (mmHg) feature, since systolic and diastolic blood pressure are already features. Additionally, we can also drop the "CVD Risk Level", since the definition of this feature isn't clearly stated. Furthermore, in the binary features 'Sex", "Smoking Status", and "Diabetes Status", we simply encoded them to "1" for Yes and "0" for No. Finally,

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

we encoded the rest of the features by assigning each value an integer from zero according to their extremity, because each of these features contains only a few possible values.

FEATURE SELECTION

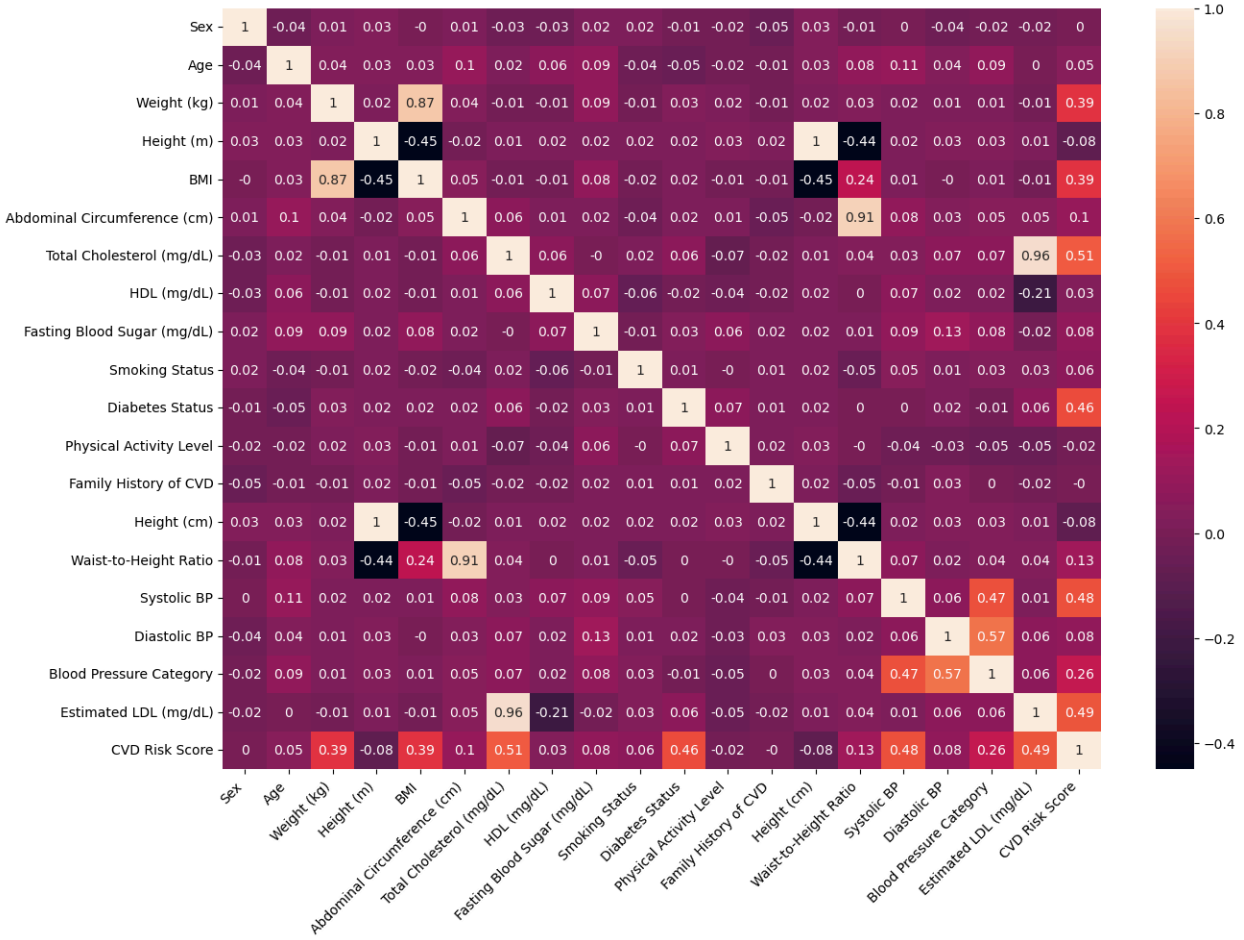

Figure 2: *Correlation matrix of dataset*
The lighter colors representing higher a correlation coefficient

Fig. 2 shows the correlations between features and that between each feature and the target variable. Focusing on the target "CVD Risk Score", we notice high correlations between target and several features, specifically Estimated LDL, Blood Pressure Category, Systolic BP, Diabetes Status, Total Cholesterol, BMI, and

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Weight. We used scientific knowledge of the biology of the human body to analyze why these features are important.

- LDL: Estimated LDL, also known as bad cholesterol, is an extremely important contributing factor to the risk of developing cardiovascular disease. This is because the rise in LDL levels directly correlates to the amount of plaque being built up in your arteries, leading to issues such as coronary artery disease and stroke (Cleveland Clinic, 2022). Similarly, this also applies to total cholesterol. In the dataset, the average ratio of LDL to HDL (the "good" type of cholesterol) is 1.5 - 2.5, meaning that the samples' cholesterol levels mostly contain LDL, in which case the higher the total cholesterol, the higher the LDL.
- Systolic Blood Pressure: Multiple studies have shown that when testing for heart disease, systolic blood pressure is much more useful than diastolic blood pressure. Systolic blood pressure is a measure of how hard the heart pumps blood, eventually leading to damaged blood vessels and cardiac disease. This dataset, therefore, matches with the recent research that Systolic blood pressure is an important indicator of heart disease (Christensen, 2021).
- BMI/Weight: High BMI and weight are essentially methods of obesity. According to Wiley et al. (2021), "Obesity contributes directly to incident cardiovascular risk factors, including dyslipidemia, type 2 diabetes, hypertension, and sleep disorders." Research has been heavily conducted observing the relationship between the two, and consistently finds that they are closely related (Wiley, 2021). Additionally, the American Heart Association (2021) states that obesity is also linked to development of type 2 diabetes, another possible explanation of the high correlation of the "Diabetes Status" feature in our dataset.
- Diabetes: Development of diabetes occurs when the body becomes unable to produce a stable amount of insulin, a hormone controlling blood sugar levels. Therefore, high blood sugar levels may damage the condition of blood vessels and nerves. Diabetes is also closely related with BMI/obesity (NIDDK, 2021).

Therefore, we only kept these features including Estimated LDL, Blood Pressure Category, Systolic BP, Diabetes Status, Total Cholesterol, BMI, and Weight for further machine learning model development.

MACHINE LEARNING MODELS

We used four different machine learning regression models (see below for details) to predict CVD risk scores. To properly test the accuracy of the models, we randomly split the whole dataset into training (⅔ of dataset) and testing (⅓ of dataset) sets. The training set was used to build models and the testing set is used to evaluate the model performances in terms of MAE (mean absolute error), which tells us on average how much we are off. The computation was done in Python Environment with the scikit-learn module.

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Model 1: Linear Regression

Linear Regression (Scikit-Learn, 2019) is a basic machine learning regression model that mathematically visualizes a straight line through all data points. New predictions are created by taking all input variables, choosing the point on the line containing these variables, and looking at the output variable. While linear regression may seem very simple, it is extremely powerful when used under the correct circumstances, which will be especially demonstrated later. A standard linear regression line is modeled as following:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n$$

Where the goal is to choose the best values for θ_i such that the values most accurately match the dataset. There are many types of linear regression, each one with a different way of choosing the best coefficients, but we used the standard Linear Regression model imported from module scikit-learn, which uses the Ordinary Least Squares Method. The Ordinary Least Squares method attempts to minimize loss, in which the lost function in this case is Mean Squared Error (MSE). The formula for MSE is just the sum of the squares of all differences between predicted and actual value, or, mathematically speaking, $MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

While linear regression may work well in some cases, it has some cons. Specifically, it is extremely reliant on the fact that our data can be viewed as a straight line. In most practical applications, data will most likely be scattered, and may even contain several outliers. Thus, linear regression in these cases will be unable to properly predict new cases. In the end, our ending MAE for Linear Regression is 0.88, which is quite good.

Model 2: Decision Tree

A decision tree is another simple yet functional model . It takes the data and creates a binary tree which splits the dataset into multiple regions. Each region contains a value, and the test data gets sorted into a region. The value of the region of a datapoint is assigned to the target value of the datapoint (Scikit-Learn, 2019). To begin with some terminology, the root of the decision tree is the first decision-making node that a datapoint will encounter, which divides the dataset into two parts. Decision nodes are nodes that are in the middle, further splitting the dataset. Finally, there are leaf nodes, which are the final results of the decision tree. Leaf nodes determine the result of a data point after multiple splits. However, the most important part of a decision tree is to decide the order of the features. To do this, the model attempts to minimize variance. The formula for variance is:

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

$$\sigma^2 = \frac{1}{n} \Sigma (y_i - \overline{y})$$

Where n is the number of values, y_i denotes a specific value, and $\overline{y}$ denotes the mean of all values. Essentially, variance is the average of the MAE of the differences between expected and actual values. To determine which feature has the least variance, the following formula is used:

$$\Delta\sigma^2 = \sigma^2_{parent} - (\frac{n_{left}}{n_{total}}\sigma^2_{left} + \frac{n_{right}}{n_{total}}\sigma^2_{right})$$

Where $\Delta\sigma^2$ represents the variance function, n_{total} represents the total number of nodes, n_{left} and n_{left} represent the number of nodes on the left and right. The decision tree calculates this value for all features, then chooses the feature which minimizes it. Note that variance is used in this case because it is a regression task. For most classification tasks, entropy should be used as a measure, and try to maximize information gain.

However, like linear regression, decision trees have flaws. First of all, making a decision tree too deep would result in overfitting. To combat this, we used the max_depth variable when creating the decision tree. After testing multiple values of max_depth, we determined that the best value was 8, which substantially decreased the MAE from around 1.09 (max_depth at default None) to around 0.99. Additionally, we updated the min_samples_leaf parameter to make every leaf node have at least a certain amount of samples. This also helps avoid overfitting. After trying different values, the min_samples_leaf of 4 works the best, slightly reducing our MAE to 0.98.

Model 3: Random Forest Regressor

The Random Forest model generalizes and minimizes overfitting, by taking multiple subsets of the entire dataset, creating decision trees for each of these subsets, and finally using the mean of all trees as our final result (Scikit-Learn, 2019). Originally, with the default settings in Random Forest, our model has a MAE of around 0.92. However, as we changed the minimum number of samples per leaf to 5 instead of the default value 1, the MAE decreased to around 0.89. Other attempts in adjusting parameters did not significantly decrease the MAE.

Model 4: XGBoost Regressor

XGBoost Regressor (also known as Xtreme Gradient Boosting regressor) is a machine learning algorithm that also builds off the Decision Tree by sequentially building small decision trees, with each tree

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

attempting to fix the errors of the previous tree. During each step, the algorithm uses the gradient of the loss function, telling the machine what to improve. Each split attempts to maximize information gain, and each small tree's information is used overall to create a big tree. XGBoost also has additional features, including handling missing values (which in our case isn't useful), regularization, a shrinkage, and subsampling. In our case, XGBoost performed with a MAE of around 0.93.

RESULTS

The following bar graph describes our results:

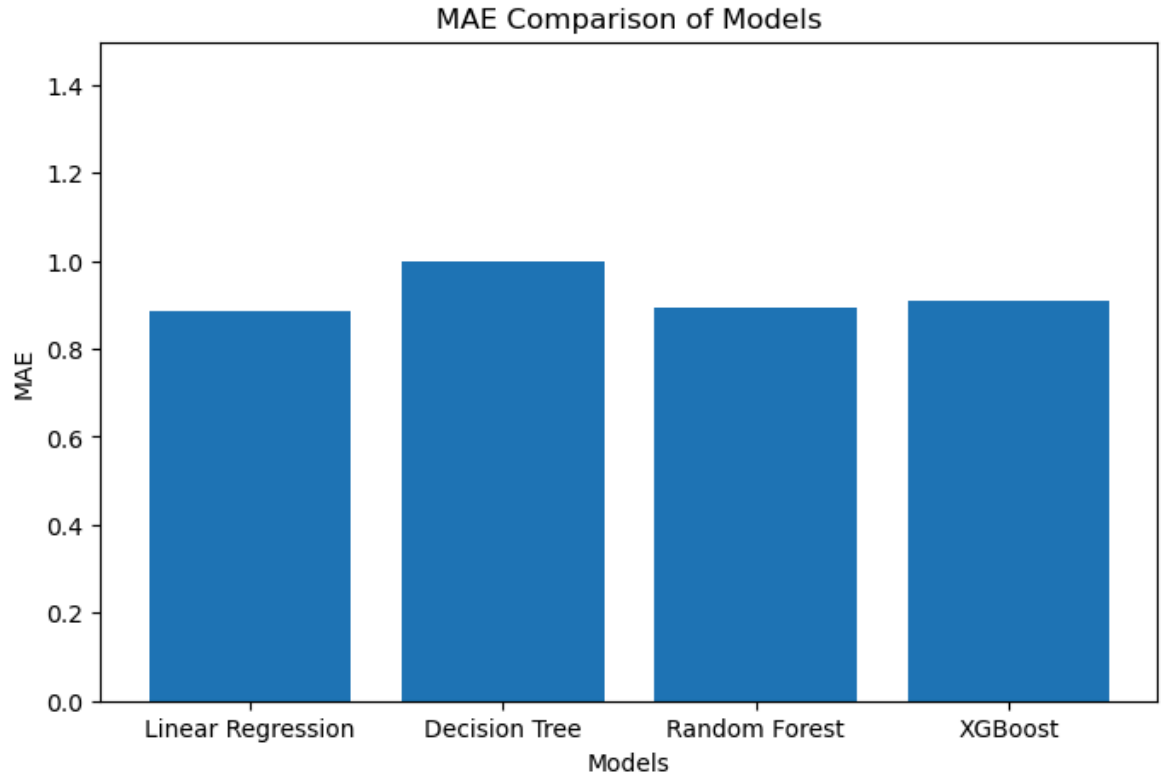

***Figure 3**: Results bar graph*
Random Forest and Linear Regression each around 0.88, with Linear Regression performing slightly better

One observation is that all models perform decently well. The target variable (CVD Risk Score) ranges from 10 to 24 (in Fig. 1), while MAEs are all below 1.0. We can conclude why each model performs in which way. Firstly, our linear regression model performs the best - meaning that there is some degree of linearity

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

between features and the target variable. This could also explain why the Decision Tree model has the highest MAE - decision trees are used in datasets with higher complexity and less linearity. Random forest and XGboost are ensembles of decision trees and improve performance from a single decision tree. Their performances are nearly as good as linear regression and can provide feature importance.

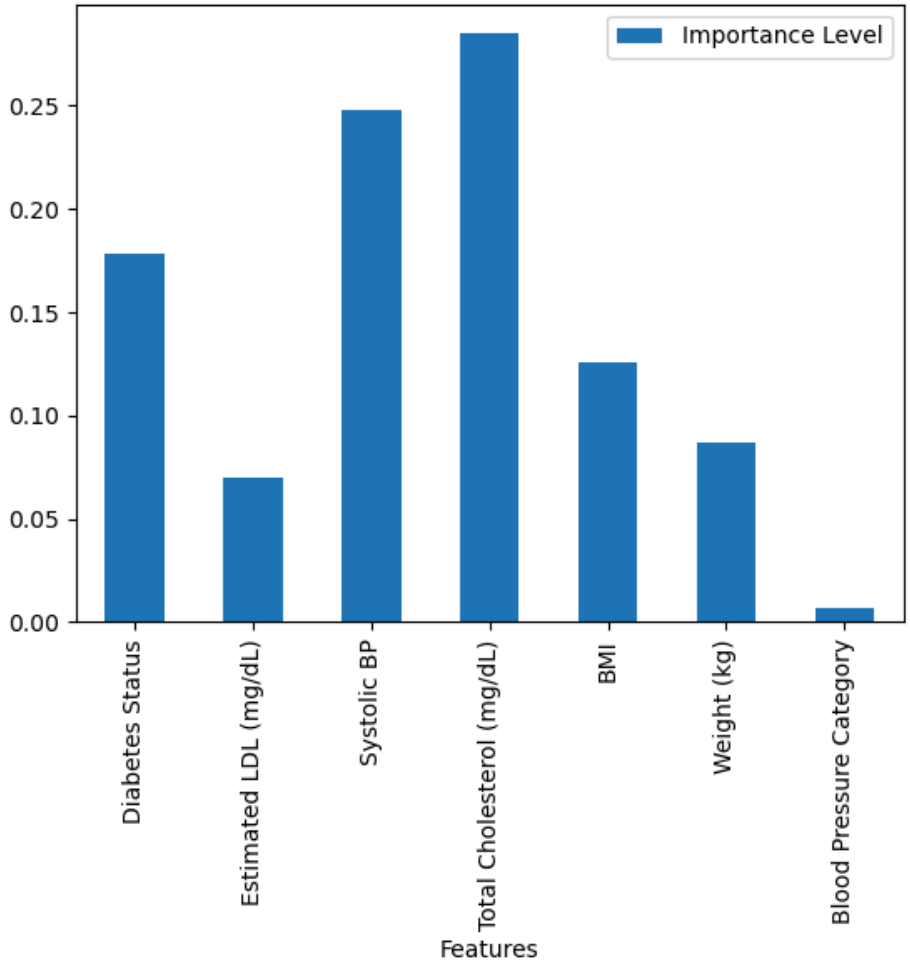

***Figure 4**: Feature importance results bar graph (according to Random Forest)*
Total Cholesterol is clearly the best, followed by Systolic BP, Diabetes Status, and BMI

These features also belong in 3 different levels of risk.

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

The first level is the most risky, shown to have a direct correlation to CVD risk, even ignoring other features, is cholesterol

- Total cholesterol is most influential, maybe because it considers the impact of both good and bad cholesterols. The majority of total cholesterol is bad cholesterol, so this feature includes the impact of LDL. Although HDL is good cholesterol, too much HDL is still harmful.
- LDL is important. However, when total cholesterol as the no.1 feature, LDL's impact is already considered so its importance is suppressed
- Second-level risky: physiological disease
    - Systolic BP and Diabetes
- Third-level risky: no disease but unhealthy signals
    - BMI

DISCUSSION

To summarize our findings, we found that Linear Regression and Random Forest are the best-performing models, each being only off by 0.88% (fig. 4). Additionally, the features that are the most important (in order from most to least) are Total Cholesterol, followed by Systolic BP, Diabetes Status, and BMI. As we see, while Blood Pressure Category had high correlation with CVD Risk Score, it has surprisingly low variance when trained because its impact is already included in Systolic BP. Additionally, Total Cholesterol is a much better indicator than just LDL (bad cholesterol), probably because the majority of total cholesterol is composed of LDL. Therefore, when the model uses total cholesterol as a feature, it already considers the impact of LDL.

We compared our results with those from other research that uses machine learning to predict CVD. Peng et al.(2023) predicts whether a person has CVD (different from calculating CVD risk score) by using a XGBH model. It was found that Systolic Blood Pressure, cholesterol levels and BMI are the top factors, which is consistent with our discovery, although Peng et al. ranked Systolic Blood Pressure as the most important factor, while we find it is the second-most important. However, they also found age and diastolic blood pressure important, while we find that it has little correlation with CVD Risk Score. Note that the difference in important features might be due to that in the scope of models. The model by Peng et al. is to differentiate CVD-risky patients from the healthy population, while our model is to estimate quantitative risk levels among the CVD-risky population.

The scope of this research should be limited to patients who have relatively high CVD risk scores. All of our data was collected in a Medical College Hospital, and thus from those who are already suffering the consequences of CVD. Figure 1 shows that the CVD risk scores in our data range from 10-24% and peaks near

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

17-18%, while the score for the healthy population is 0-5%. Therefore, our model in the future could be used together with another model skilled at differentiating healthy and unhealthy patients. The later model first identifies an unhealthy population which typically has high CVD risk scores. Our model could then further deep dive the unhealthy population to predict different CVD risks, helping physicians develop different prevention, diagnosis and treatment strategies based on their risk levels.

Additionally, this research could be further improved as more data become available. Current data was gathered within only a year (Kaggle 2025). As time goes on, if more data is collected to include more features and increase the sample size from currently around 800 to thousands or more, we expect our models could be improved with better accuracy and the scope could be even expanded.

REFERENCES

Christensen, T. (2021, March 1). *Which blood pressure number matters most? The answer might depend on your age*. Www.heart.org. https://www.heart.org/en/news/2021/03/01/which-blood-pressure-number-matters-most-the-answer-might-depend-on-your-age

Cleveland Clinic. (2022, October 27). *LDL cholesterol: What it is & how to lower it*. Cleveland Clinic. https://my.clevelandclinic.org/health/articles/24391-ldl-cholesterol

Dumlao, J. (2025). *CAIR-CVD-2025: Cardiovascular Risk from Bangladesh*. Kaggle.com. https://www.kaggle.com/datasets/jocelyndumlao/cair-cvd-2025-cardiovascular-risk-from-bangladesh

Mezzatesta, S., Torino, C., Meo, P. D., Fiumara, G., & Vilasi, A. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. *Computer Methods and Programs in Biomedicine*, *177*, 9–15. https://doi.org/10.1016/j.cmpb.2019.05.005

Mohd Faizal, A. S., Thevarajah, T. M., Khor, S. M., & Chang, S.-W. (2021). A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. *Computer Methods and Programs in Biomedicine*, *207*, 106190. https://doi.org/10.1016/j.cmpb.2021.106190

NIDDK. (2021, April). *Diabetes, Heart Disease, and Stroke | NIDDK*. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

P. Unnikrishnan, Kumar, D., Arjunan, S. P., Kumar, H., Mitchell, P., & Kawasaki, R. (2016). Development of Health Parameter Model for Risk Prediction of CVD Using SVM. *Computational and Mathematical Methods in Medicine*, *2016*, 1–7. https://doi.org/10.1155/2016/3016245

Pal, M., Parija, S., Panda, G., Dhama, K., & Mohapatra, R. K. (2022). Risk prediction of cardiovascular disease using machine learning classifiers. *Open Medicine*, *17*(1), 1100–1113. https://doi.org/10.1515/med-2022-0508

Patel, J. (2021). South Asian cardiovascular disease: Dispelling stereotypes and disparity. *American Journal of Preventive Cardiology*, *7*, 100189. https://doi.org/10.1016/j.ajpc.2021.100189

Peng, M., Hou, F., Cheng, Z., Shen, T., Liu, K., Zhao, C., & Zheng, W. (2023). Prediction of cardiovascular disease risk based on major contributing features. *Scientific Reports*, *13*(1), 4778. https://doi.org/10.1038/s41598-023-31870-8

Prosperity Institute. (2023, February 13). *Rankings :: Legatum Prosperity Index 2023*. Legatum Prosperity Index 2023. https://index.prosperity.com/rankings

Roh, E., Chung, H. S., Lee, J. S., Kim, J. A., Lee, Y.-B., Hong, S., Kim, N. H., Yoo, H. J., Seo, J. A., Kim, S. G., Kim, N. H., Baik, S. H., & Choi, K. M. (2019). Total cholesterol variability and risk of atrial fibrillation: A nationwide population-based cohort study. *PLOS ONE*, *14*(4), e0215687. https://doi.org/10.1371/journal.pone.0215687

Schultz, W. M., Kelli, H. M., Lisko, J. C., Varghese, T., Shen, J., Sandesara, P., Quyyumi, A. A., Taylor, H. A., Gulati, M., Harold, J. G., Mieres, J. H., Ferdinand, K. C., Mensah, G. A., & Sperling, L. S. (2018). Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. *Circulation*, *137*(20), 2166–2178. https://doi.org/10.1161/circulationaha.117.029652

Scikit-Learn. (2019). *User guide: contents — scikit-learn 0.22.1 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/user_guide.html

Shaw, L. J. (2012). An Approach to Asymptomatic and Atypically or Typically Symptomatic Women with Cardiac Disease. *Interventional Cardiology Clinics*, *1*(2), 157–163. https://doi.org/10.1016/j.iccl.2012.02.001

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

SCHOLARLY REVIEW

SCHOLARLY REVIEW JOURNAL

E-ISSN: 2996-8380
Published by Leadership & Innovation Lab

WINTER 2025/2026

Wiley, T. M. P., Poirier, P., Burke, L. E., Després, J.-P., Larsen, P. G., Lavie, C. J., Lear, S. A., Ndumele, C. E., Neeland, I. J., Sanders, P., & St-Onge, M.-P. (2021). Obesity and Cardiovascular disease: a Scientific Statement from the American Heart Association. *Circulation*, *143*(21), 984–1010. https://doi.org/10.1161/cir.0000000000000973

World Health Organization. (2021, June 11). *Cardiovascular diseases (CVDs)*. World Health Organisation . https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

WorldOMeter. (2019). *Population of Southern Asia (2019) - Worldometers*. Worldometers.info. https://www.worldometers.info/world-population/southern-asia-population/