Project – Report

Topic: Empowering Early Detection of Heart Attack Risks with Machine Learning

Student: Aarti Gupta
Machine Learning Project

1. Introduction

Heart attacks are one of the leading causes of death worldwide. Early identification of individuals at risk can significantly improve outcomes through timely intervention. This project leverages machine learning to predict heart attack risks based on individual health indicators, helping in early diagnosis and preventive care.

2. Objective

To develop a reliable machine learning model that can predict whether an individual is at risk of heart disease using health-related features.

3. Dataset Description

- **Source:** Kaggle – Personal Key Indicators of Heart Disease by Kamil Pytlak
- **Size:** 319,795 rows
- **Features:**
  - Demographic: Age, Sex, Race
  - Behavioral: Smoking, Alcohol, Physical Activity, Sleep, Mental Health
  - Medical: BMI, Stroke, Diabetes, High Blood Pressure, Cholesterol
- **Target Variable:** HeartDisease (0 = No, 1 = Yes)

4. Methodology

Data Preprocessing

- Categorical encoding (Label Encoding & One-Hot)
- Handling class imbalance
- Feature scaling
- Train-test split (80–20)

Exploratory Data Analysis (EDA)

- Visualized distribution of target and features
- Correlation matrix
- Identified risk patterns based on lifestyle and medical history

Model Building

Eight models were implemented:

- Logistic Regression
- K-Nearest Neighbors
- Naive Bayes
- Decision Tree
- Random Forest
- AdaBoost
- Gradient Boosting
- XGBoost

Evaluation Metrics

Models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score

5. Results and Interpretation

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.84	0.76	0.78	0.77
Random Forest	0.83	0.74	0.80	0.77
AdaBoost	**0.85**	0.78	0.79	0.78
Gradient Boost	0.84	0.77	0.81	0.79
XGBoost	0.84	0.76	0.81	0.78

Key Insights

- **AdaBoost** had the highest test accuracy (0.85), closely followed by Gradient Boost and XGBoost.
- **Recall and accuracy were consistent** between train and test sets, indicating good generalization.
- Precision varied slightly, suggesting the model occasionally predicted false positives.

6. Conclusion

The project successfully built machine learning models that can detect heart disease risk with good accuracy. AdaBoost and Gradient Boost performed best. The results can be used in clinical settings or awareness tools to help individuals take preventive actions.

7. Future Scope

- Threshold tuning to improve precision
- Deploying model using Streamlit or Flask
- Incorporating real-time data or electronic health records
- Using SHAP or LIME for model explainability

8. References

Pytlak, K. (2021). *Personal Key Indicators of Heart Disease* [Dataset]. Kaggle. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease