A Course Based Project Report on

Heart Disease Prediction using Machine Learning

Submitted to the

Department of Information Technology

in partial fulfillment of the requirements for the completion of course
Machine Learning Laboratory (22PC2CS302)

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

Submitted by

LAKSHITA	22071A12F6
S. JYOTHSNA RAO	22071A12J7
T. THANMAYEE	22071A12J8
T. KIRANMAYI	22071A12J9

Under the guidance of

Dr.G Madhu

(Course Instructor)

Professor, Department of IT, VNRVJIET

TAMASOMA JYOTIRGAMAYA
Estd.1995

DEPARTMENT OF INFORMATION TECHNOLOGY

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING & TECHNOLOGY

An Autonomous Institute, NAAC Accredited with 'A++' Grade, NBA

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet (S.O), Hyderabad – 500 090, TS, India

MAY 2025

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institute, NAAC Accredited with 'A++' Grade, NBA Accredited for CE, EEE, ME, ECE, CSE, EIE, IT B. Tech Courses, Approved by AICTE, New Delhi, Affiliated to JNTUH, Recognized as "College with Potential for Excellence" by UGC, ISO 9001:2015 Certified, QS I GUAGE Diamond Rated

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet(SO), Hyderabad-500090, TS, India

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project report entitled "HEART DISEASE PREDICTION USING MACHINE LEARNING" is a bonafide work done under our supervision and is being submitted by Lakshita (22071A12F6) , S . Jyothsna Rao (22071A12J7) ,T. Thanmayee (22071A12J8) , T. Kiranmayi (22071A12J9) in partial fulfilment for the award of the degree of Bachelor of Technology in Information Technology, of the VNRVJIET, Hyderabad during the academic year 2024-2025.

Dr. G. Madhu

Professor

Department of IT

Dr. N Mangathayaru

Professor & HOD

Department of IT

Course based Projects Reviewer

VALLURUPALLI NAGESWARA RAO VIGNANA JYOTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

An Autonomous Institute, NAAC Accredited with 'A++' Grade,

Vignana Jyothi Nagar, Pragathi Nagar, Nizampet(SO), Hyderabad-500090, TS, India

DEPARTMENT OF INFORMATION TECHNOLOGY

DECLARATION

We declare that the course based project work entitled "**HEART DISEASE PREDICTION USING MACHINE LEARNING**" submitted in the Department of Information Technology, Vallurupalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Hyderabad, in partial fulfilment of the requirement for the award of the degree of **Bachelor of Technology in Information Technology** is a bonafide record of our own work carried out under the supervision of **Dr.G Madhu, Professor, Department of IT, VNRVJIET.** Also, we declare that the matter embodied in this thesis has not been submitted by us in full or in any part thereof for the award of any degree/diploma of any other institution or university previously.

Place: Hyderabad.

Lakshita	**(22071A12F6)**
S. Jyothsna Rao	**(22071A12J7)**
T. Thanmayee	**(22071A12J8)**
T. Kiranmayi	**(22071A12J9)**

ACKNOWLEDGEMENT

We express our deep sense of gratitude to our beloved President, Sri. D. Suresh Babu, VNR Vignana Jyothi Institute of Engineering & Technology for the valuable guidance and for permitting us to carry out this project.

With immense pleasure, we record our deep sense of gratitude to our beloved Principal, Dr.C.D Naidu, for permitting us to carry out this project.

We express our deep sense of gratitude to our beloved Professor **Dr. N MANGATHAYARU**, Professor and Head, Department of Information Technology, VNR Vignana Jyothi Institute of Engineering & Technology, Hyderabad- 500090 for the valuable guidance and suggestions, keen interest and through encouragement extended throughout the period of project work.

We take immense pleasure to express our deep sense of gratitude to our beloved Guide,

Dr. G Madhu, Professor in Information Technology, VNR Vignana JyothiInstitute of Engineering & Technology, Hyderabad, for his/her valuable suggestions andrare insights, for constant source of encouragement and inspiration throughout my projectwork.

We express our thanks to all those who contributed for the successful completion of our project work.

Lakshita	**(22071A12F6)**
S. Jyothsna Rao	**(22071A12J7)**
T. Thanmayee	**(22071A12J8)**
T. Kiranmayi	**(22071A12J9)**

ABSTRACT

Heart disease remains one of the leading causes of mortality worldwide, making early detection and accurate prediction crucial for effective treatment and prevention. Traditional diagnostic methods rely heavily on clinical examinations, patient history, and basic statistical analysis, which are limited by their dependence on human interpretation and inability to identify complex patterns in large datasets.

This project implements a Heart Disease Prediction system using Machine Learning algorithms to overcome these limitations. By leveraging three powerful classification techniques—Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN)—our system can effectively analyze patient medical data and provide accurate predictions for heart disease diagnosis.

The methodology involves comprehensive data preprocessing of patient records including handling missing values and feature scaling, development of optimized machine learning models, and rigorous evaluation using metrics such as accuracy, precision, recall, and F1-score. Through comparative analysis, we determine which algorithm delivers the most reliable predictions for heart disease detection.

Our results demonstrate that machine learning approaches can achieve high accuracy in predicting heart disease risk, providing a valuable tool for medical professionals to support clinical decision-making. The comparison between the different algorithms highlights their respective strengths and limitations in the context of medical diagnosis, offering insights into the most effective approaches for similar healthcare applications.

By combining data preprocessing techniques with powerful classification algorithms, our Heart Disease Prediction system represents a significant step forward in computer-aided diagnosis technologies, with potential to improve early detection rates and patient outcom

TABLE OF CONTENTS

1. INTRODUCTION

Heart disease is still among the world's top causes of death, and therefore early detection and proper prediction are essential to ensure proper treatment and prevention. The conventional way of heart disease diagnosis uses clinical examination, patient history, and statistical models. These, though, are limited by their need for human interpretation and capability to recognize intricate patterns in large sets of data.

Machine learning has emerged as a powerful tool in medical diagnosis, offering the ability to analyze complex datasets and identify patterns that might not be apparent through traditional statistical methods. By training algorithms on historical patient data, machine learning models can learn to recognize risk factors and combinations of features that are indicative of heart disease, potentially leading to earlier and more accurate diagnoses.

The existing systems for heart disease prediction include traditional statistical approaches and basic rule-based systems. While these approaches have been used for decades, they often fail to capture the complex interrelationships between multiple risk factors. Recent advancements in machine learning have led to the development of more sophisticated prediction models that can analyze numerous variables simultaneously and detect subtle patterns that might be missed by conventional methods.

Our proposed system leverages three well-established machine learning algorithms—Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN)—to create a comprehensive heart disease prediction tool. Logistic Regression provides a probabilistic approach to classification, making it suitable for risk assessment. Random Forest, an ensemble learning method, combines multiple decision trees to improve accuracy and handle complex relationships in the data. KNN, a non-parametric method, classifies cases based on similarity to known examples, making it effective for detecting patterns in medical data.

By implementing and comparing these diverse algorithms, our system aims to identify the most effective approach for heart disease prediction while providing insights into the relative importance of different risk factors.

2. OBJECTIVE

The primary objective of this project is to develop a robust and accurate Heart Disease Prediction system using Deep Learning techniques, specifically Artificial Neural Networks (ANNs). This system aims to assist medical professionals in early detection and diagnosis of heart disease, potentially leading to timely intervention and improved patient outcomes.

The second objective is to implement a comprehensive data engineering pipeline that efficiently processes and prepares medical data for deep learning analysis. This includes handling missing values, normalizing features, and transforming categorical variables to ensure optimal model performance. By establishing a structured approach to data preprocessing, we aim to create a system that can handle diverse medical datasets and maintain high prediction accuracy.

The third objective is to compare the performance of our deep learning approach with traditional machine learning methods to demonstrate the advantages of using neural networks for this application. Through extensive evaluation using metrics such as accuracy, precision, recall, and F1-score, we aim to provide quantitative evidence of the superiority of our approach in terms of prediction capabilities.

The final objective is to develop a user-friendly interface using Streamlit that allows healthcare professionals to easily input patient data and receive instant predictions.

3. SOURCE CODE

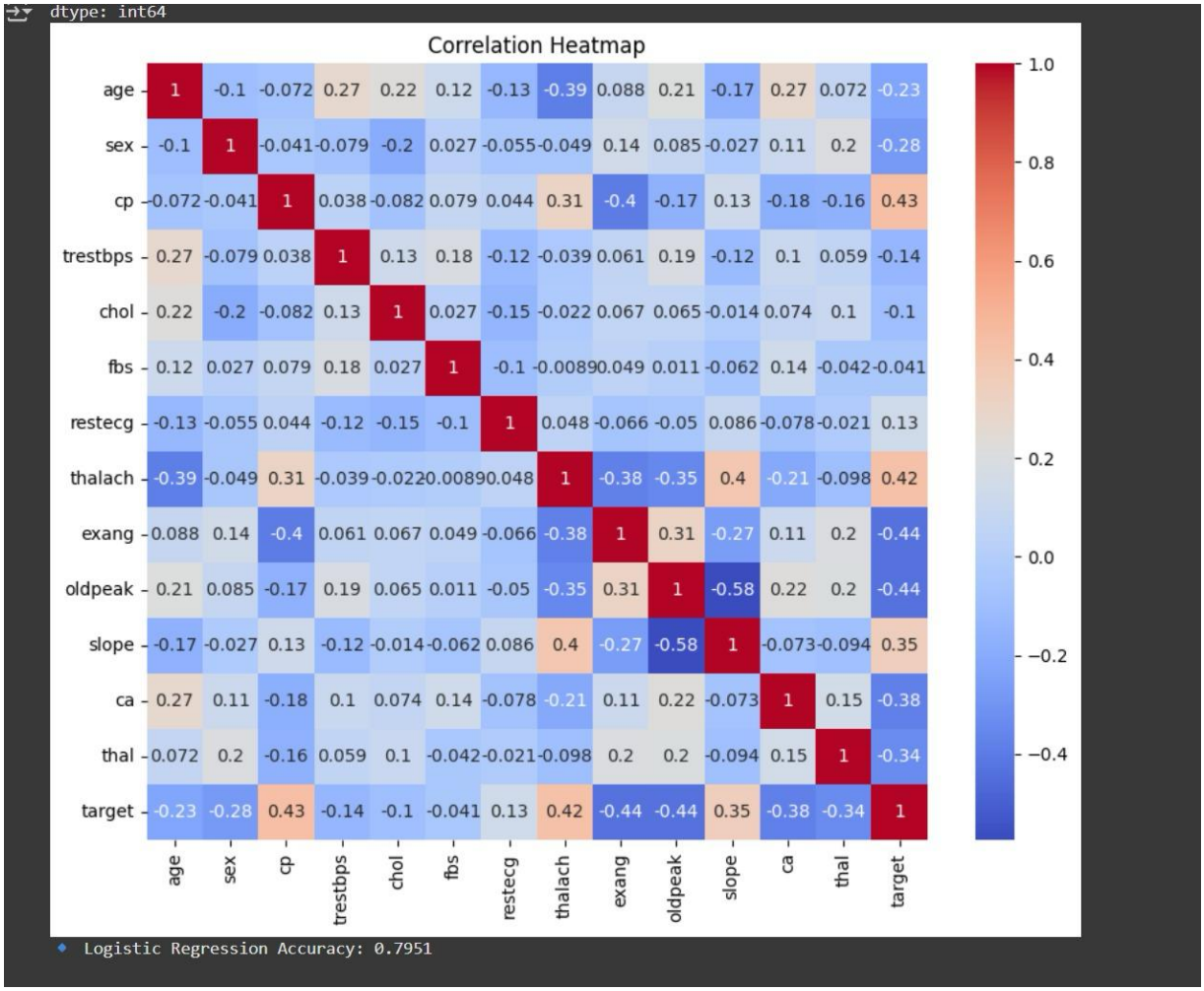
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv('/content/heart.csv')  # Change path if needed

# Check for missing values
print(df.isnull().sum())

# Fill missing values if any
df.fillna(df.median(numeric_only=True), inplace=True)

# Encode categorical columns if any (optional)
cat_cols = df.select_dtypes(include='object').columns
if len(cat_cols):
    df = pd.get_dummies(df, columns=cat_cols, drop_first=True)

# Heatmap (optional)
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()

# Feature-target split
X = df.drop('target', axis=1)
y = df['target']

# Scale the features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

 ------------------------- Logistic Regression -------------------------

# Initialize and train Logistic Regression model
log_model = LogisticRegression()
log_model.fit(X_train, y_train)

# Predict
y_pred_log = log_model.predict(X_test)
```

```
# Accuracy
log_accuracy = accuracy_score(y_test, y_pred_log)
print(f"□ Logistic Regression Accuracy: {log_accuracy:.4f}")

# Confusion Matrix
log_cm = confusion_matrix(y_test, y_pred_log)
print("\nConfusion Matrix:\n", log_cm)
print("\nClassification Report:\n", classification_report(y_test, y_pred_log))

# Confusion Matrix Plot
plt.figure(figsize=(6,4))
sns.heatmap(log_cm, annot=True, fmt='d', cmap='Blues', xticklabels=['No Disease', 'Disease'],
yticklabels=['No Disease', 'Disease'])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Logistic Regression - Confusion Matrix")
plt.show()

-------------------------- Random Forest --------------------------

# Initialize and train Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predict
y_pred_rf = rf_model.predict(X_test)

# Accuracy
rf_accuracy = accuracy_score(y_test, y_pred_rf)
print(f"□ Random Forest Accuracy: {rf_accuracy:.4f}")

# Confusion Matrix
rf_cm = confusion_matrix(y_test, y_pred_rf)
print("\nConfusion Matrix:\n", rf_cm)
print("\nClassification Report:\n", classification_report(y_test, y_pred_rf))

# Confusion Matrix Plot
plt.figure(figsize=(6,4))
sns.heatmap(rf_cm, annot=True, fmt='d', cmap='Purples', xticklabels=['No Disease', 'Disease'],
yticklabels=['No Disease', 'Disease'])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Random Forest - Confusion Matrix")
plt.show()

-------------------------- K-Nearest Neighbors (KNN) --------------------------

# Initialize and train K-Nearest Neighbors (KNN) model
knn_model = KNeighborsClassifier(n_neighbors=5)
knn_model.fit(X_train, y_train)

# Predict
```

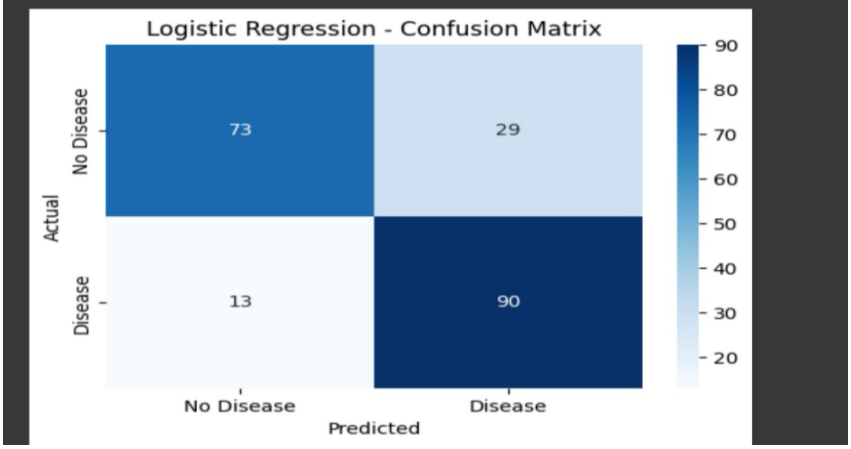
```
y_pred_knn = knn_model.predict(X_test)

# Accuracy
knn_accuracy = accuracy_score(y_test, y_pred_knn)
print(f"□ K-Nearest Neighbors Accuracy: {knn_accuracy:.4f}")

# Confusion Matrix
knn_cm = confusion_matrix(y_test, y_pred_knn)
print("\nConfusion Matrix:\n", knn_cm)
print("\nClassification Report:\n", classification_report(y_test, y_pred_knn))

# Confusion Matrix Plot
plt.figure(figsize=(6,4))
sns.heatmap(knn_cm, annot=True, fmt='d', cmap='Oranges', xticklabels=['No Disease', 'Disease'],
yticklabels=['No Disease', 'Disease'])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("K-Nearest Neighbors - Confusion Matrix")
plt.show
```

4. OUTPUT

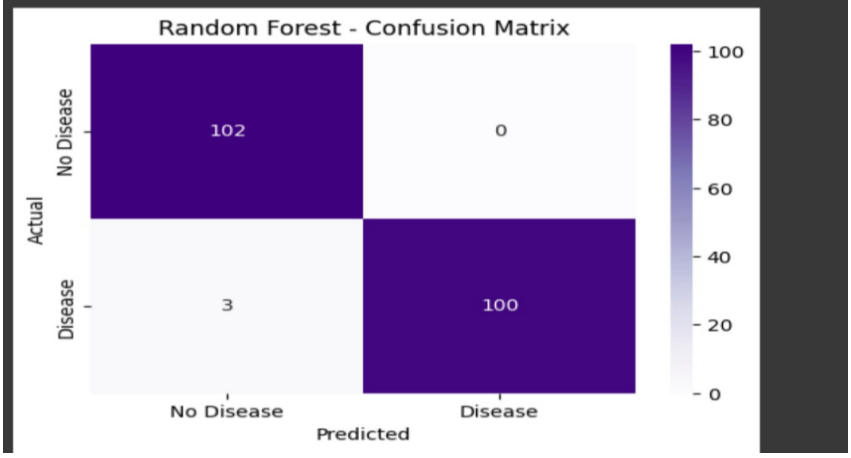
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```



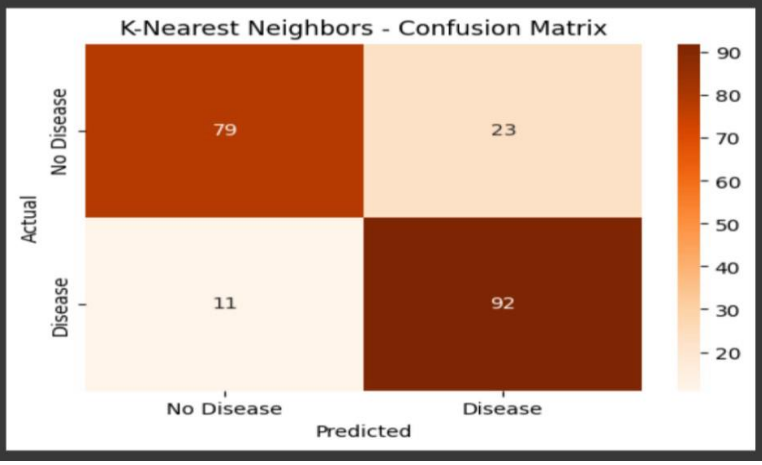
```
🔹 Logistic Regression Accuracy: 0.7951

Confusion Matrix:
 [[73 29]
 [13 90]]

Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.72      0.78       102
           1       0.76      0.87      0.81       103

    accuracy                           0.80       205
   macro avg       0.80      0.79      0.79       205
weighted avg       0.80      0.80      0.79       205
```




```
🔹 Random Forest Accuracy: 0.9854

Confusion Matrix:
 [[102   0]
 [  3 100]]

Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       102
           1       1.00      0.97      0.99       103

    accuracy                           0.99       205
   macro avg       0.99      0.99      0.99       205
weighted avg       0.99      0.99      0.99       205
```



```
🔹 K-Nearest Neighbors Accuracy: 0.8341

Confusion Matrix:
 [[79 23]
 [11 92]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.77      0.82       102
           1       0.80      0.89      0.84       103

    accuracy                           0.83       205
   macro avg       0.84      0.83      0.83       205
weighted avg       0.84      0.83      0.83       205
```



5. CONCLUSION

In this study on heart disease prediction using machine learning, we evaluated the performance of three widely used classification algorithms: Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN). Each model exhibited its own advantages in terms of accuracy, interpretability, and computational behavior.

Logistic Regression served as a reliable baseline model. Its straightforward implementation and interpretability made it effective for understanding how individual features influenced heart disease risk. However, its linear decision boundaries limited its ability to capture complex relationships within the data.

K-Nearest Neighbors offered an intuitive approach by classifying cases based on proximity to known instances. It performed reasonably well when the features were properly scaled and the value of *k* carefully tuned. Nonetheless, its prediction speed and sensitivity to noise in the data posed challenges, especially as the dataset size increased.

Among the three, **Random Forest consistently outperformed the others in terms of predictive accuracy and robustness**. As an ensemble learning method, it effectively captured non-linear patterns, handled feature interactions automatically, and was less prone to overfitting. Moreover, it provided valuable insights through feature importance scores, aiding in the interpretability of the model from a data perspective—even if the internal mechanics remain complex.

In conclusion, Random Forest emerged as the most effective and reliable model for predicting heart disease, combining strong predictive performance with robustness and generalizability. It stands out as a preferred choice for deployment in real-world healthcare applications where accuracy and dependability are paramount.

.

6. REFERENCES

1. **Dua, D., & Graff, C. (2019). UCI Machine Learning Repository: Heart Disease Dataset.**
   – Dataset source used for training and testing the prediction models.
   https://archive.ics.uci.edu/ml/datasets/heart+Disease
2. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.**
   – Core reference for implementing machine learning models such as Logistic Regression, Random Forest, and KNN.
   https://scikit-learn.org/
3. **McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.**
   – Reference for using the pandas library for data loading, cleaning, and manipulation.
   https://pandas.pydata.org/
4. **Oliphant, T. E. (2006). A guide to NumPy. *Trelgol Publishing.*
   – Used for numerical computations and handling arrays with NumPy.
   https://numpy.org/
5. **Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.**
   – Used for plotting confusion matrices and accuracy comparisons.
   https://matplotlib.org/
6. **Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.**
   – Used for enhanced heatmap and visualization generation (correlation matrix, confusion matrices).
   https://seaborn.pydata.org/
7. **James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.**
   – A comprehensive guide to classification methods including Logistic Regression and KNN.