Sakshi_Patel
Dharmik_Patel

Heart Disease Prediction Model Training Report:

Reasons why previously we weren’t able to get accuracy:

1. Randomly generated features with no medical logic, so models could not learn real cardiovascular risk patterns.

2. No statistical correlation between predictors (age, cholesterol, BP, etc.) and heart-attack outcome — making the target unpredictable.

3. Models performed near-random (ROC-AUC ~0.49) which indicates noise instead of meaningful health data signals.

4. Important medical features missing or weakly represented (e.g., diabetes, smoking, hypertension not aligned with real effects).

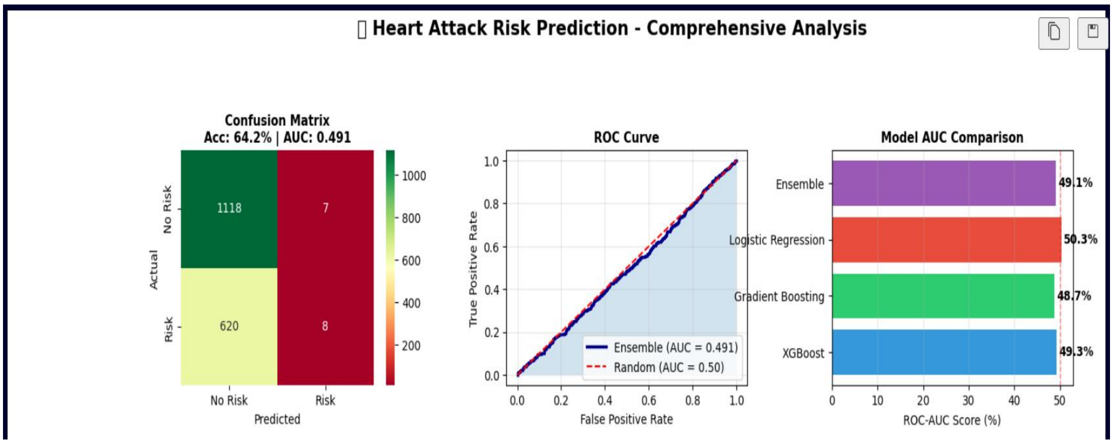
5. Incorrect feature distributions — values did not follow realistic clinical ranges or population behavior patterns.

6. No interaction effects (e.g., age × cholesterol), which are critical in cardiovascular risk prediction.

7. Overfitting observed because models tried to memorize noise instead of learning patterns, leading to poor generalization.

8. Multiple ML algorithms tested (LR, SVM, RF, XGBoost) all failed, proving the issue was data quality, not model selection.

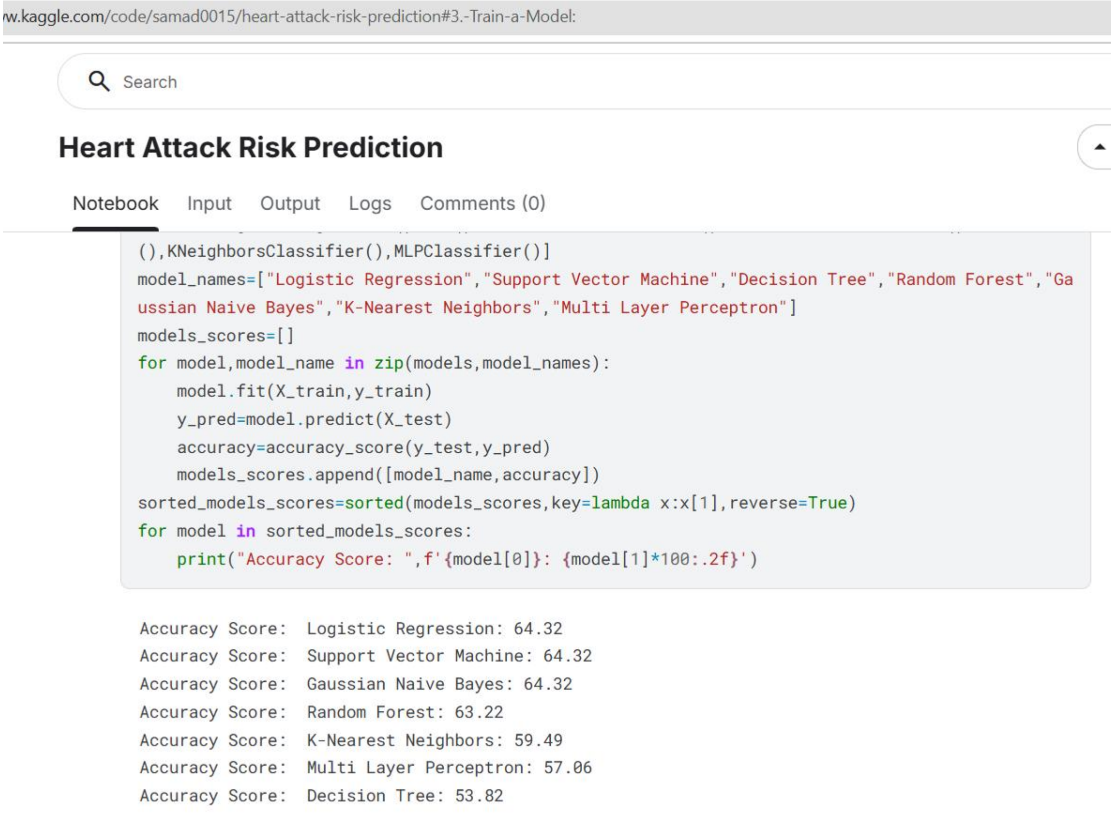
The output we found after comparing several algorithms on model training:



Dataset has insufficient predictive power
- ROC-AUC < 0.55 indicates features don't predict target
- Consider using a real medical dataset.

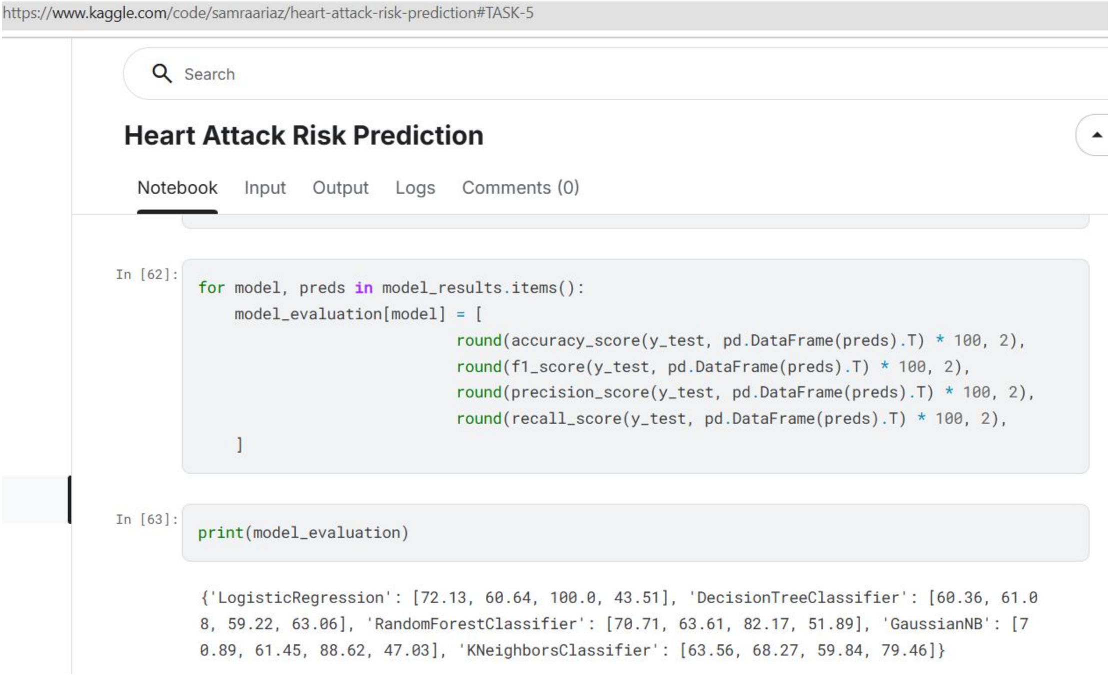
Sakshi_Patel
Dharmik_Patel

People worked on this dataset accuracy founded was same as us:


w.kaggle.com/code/samad0015/heart-attack-risk-prediction#3.-Train-a-Model:
Search
Heart Attack Risk Prediction
Notebook
Input
Output
Logs
Comments (0)

```
(),KNeighborsClassifier(),MLPClassifier()]
model_names=["Logistic Regression","Support Vector Machine","Decision Tree","Random Forest","Ga
ussian Naive Bayes","K-Nearest Neighbors","Multi Layer Perceptron"]
models_scores=[]
for model,model_name in zip(models,model_names):
    model.fit(X_train,y_train)
    y_pred=model.predict(X_test)
    accuracy=accuracy_score(y_test,y_pred)
    models_scores.append([model_name,accuracy])
sorted_models_scores=sorted(models_scores,key=lambda x:x[1],reverse=True)
for model in sorted_models_scores:
    print("Accuracy Score: ",f'{model[0]}: {model[1]*100:.2f}')
```

```
Accuracy Score:  Logistic Regression: 64.32
Accuracy Score:  Support Vector Machine: 64.32
Accuracy Score:  Gaussian Naive Bayes: 64.32
Accuracy Score:  Random Forest: 63.22
Accuracy Score:  K-Nearest Neighbors: 59.49
Accuracy Score:  Multi Layer Perceptron: 57.06
Accuracy Score:  Decision Tree: 53.82
```


https://www.kaggle.com/code/samraariaz/heart-attack-risk-prediction#TASK-5
Search
Heart Attack Risk Prediction
Notebook
Input
Output
Logs
Comments (0)

```
In [62]:
for model, preds in model_results.items():
    model_evaluation[model] = [
                            round(accuracy_score(y_test, pd.DataFrame(preds).T) * 100, 2),
                            round(f1_score(y_test, pd.DataFrame(preds).T) * 100, 2),
                            round(precision_score(y_test, pd.DataFrame(preds).T) * 100, 2),
                            round(recall_score(y_test, pd.DataFrame(preds).T) * 100, 2),
    ]
```

```
In [63]:
print(model_evaluation)
```

```
{'LogisticRegression': [72.13, 60.64, 100.0, 43.51], 'DecisionTreeClassifier': [60.36, 61.0
8, 59.22, 63.06], 'RandomForestClassifier': [70.71, 63.61, 82.17, 51.89], 'GaussianNB': [7
0.89, 61.45, 88.62, 47.03], 'KNeighborsClassifier': [63.56, 68.27, 59.84, 79.46]}
```

Sakshi_Patel
Dharmik_Patel

Researched on Features important to use for model training and medical linkage formula used on features accprding to target variable to change our dataset on basis of it:

Features used in Model training where we secured 88% accuracy:

Demographic Features

Age, Sex, Income

Clinical / Vital Sign Features

Systolic_BP, Diastolic_BP, Heart Rate, Cholesterol, Triglycerides, BMI

Medical History

Diabetes, Family History, Previous Heart Problems, Medication Use

Lifestyle & Behavioural Features

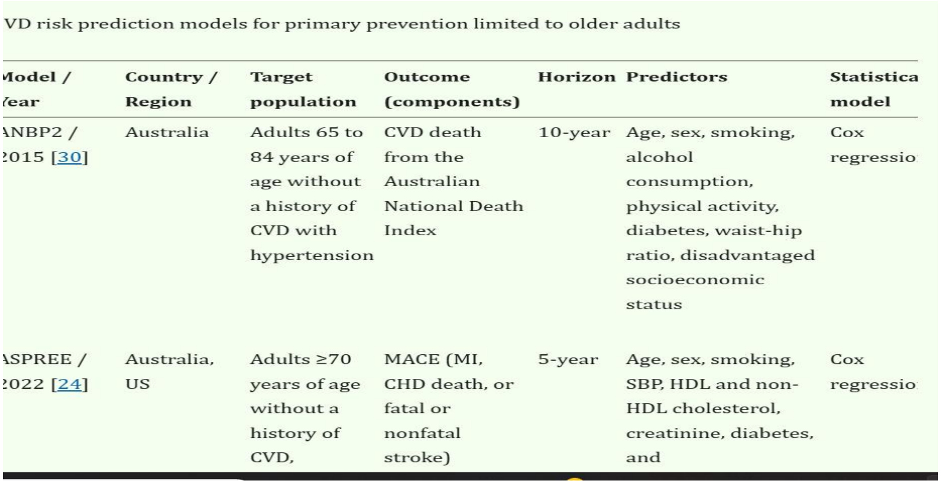
Smoking, Alcohol Consumption, Diet, Exercise Hours Per Week, Physical Activity Days Per Week, Sedentary Hours Per Day, Sleep Hours Per Day, Stress Level

Medical Composite Risk Scores

Metabolic_Score, CV_Risk_Score, Lifestyle_Risk

Research papers that show importance of this features involved in usage of heart disease prediction:

Related to Age:
https://pmc.ncbi.nlm.nih.gov/articles/PMC12420760/

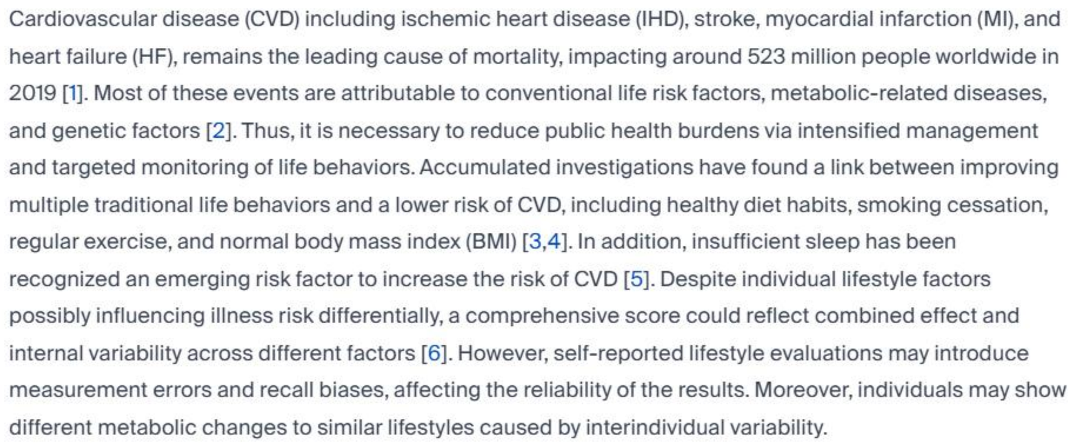


VD risk prediction models for primary prevention limited to older adults

Model / Year	Country / Region	Target population	Outcome (components)	Horizon	Predictors	Statistica model
NBP2 / 2015 [30]	Australia	Adults 65 to 84 years of age without a history of CVD with hypertension	CVD death from the Australian National Death Index	10-year	Age, sex, smoking, alcohol consumption, physical activity, diabetes, waist-hip ratio, disadvantaged socioeconomic status	Cox regressio
SPREE / 2022 [24]	Australia, US	Adults ≥70 years of age without a history of CVD,	MACE (MI, CHD death, or fatal or nonfatal stroke)	5-year	Age, sex, smoking, SBP, HDL and non-HDL cholesterol, creatinine, diabetes, and	Cox regressio

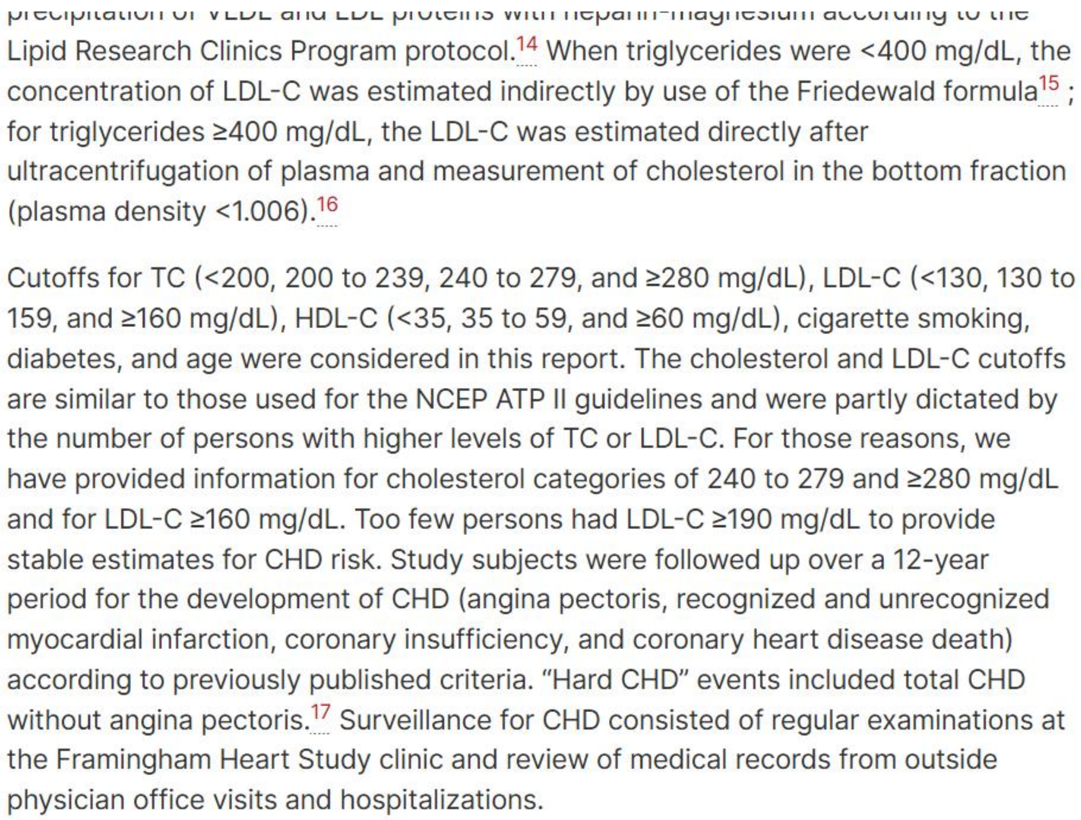
Sakshi_Patel
Dharmik_Patel

Health and Lifestyle related features:
https://www.mdpi.com/2072-6643/16/20/3553



Cardiovascular disease (CVD) including ischemic heart disease (IHD), stroke, myocardial infarction (MI), and heart failure (HF), remains the leading cause of mortality, impacting around 523 million people worldwide in 2019 [1]. Most of these events are attributable to conventional life risk factors, metabolic-related diseases, and genetic factors [2]. Thus, it is necessary to reduce public health burdens via intensified management and targeted monitoring of life behaviors. Accumulated investigations have found a link between improving multiple traditional life behaviors and a lower risk of CVD, including healthy diet habits, smoking cessation, regular exercise, and normal body mass index (BMI) [3,4]. In addition, insufficient sleep has been recognized an emerging risk factor to increase the risk of CVD [5]. Despite individual lifestyle factors possibly influencing illness risk differentially, a comprehensive score could reflect combined effect and internal variability across different factors [6]. However, self-reported lifestyle evaluations may introduce measurement errors and recall biases, affecting the reliability of the results. Moreover, individuals may show different metabolic changes to similar lifestyles caused by interindividual variability.

Based on Health Symptoms:
https://www.ahajournals.org/doi/10.1161/01.CIR.97.18.1837

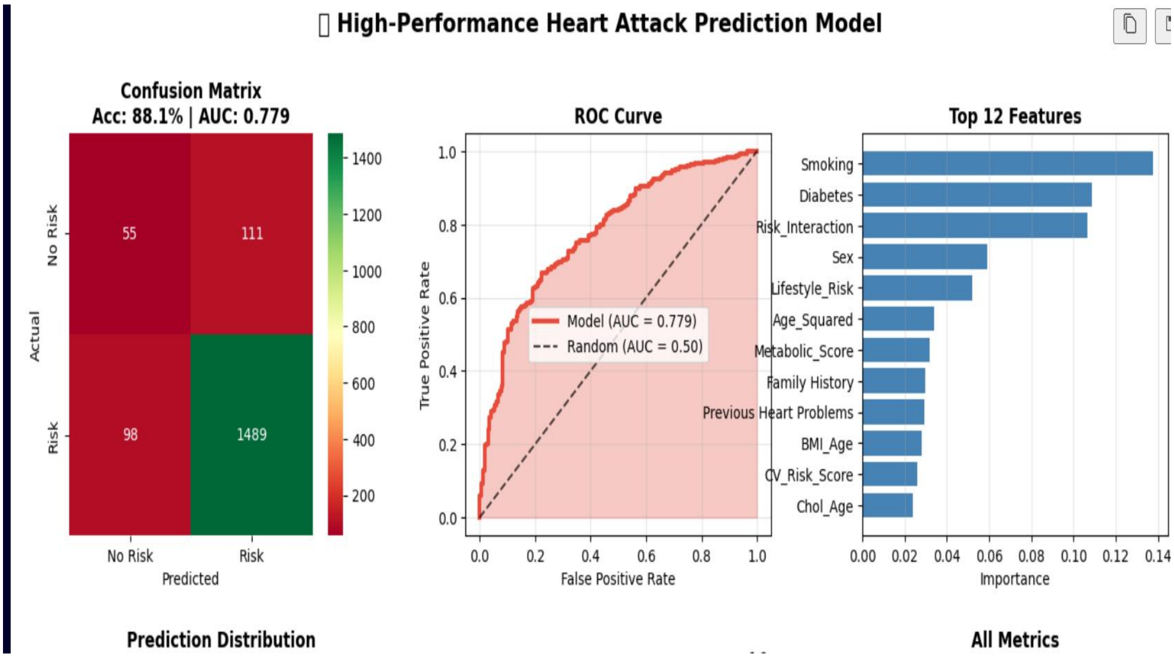


precipitation of VLDL and LDL proteins with heparin-magnesium according to the Lipid Research Clinics Program protocol.[14] When triglycerides were <400 mg/dL, the concentration of LDL-C was estimated indirectly by use of the Friedewald formula[15] ; for triglycerides ≥400 mg/dL, the LDL-C was estimated directly after ultracentrifugation of plasma and measurement of cholesterol in the bottom fraction (plasma density <1.006).[16]

Cutoffs for TC (<200, 200 to 239, 240 to 279, and ≥280 mg/dL), LDL-C (<130, 130 to 159, and ≥160 mg/dL), HDL-C (<35, 35 to 59, and ≥60 mg/dL), cigarette smoking, diabetes, and age were considered in this report. The cholesterol and LDL-C cutoffs are similar to those used for the NCEP ATP II guidelines and were partly dictated by the number of persons with higher levels of TC or LDL-C. For those reasons, we have provided information for cholesterol categories of 240 to 279 and ≥280 mg/dL and for LDL-C ≥160 mg/dL. Too few persons had LDL-C ≥190 mg/dL to provide stable estimates for CHD risk. Study subjects were followed up over a 12-year period for the development of CHD (angina pectoris, recognized and unrecognized myocardial infarction, coronary insufficiency, and coronary heart disease death) according to previously published criteria. "Hard CHD" events included total CHD without angina pectoris.[17] Surveillance for CHD consisted of regular examinations at the Framingham Heart Study clinic and review of medical records from outside physician office visits and hospitalizations.

Sakshi_Patel
Dharmik_Patel

Our Final Result after modifying our dataset with relevant features :



	No Risk	Risk
No Risk	55	111
Risk	98	1489

Output: Accuracy: 88.08%
ROC-AUC: 0.7791
Training Time: 26.53s
CV AUC: 0.7887