

Exploring Machine Learning Applications in Predicting Heart Disease risk

A Comparative Study of Logistic Regression, SVM, and Decision Trees

By: Asiana Holloway

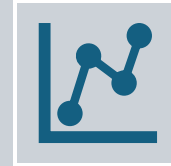
Introduction

- Heart disease is one of the leading causes of death globally.
- Early diagnosis can significantly improve health outcomes.
- Artificial intelligence (AI) and machine learning (ML) have demonstrated significant potential in healthcare.
- This project will explore the application of AI/ML in risk prediction for heart disease.
- Models evaluated: Logistic Regression, SVM, Decision Trees.

Clinical Challenge and Research



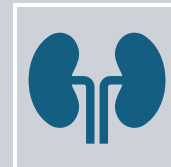
Traditional diagnostic methods may overlook early risk patterns.



ML offers an opportunity to create predictive models using clinical data.



Past research (e.g, Weng et al., 2017) has shown ML models can outperform standard tools.



This study will compare ML models using the UCI Heart Disease dataset from Kaggle.

Model Development and Workflow Overview

- The Python code was written in Google Colab using standard libraries such as Pandas, Scikit-learn, and Matplotlib

Purpose:

- Preprocess and scale the dataset
- Train three models (Logistic Regression, SVM, Decision Tree)
- Evaluate and compare performance using accuracy, precision, recall, F1, and AUC metrics.

Heart Disease Risk Prediction: Logistic Regression, SVM, Decision Tree

```
# Install & import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score, classification_report, confusion_matrix
```

```
df = pd.read_csv('heart_cleveland_upload.csv')
df.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	condition
0	69	1	0	160	234	1	2	131	0	0.1	1	1	0	0
1	69	0	0	140	239	0	0	151	0	1.8	0	2	0	0
2	66	0	0	150	226	0	0	114	0	2.6	2	0	0	0
3	65	1	0	138	282	1	2	174	0	1.4	1	1	0	1
4	64	1	0	110	211	0	2	144	1	1.8	1	0	0	0

```
[ ] # Preprocess the data
X = df.drop('condition', axis=1)
y = df['condition']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Model Performance and Metrics

Logistic Regression:	Accuracy: 0.73
	Precision: 0.70
	Recall: 0.75
	F1: 0.72
	AUC: 0.84
SVM	Accuracy: 0.73
	Precision: 0.69
	Recall: 0.79
	F1: 0.73
	AUC: 0.83
Decision Tree:	Accuracy: 0.77
	Precision: 0.73
	Recall: 0.79
	F1: 0.76
	AUC: 0.77

```
[ ] # Train the models
lr = LogisticRegression().fit(X_train_scaled, y_train)
svm = SVC(probability=True).fit(X_train_scaled, y_train)
dt = DecisionTreeClassifier().fit(X_train, y_train)

# Evaluate models
models = {'Logistic Regression': lr, 'SVM': svm, 'Decision Tree': dt}
for name, model in models.items():
    if name == 'Decision Tree':
        preds = model.predict(X_test)
        proba = model.predict_proba(X_test)[: , 1]
    else:
        preds = model.predict(X_test_scaled)
        proba = model.predict_proba(X_test_scaled)[: , 1]
    print(f"\n{name} Metrics:")
    print("Accuracy:", accuracy_score(y_test, preds))
    print("Precision:", precision_score(y_test, preds))
    print("Recall:", recall_score(y_test, preds))
    print("F1 Score:", f1_score(y_test, preds))
    print("ROC AUC Score:", roc_auc_score(y_test, proba))
```

RESULTS

Logistic Regression Metrics:
Accuracy: 0.7333333333333333
Precision: 0.7
Recall: 0.75
F1 Score: 0.7241379310344828
ROC AUC Score: 0.8415178571428571

SVM Metrics:
Accuracy: 0.7333333333333333
Precision: 0.6875
Recall: 0.7857142857142857
F1 Score: 0.7333333333333333
ROC AUC Score: 0.8337053571428572

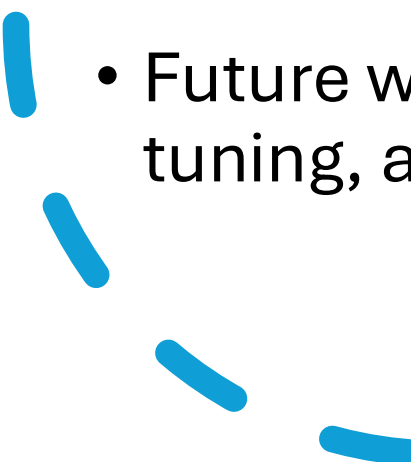
Decision Tree Metrics:
Accuracy: 0.7666666666666667
Precision: 0.7333333333333333
Recall: 0.7857142857142857
F1 Score: 0.7586206896551724
ROC AUC Score: 0.7678571428571428

Literature Benchmarks and Validation

- Weng et al. (2017) showed ML improves cardiovascular risk prediction using clinical data.
- Deo (2015) emphasized ML's role in reducing diagnostic errors.
- Gudadhe et al. (2010) and IEEE (2022) supported SVM and Decision Tree effectiveness.
- My model showed similar or slightly lower AUC and recall.
- Decision tree had the highest F1 score, supporting everyday medical use.
- results align with prior studies, providing clinical potential.



Conclusion

- All three models performed well with slight differences.
 - Decision Tree showed the best balance of metrics.
 - This supports the integration of ML into healthcare diagnostics.
 - Findings align with prior studies and support the use of ML in clinical risk prediction.
 - Future work could explore additional datasets, hyperparameter tuning, and ensemble methods.
- 

References:

- Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930. <https://doi.org/10.1161/CIRCULATIONAHA.115.001593>
- 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>
- Author(s). (2022). A comparison based study of supervised machine learning algorithms for prediction of heart disease. *Proceedings of the 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*. <https://ieeexplore.ieee.org/document/9844328/>
- Heart Disease UCI Dataset – Kaggle. (n.d.). <https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci>
- Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebraei, H., Shoeibi, A., ... & Acharya, U. R. (2018). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, 100, 103–111. <https://doi.org/10.1016/j.combiomed.2018.05.013>
- Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5), 304–310. [https://doi.org/10.1016/0002-9149\(89\)90524-9](https://doi.org/10.1016/0002-9149(89)90524-9)
- Gudadhe, M., Wankhade, K., & Dongre, S. (2010). Decision support system for heart disease based on support vector machine and artificial neural network. *International Journal of Computer Applications*, 7(13), 1–5. <https://ieeexplore.ieee.org/document/5640377>