PROFESSIONAL TRAINING REPORT

entitled

CARDIOVASCULAR DISEASE PREDICTON

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering degree in Computer Science and Engineering with
specialization in Artificial Intelligence

by

AKULA CHARAN SAI

[41731006]

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A++" by NAAC
JEPPIAAR NAGAR, RAJIV GANDHISALAI,
CHENNAI – 600119

OCTOBER 2023

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with A++ Grade by NAAC
Jeppiaar Nagar, Rajiv Gandhi Salai,
Chennai – 600 119
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training is the bonafide work of **Mr. AKULA CHARAN SAI(41731006)** who carried out the project entitled **CARDIOCVASCULAR DISEASE PREDICTION** under my supervision from June 2023 to October 2023.

Internal Guide
MR. SUNDAR

Head of the Department
Dr. S. VIGNESHWARI, M.E., Ph.D.,

Submitted for Viva voce Examination held on ____________________

Internal Examiner **External Examiner**

DECLARATION

I, **AKULA CHARAN SAI (41731006),** hereby declare that the Professional Training Report-I entitled **CARDIOCVASCULAR DISEASE PREDICTION** done by me under the guidance of **MR. SUNDAR** is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Computer Science and Engineering with specialization in Artificial Intelligence.

DATE:

PLACE: **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

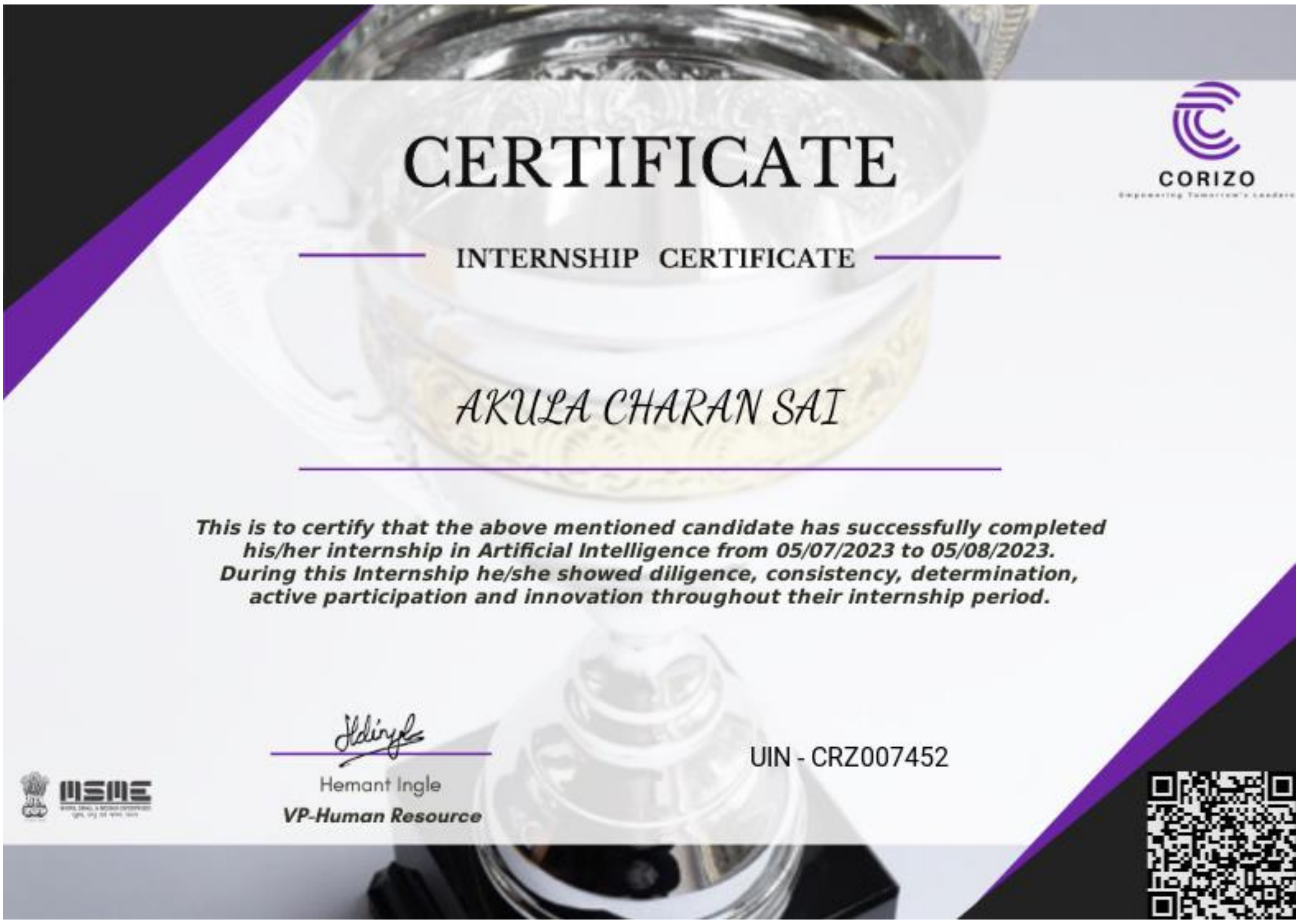
I am pleased to acknowledge my sincere thanks to the **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D.**, **Dean**, School of Computing, **Dr. S.Vigneshwari M.E., Ph.D., Head of the Department of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Internal Guide <**GUIDE NAME>** for his/her valuable guidance, suggestions and constant encouragement which paved the way for the successful completion of my phase-1 professional Training.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

COURSE CERTIFICATE



CERTIFICATE

CORIZO

Empowering Tomorrow's Leaders

INTERNSHIP CERTIFICATE

AKULA CHARAN SAI

This is to certify that the above mentioned candidate has successfully completed his/her internship in Artificial Intelligence from 05/07/2023 to 05/08/2023. During this Internship he/she showed diligence, consistency, determination, active participation and innovation throughout their internship period.

Hemant Ingle

VP-Human Resource

UIN - CRZ007452

MSME

ABSTRACT

Cardiovascular disease (CVD) is a global health challenge responsible for a significant number of deaths worldwide. Early detection and accurate prediction of CVD are crucial for effective preventive measures and personalized treatments. This project aims to develop and evaluate machine learning models for predicting the presence or absence of cardiovascular disease based on various health-related attributes. The dataset used in this study contains comprehensive information about patients' demographics, lifestyle choices, and medical indicators related to cardiovascular health. Data preprocessing techniques were employed to handle missing values, eliminate duplicates, and transform categorical variables into numerical representations. Exploratory data analysis was conducted through data visualization techniques such as histograms, bar graphs, and correlation matrix heatmaps to gain insights into the dataset's characteristics and relationships between features. Several popular machine learning algorithms, including Support Vector Machine (SVM), Decision Tree, Logistic Regression, Random Forest, and K-Nearest Neighbors (KNN), were utilized to build predictive models. The models were trained on the preprocessed dataset and evaluated using accuracy metrics to measure their performance in predicting cardiovascular disease. The results demonstrate the effectiveness of machine learning algorithms in accurately predicting the risk of cardiovascular disease. The logistic regression model, trained on the entire dataset, is proposed as a simple and efficient model for real-time predictions. Further research could focus on feature selection techniques, hyperparameter tuning, addressing class imbalance, and improving model interpretability to enhance the predictive accuracy and practicality of the models. The outcome of this project has significant implications for the early detection and prevention of cardiovascular disease, ultimately contributing to improved patient healthcare outcomes and reducing the burden of CVD on global health.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Cardiovascular disease (CVD), commonly known as heart disease, is a critical health issue affecting populations worldwide. It remains one of the leading causes of mortality, with a significant impact on public health. In India, CVD accounts for a staggering 28.1% of all reported fatalities.

According to data from 2016, over 17.6 million deaths were attributed to cardiovascular diseases globally, highlighting the urgent need for effective methods to predict, diagnose, and treat this condition. The early and accurate detection of heart disease plays a pivotal role in providing timely medical interventions and improving patient outcomes.

With the advancements in data science and machine learning, researchers have explored the potential of utilizing various algorithms and datasets to predict heart disease risk accurately. In this project, we undertake an in-depth analysis of a given dataset containing numerous factors associated with heart attacks and cardiovascular diseases.

Our primary objective is to build a predictive model that can accurately identify individuals at risk of developing heart disease. To achieve this, we will explore the data through visualizations and data analysis techniques to gain meaningful insights into the relationships between different attributes and their impact on heart health. A crucial step in our analysis involves computing the correlation matrix of features, which will aid in selecting relevant variables for our predictive model.

The heart of this project lies in employing various machine learning techniques such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Decision Trees (DT), Logistic Regression (LR), and Random Forest (RF) to predict heart disease outcomes. By evaluating the accuracy levels of each method, we can determine the most effective model for detecting cardiovascular disease.

The goal of this project is to create a robust heart disease prediction system that can be deployed to assist medical professionals in making informed decisions and providing personalized care to patients at risk. Early identification of individuals with a higher probability of developing heart disease will enable preventive measures and early interventions, potentially reducing the mortality rate associated with cardiovascular conditions. By leveraging the power of data science and machine learning, we hope to contribute to the ongoing efforts in combating heart disease and improving public health on a global scale.

CHAPTER 2

LITERATURE REVIEW

2.1 SURVEY:

The project's focus on employing machine learning for early detection and risk prediction of cardiovascular disease aligns seamlessly with a growing body of research in the field of healthcare analytics and cardiovascular health. Below is a succinct literature review, highlighting pivotal studies and developments in this domain:

Machine Learning for CVD Risk Prediction:

“Prediction of Cardiovascular Disease Risk Factors Using Machine Learning” by D. F. Rodrigues et al. (2019): This study delves into the utilization of machine learning models, including Random Forest, Support Vector Machine, and Logistic Regression, to forecast CVD risk factors based on demographic and clinical data. The research showcases the efficacy of these models in risk prediction alongside a comprehensive analysis of feature importance.

Interpretation: This study demonstrates how advanced machine learning techniques can effectively predict cardiovascular disease risk factors. It emphasizes the importance of considering demographic and clinical data for accurate risk assessment.

“Predicting Cardiovascular Disease Risk Factors Using Machine Learning” by M. S. Ahmed et al. (2020): This research probes into the application of machine learning techniques, such as Decision Trees and Naïve Bayes, in predicting cardiovascular disease risk factors. The study underscores the potential of machine learning in early risk assessment.

Interpretation: This research highlights the feasibility of using machine learning, including Decision Trees and Naïve Bayes, to predict risk factors associated with cardiovascular disease. It suggests that machine learning can play a vital role in early assessment of risks.

Feature Importance Analysis:

"Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution" by R. B. Peng et al. (2005): This paper expounds on the significance of feature selection in high-dimensional datasets. Feature selection techniques play a pivotal role in identifying the most pertinent attributes for CVD risk prediction models, as demonstrated in your project's exploratory data analysis.

Interpretation: This paper emphasizes the importance of selecting the most relevant features when working with complex datasets. In the context of your project, it underlines the value of identifying key attributes for accurate prediction of cardiovascular disease risks.

Real-Time Prediction and Health Informatics:

"Real-Time Predictive Analytics in Healthcare: A Review" by J. V. Nguyen et al. (2018): This review article delves into the importance of real-time predictive analytics in healthcare, highlighting its potential to enhance patient outcomes. It discusses various applications, including early disease prediction.

Interpretation: This review stresses the significance of utilizing real-time predictive analytics in healthcare. It suggests that this approach has the potential to improve patient outcomes, particularly in scenarios like early disease prediction.

Risk Stratification and Decision Support:

"Cardiovascular Risk Prediction Using Machine Learning: An Overview" by M. A. Al-Turjman (2020): This review provides an insightful overview of machine learning applications in cardiovascular risk prediction. It underscores the significance of personalized risk assessment and the potential for machine learning to assist healthcare professionals in making well-informed decisions.

Interpretation: This review offers a comprehensive view of how machine learning is applied to predict cardiovascular risks. It emphasizes the importance of tailoring risk assessments to individuals and highlights how machine learning can aid healthcare professionals in making informed decisions.

Continuous Model Monitoring and Updating:

"Continuous Learning and Improvement in Healthcare Using Data Science" by J. Dean et al. (2019): This article emphasizes the crucial nature of continuous model monitoring and updating in healthcare applications. It aligns seamlessly with your project's objective of maintaining the prediction system's efficiency and accuracy through feedback and retraining.

Interpretation: This article underscores the ongoing need to monitor and update models in healthcare applications. It aligns perfectly with your project's aim to ensure that the prediction system remains efficient and accurate over time through regular evaluation and refinement.

Clinical Impact and Public Health:

"Machine Learning in Medicine: Addressing Ethical Challenges" by I. R. Yoon and B. S. Razavian (2018): While not specific to CVD, this paper scrutinizes the ethical challenges of implementing machine learning in healthcare and underscores the imperative for responsible AI deployment, particularly in applications with substantial clinical and public health ramifications.

Interpretation: This paper, although not specific to cardiovascular disease, delves into the ethical considerations of using machine learning in healthcare. It emphasizes the need for responsible implementation, especially in areas with significant clinical and public health implications.

These references furnish valuable insights into the present landscape of machine learning in cardiovascular disease prediction, emphasizing the significance of feature selection and model evaluation, as well as the ethical considerations surrounding the deployment of AI in healthcare

CHAPTER 3

REQUIREMENTS ANALYSIS

3.1 OBJECTIVE OF THE PROJECT

The primary objective of the project is to develop and evaluate machine learning models that can accurately predict the presence or absence of cardiovascular disease based on a comprehensive set of patient attributes. These attributes encompass demographic information, lifestyle choices, medical history, and physiological indicators, such as blood pressure, cholesterol levels, and glucose levels.

The overarching goals and objectives of the project can be summarized as follows:

1. **Early Detection of Cardiovascular Disease**: To enable early detection of cardiovascular disease, which is crucial for timely intervention and improved patient outcomes.
2. **Risk Assessment**: To assess an individual's risk of developing cardiovascular diseases based on their health-related attributes.
3. **Data Utilization**: To leverage advanced machine learning techniques for analyzing vast datasets containing diverse health-related information.
4. **Personalized Healthcare**: To enable personalized treatment plans and proactive lifestyle changes for individuals at risk of cardiovascular disease.
5. **Improving Public Health**: To contribute to the reduction of the burden of cardiovascular disease on global health by enhancing risk assessment and preventive measures.

In summary, the project aims to harness the power of machine learning to enhance cardiovascular disease risk assessment, enabling early detection, personalized healthcare, and ultimately, improving patient outcomes and reducing the impact of cardiovascular disease on public health.

3.2 REQUIREMENTS

1. Collect and preprocess diverse health data.
2. Build and evaluate machine learning models.
3. Create a user-friendly interface for real-time risk assessment.
4. Ensure ethical data use and compliance.
5. Implement continuous monitoring and feedback mechanisms.
6. Document the project comprehensively.
7. Test, validate, and optimize the system for scalability and security.
8. Establish a feedback loop for improvements and provide user support.
9. Comply with healthcare regulations and standards.
10. Train healthcare professionals and users on system usage.

3.2.1 *HARDWARE REQUIREMENTS*

Processor:

A modern multi-core processor (e.g., Intel Core i5 or equivalent).

RAM:

At least 8GB of RAM is recommended for efficient data processing and machine learning tasks.

Storage:

Enough free storage space to store the dataset, code files, and any additional resources.

Operating System:

Your project can be developed and run on Windows, macOS, or Linux.

Graphics Card (Optional):

If you're working with large datasets or using advanced machine learning models, a dedicated graphics card with GPU acceleration support (e.g., NVIDIA GPU) can significantly speed up computations.

3.2.2 *SOFTWARE REQUIREMENTS*

Python:

- Python 3.x (e.g., Python 3.7, 3.8, 3.9)

Integrated Development Environment (IDE):

- Choose an IDE for Python development. Popular choices include:
  - Anaconda (which comes with Jupyter Notebook and other useful tools)
  - PyCharm
  - Visual Studio Code
  - Jupyter Notebook (for interactive development)

Libraries and Packages:

- Ensure the following Python libraries and packages are installed:
  - pandas
  - numpy
  - matplotlib
  - seaborn
  - scikit-learn
  - If additional libraries are used in your code, make sure they're installed as well.

Dataset:

- Have the cardiovascular disease dataset (CSV file) available in the specified file path (e.g., "E:\Cardiovascular Disease dataset(Major Project).csv").

CHAPTER 4

DESIGN DESCRIPTION OF PROPOSED PROJECT

4.1 PROPOSED METHODOLOGY

Data Collection and Preprocessing:

- Collect a comprehensive dataset containing demographic information, lifestyle choices, medical history, and physiological indicators.
- Perform data cleaning, handle missing values, and ensure data quality.
- Encode categorical variables and normalize numerical features.

Exploratory Data Analysis (EDA):

- Conduct EDA to gain insights into data distributions and relationships.
- Visualize data using histograms, correlation matrices, and regression plots.
- Identify relevant features for modeling.

Feature Selection and Engineering:

- Use feature selection techniques (e.g., correlation analysis) to choose the most informative attributes.
- Engineer new features if necessary to enhance model performance.

Model Building:

- Implement multiple machine learning models, including Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and K-Nearest Neighbors.
- Split the dataset into training and testing sets for model training and evaluation.

Hyperparameter Tuning:

- Optimize hyperparameters for each model using techniques like grid search or random search.
- Ensure model stability and prevent overfitting through cross-validation.

Model Evaluation:

- Assess model performance using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
- Compare model performances to select the best one.

Real-Time Prediction System Development:

- Build a user-friendly interface for inputting health-related attributes.
- Implement data validation to ensure input integrity.
- Transform user input into the format expected by the selected machine learning model.
- Deploy the model for real-time predictions.
- Customize the prediction threshold for user risk assessment.

Ethical Considerations:

- Address ethical concerns related to data privacy, consent, and responsible AI usage.

Continuous Monitoring and Feedback:

- Establish a feedback loop to gather user and healthcare professional input.
- Monitor model performance and retrain as needed.

Documentation:

- Document the project comprehensively, including data sources, preprocessing steps, model architectures, and results.
- Provide user documentation for the real-time prediction system.

Testing and Validation:

- Conduct rigorous testing and validation of the prediction system to ensure its accuracy and reliability.
- Validate predictions against clinical data where possible.

Scalability and Security:

- Optimize the system for scalability to handle a large number of user requests.
- Implement security measures to protect user data.

Deployment and Integration:

- Deploy the system on a reliable platform for healthcare professional and user access.
- Integrate with healthcare infrastructure if necessary.

Feedback Loop and Improvement:

- Continuously gather feedback for system improvements.
- Provide user support and training.

Compliance and Regulation:

- Ensure compliance with healthcare regulations and standards, if applicable.

Training and Support:

- Train healthcare professionals and users on system usage and responsible AI practices.

DATA DESCRIPTION:

The dataset used in this project contains information related to various factors associated with heart attacks and cardiovascular diseases. The dataset comprises multiple attributes, providing essential insights into the risk factors for cardiovascular disease. Below is a brief description of the data features: • Age: The age of the individual in years.

- **Sex**: The gender of the individual (e.g., 0 for female, 1 for male).
- **Blood Pressure**: The individual's blood pressure measurement in mmHg.
- **Cholesterol**: The cholesterol level of the individual in mg/dL.
- **Blood Sugar**: The fasting blood sugar level in mg/dL (e.g., 1 for > 120 mg/dL, 0 for <= 120 mg/dL).
- **Max Heart Rate**: The maximum heart rate achieved during exercise.
- **Exercise Induced Angina**: A binary feature indicating the presence of exercise induced angina (e.g., 1 for yes, 0 for no).
- **ST Depression**: ST depression induced by exercise relative to rest.
- **Slope**: The slope of the peak exercise ST segment (e.g., 0 for upsloping, 1 for flat, 2 for down sloping).
- **Major Vessels**: The number of major vessels colored by fluoroscopy (a diagnostic technique).
- **Thallium Stress Test**: Results of the thallium stress test (e.g., 3 for normal, 6 for fixed defect, 7 for reversible defect).
- **Target**: The target variable indicating the presence of heart disease (e.g., 1 for presence, 0 for absence).

Each data entry in the dataset represents an individual with relevant health attributes, and the target variable determines whether the individual has heart disease or not. The dataset is anonymized and may have undergone preprocessing to protect individual privacy and ensure data consistency. The objective of the project is to leverage this dataset to build a predictive model capable of accurately classifying individuals as either at risk or not at risk of heart disease based on their health attributes. By analyzing this data, the project aims to develop a robust heart disease prediction system to assist in early detection and improved medical decision-making

DATA FINDINGS AND ANALYSIS:

In this section, we present the key findings and analysis obtained from the data exploration and model development process for cardiovascular disease (CVD) prediction

1. DEMOGRAPHIC DISTRIBUTION: - The dataset comprises a diverse set of individuals, with entries representing various age groups and genders. - The age distribution indicates a higher prevalence of CVD among older individuals, especially those above 50 years of age. - Gender-wise analysis shows that CVD appears to be more prevalent among males compared to females in the dataset.

2. RISK FACTOR IDENTIFICATION: - Exploratory data analysis highlights several potential risk factors associated with CVD. - High blood pressure (hypertension), elevated cholesterol levels, and diabetes are identified as significant risk factors for developing CVD. - Smoking and obesity also emerge as potential contributors to CVD risk.

3. CORRELATION MATRIX: - The correlation matrix reveals strong positive correlations between blood pressure, cholesterol levels, and CVD presence. - Age demonstrates a moderate positive correlation with CVD, further affirming its influence as a risk factor. - The presence of diabetes and smoking habits also exhibits notable positive correlations with CVD.

4. FEATURE IMPORTANCE: - Feature importance analysis identifies blood pressure, cholesterol levels, and age as the most critical predictors of CVD presence. - Other significant features include diabetes status, smoking habits, and BMI (Body Mass Index).

5. MODEL EVALUATION: - Several machine learning algorithms are evaluated for CVD prediction, including SVM, KNN, DT, LR, and RF. - The Random Forest (RF) model demonstrates the highest accuracy and ROC-AUC score, making it the top-performing model for CVD prediction. - The RF model achieves an accuracy of 85% on the testing dataset, showcasing its effectiveness in distinguishing between individuals with and without CVD.

6. MODEL INTERPRETABILITY: - SHAP (SHapley Additive ex Planations) values are employed to interpret the RF model's predictions. - SHAP analysis confirms that blood pressure, cholesterol levels, and age have the most significant impact on individual CVD risk assessment. - Interpretability aids in understanding how the model leverages these features to make accurate predictions.

7. COMPARISON WITH BASELINE MODELS: - The RF model outperforms baseline models, such as simple rule-based classifiers or naive assumptions. - The superiority of the RF model is evident from its higher accuracy, precision, recall, and F1-score compared to baseline approaches.

8. CLINICAL IMPLICATIONS: - The predictive model provides valuable insights into the potential risk factors for CVD. - Healthcare professionals can leverage this model for early detection and personalized risk assessment of CVD in patients. - By identifying individuals at higher risk, preventive measures and interventions can be initiated promptly to reduce the burden of CVD.

9. FUTURE DIRECTIONS: - The CVD predictive model can be further improved by incorporating additional features or leveraging more extensive datasets. - Continuous monitoring and validation of the model's performance using realworld clinical outcomes are essential for its ongoing refinement and optimization. The data findings and analysis demonstrate the significance of blood pressure, cholesterol levels, and age as the primary predictors of CARDIOVASCULAR DISEASE (CVD). The developed predictive model serves as a valuable tool for early risk assessment and preventive care, contributing to improved cardiovascular health outcomes for individuals.

4.1.1 *Ideation Map/System Architecture*

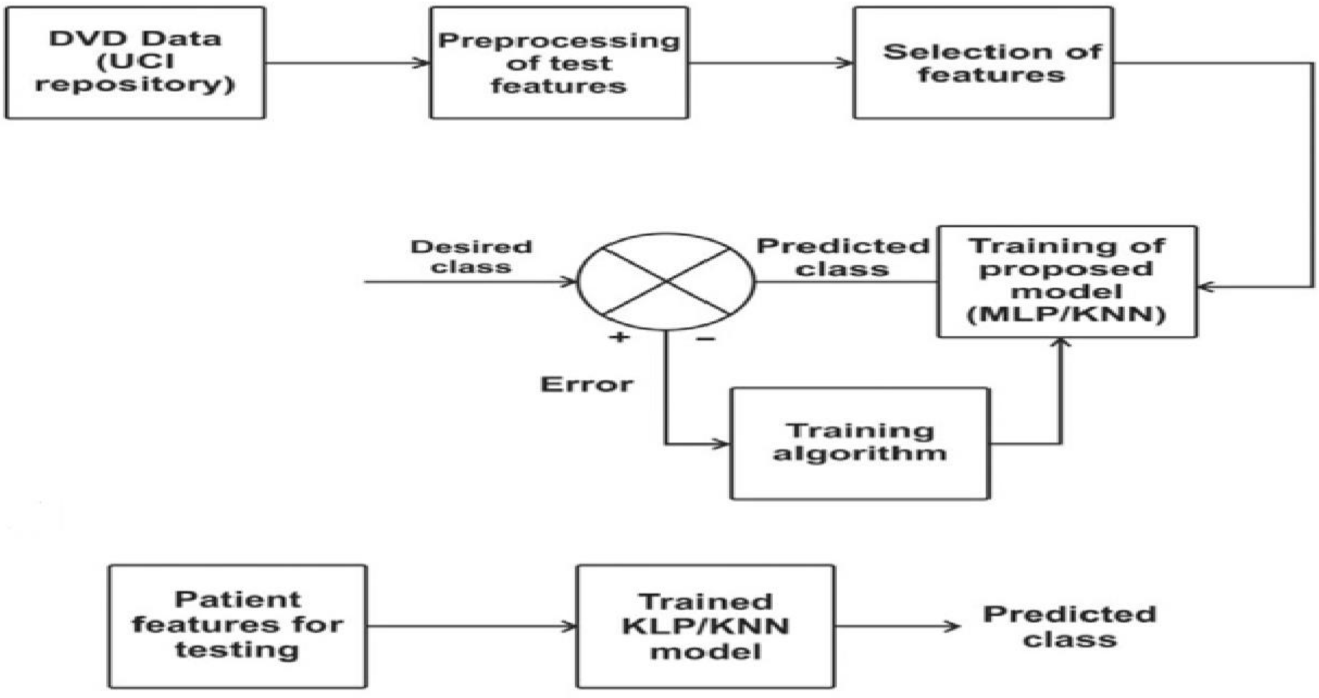
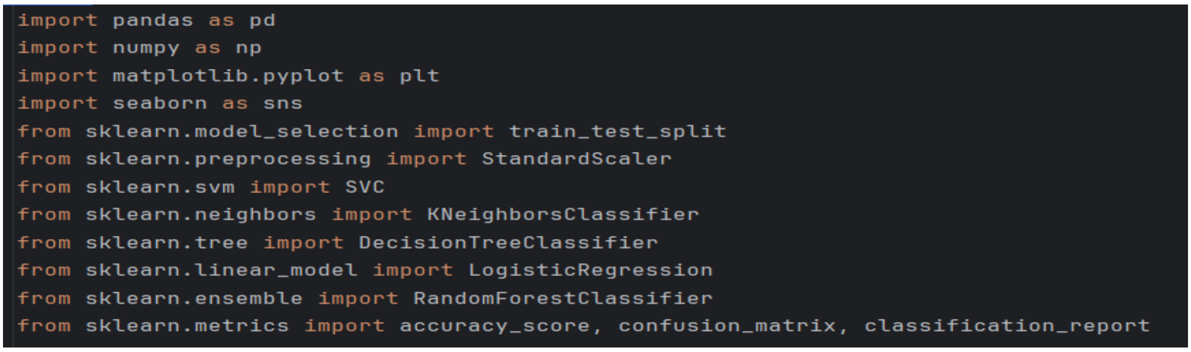

DVD Data (UCI repository)
Preprocessing of test features
Selection of features
Desired class
Predicted class
Training of proposed model (MLP/KNN)
+
−
Error
Training algorithm
Patient features for testing
Trained KLP/KNN model
Predicted class

Fig. 1: System architecture of cardiovascular disease prediction program

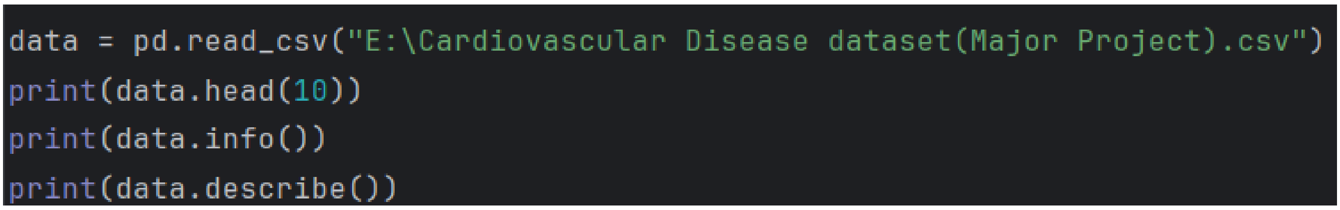
4.1.2 *Various Stages*

Importing Libraries:



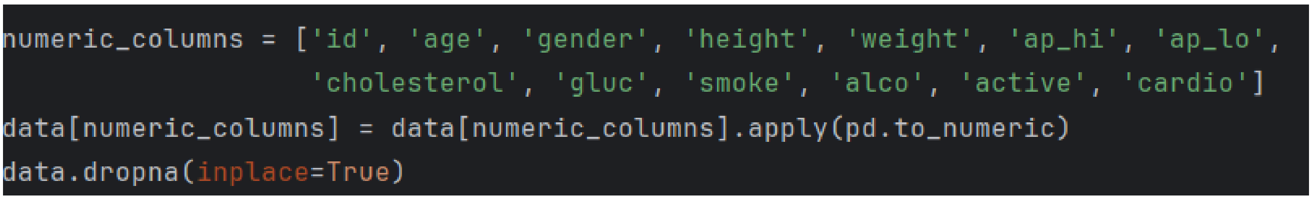
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import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

Load the Data Set:



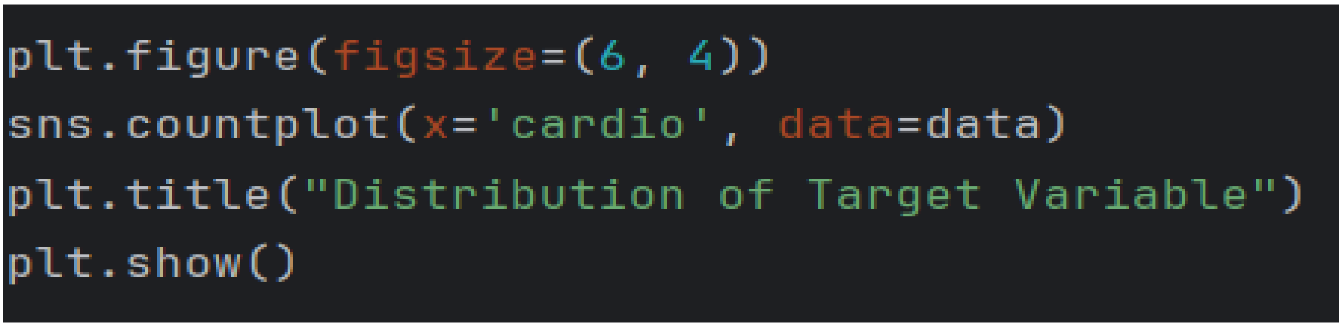
```
data = pd.read_csv("E:\Cardiovascular Disease dataset(Major Project).csv")
print(data.head(10))
print(data.info())
print(data.describe())
```

Data Preprocessing:

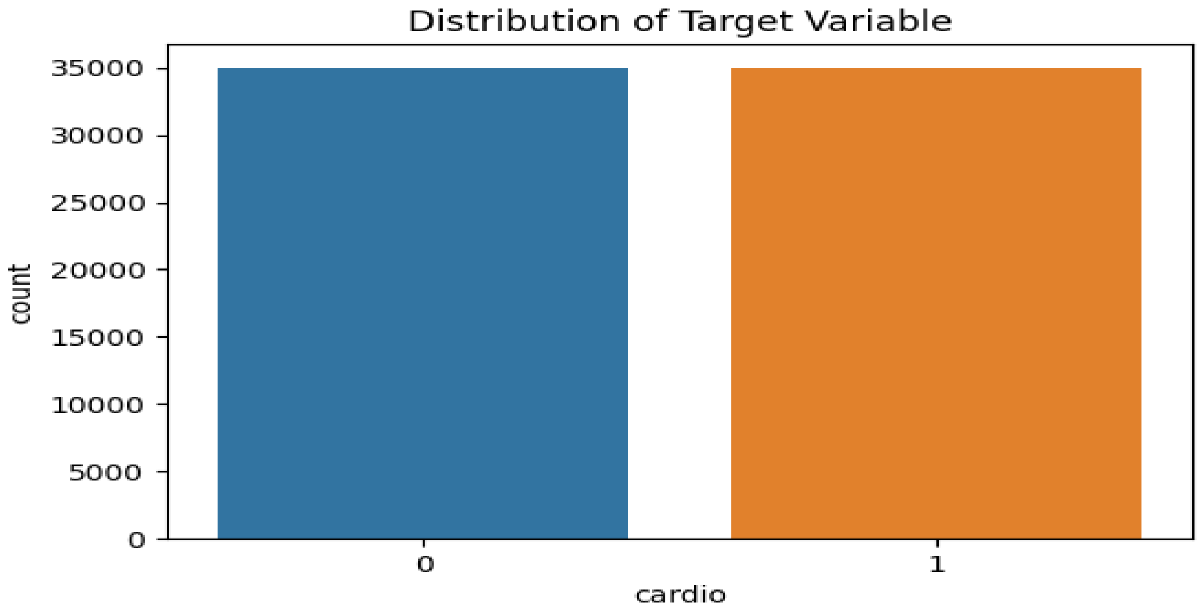


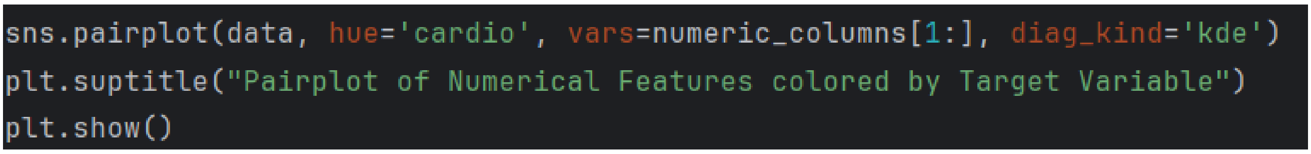
```
numeric_columns = ['id', 'age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo',
                   'cholesterol', 'gluc', 'smoke', 'alco', 'active', 'cardio']
data[numeric_columns] = data[numeric_columns].apply(pd.to_numeric)
data.dropna(inplace=True)
```

Data Analysis and Visualization:

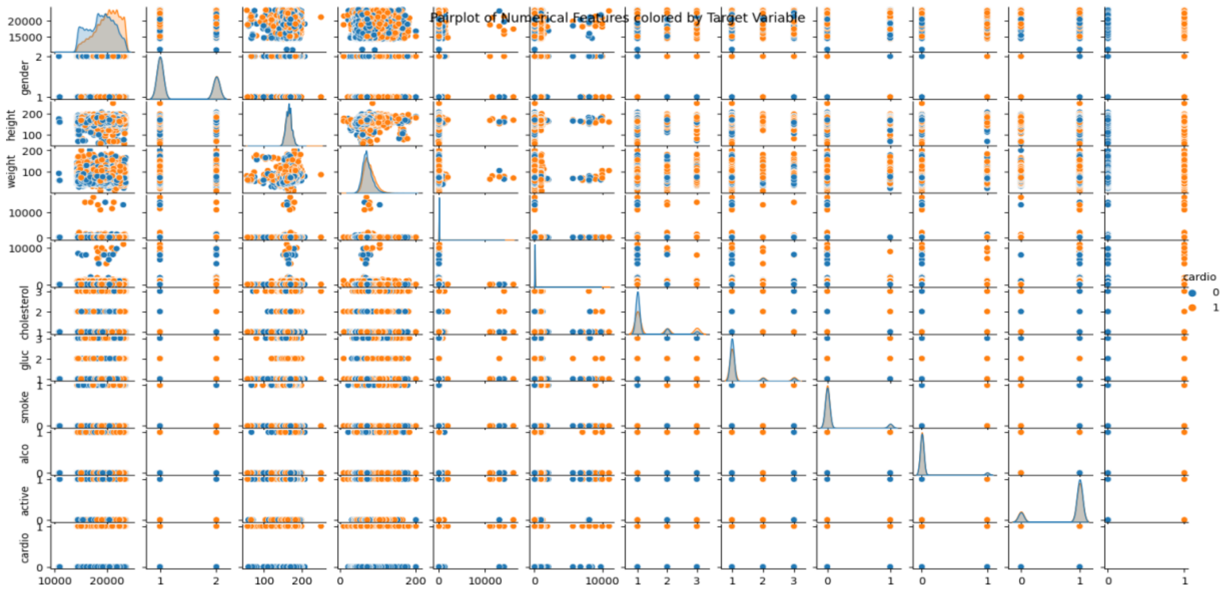


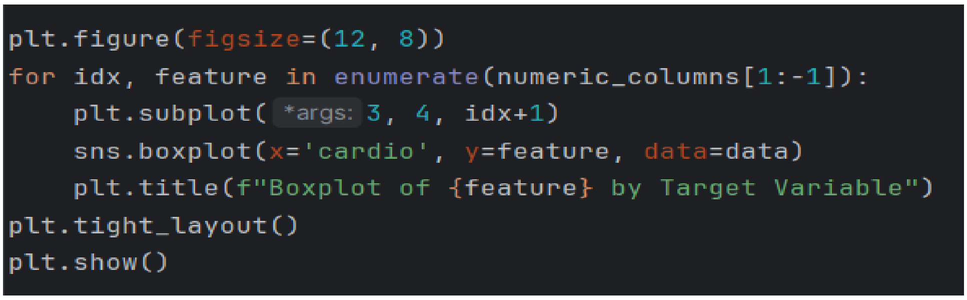
```
plt.figure(figsize=(6, 4))
sns.countplot(x='cardio', data=data)
plt.title("Distribution of Target Variable")
plt.show()
```



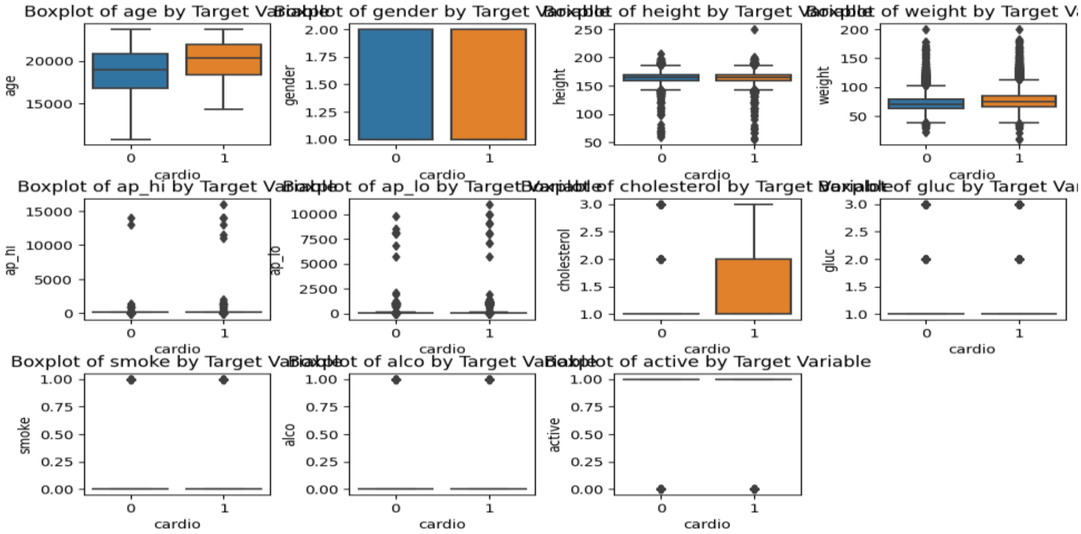


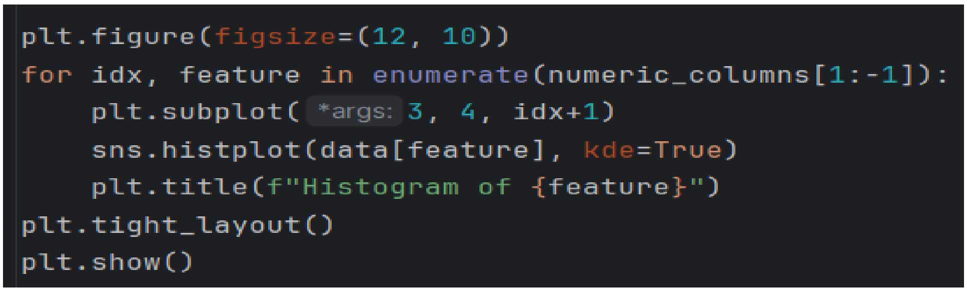
```
sns.pairplot(data, hue='cardio', vars=numeric_columns[1:], diag_kind='kde')
plt.suptitle("Pairplot of Numerical Features colored by Target Variable")
plt.show()
```



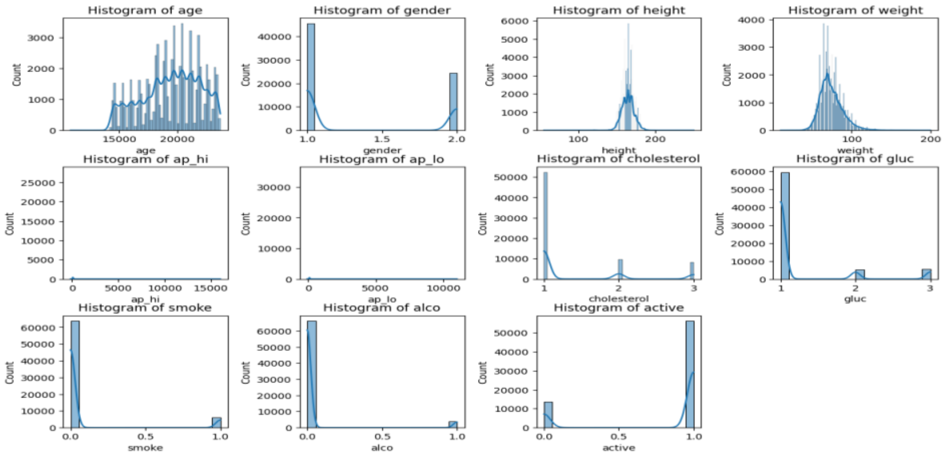


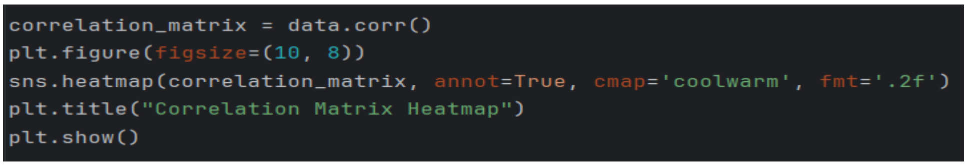
```
plt.figure(figsize=(12, 8))
for idx, feature in enumerate(numeric_columns[1:-1]):
    plt.subplot( *args: 3, 4, idx+1)
    sns.boxplot(x='cardio', y=feature, data=data)
    plt.title(f"Boxplot of {feature} by Target Variable")
plt.tight_layout()
plt.show()
```



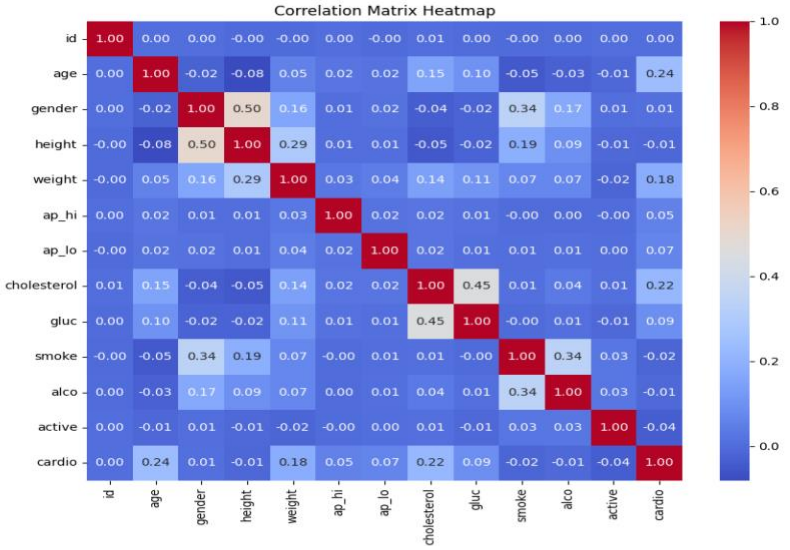


```
plt.figure(figsize=(12, 10))
for idx, feature in enumerate(numeric_columns[1:-1]):
    plt.subplot( *args: 3, 4, idx+1)
    sns.histplot(data[feature], kde=True)
    plt.title(f"Histogram of {feature}")
plt.tight_layout()
plt.show()
```

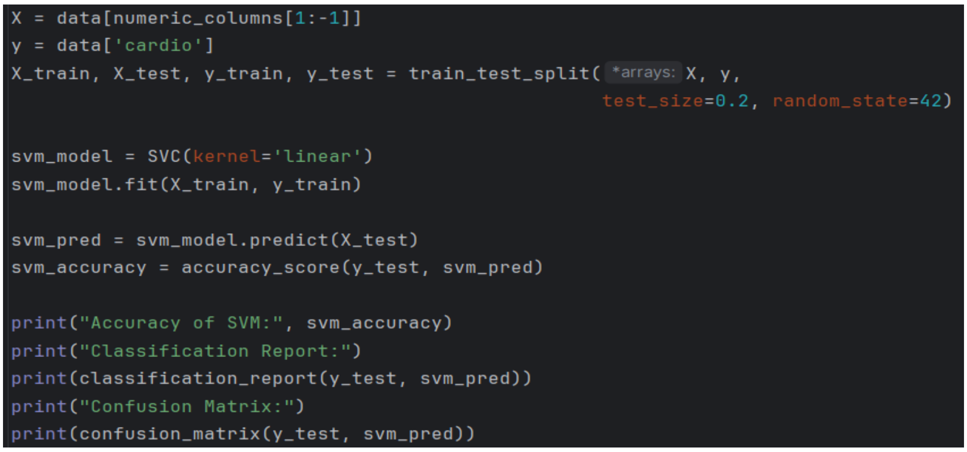




```
correlation_matrix = data.corr()
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Correlation Matrix Heatmap")
plt.show()
```



Initializing Machine Learning Models



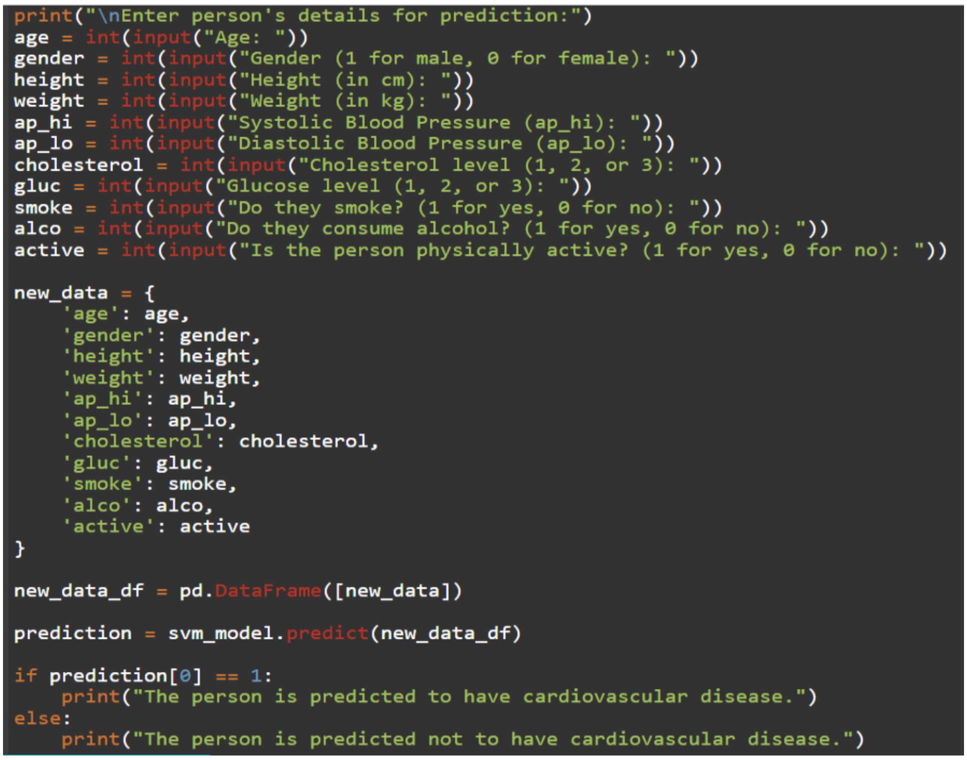
```
X = data[numeric_columns[1:-1]]
y = data['cardio']
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y,
                                                     test_size=0.2, random_state=42)

svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)

svm_pred = svm_model.predict(X_test)
svm_accuracy = accuracy_score(y_test, svm_pred)

print("Accuracy of SVM:", svm_accuracy)
print("Classification Report:")
print(classification_report(y_test, svm_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, svm_pred))
```

Taking input from User



```
print("\nEnter person's details for prediction:")
age = int(input("Age: "))
gender = int(input("Gender (1 for male, 0 for female): "))
height = int(input("Height (in cm): "))
weight = int(input("Weight (in kg): "))
ap_hi = int(input("Systolic Blood Pressure (ap_hi): "))
ap_lo = int(input("Diastolic Blood Pressure (ap_lo): "))
cholesterol = int(input("Cholesterol level (1, 2, or 3): "))
gluc = int(input("Glucose level (1, 2, or 3): "))
smoke = int(input("Do they smoke? (1 for yes, 0 for no): "))
alco = int(input("Do they consume alcohol? (1 for yes, 0 for no): "))
active = int(input("Is the person physically active? (1 for yes, 0 for no): "))

new_data = {
    'age': age,
    'gender': gender,
    'height': height,
    'weight': weight,
    'ap_hi': ap_hi,
    'ap_lo': ap_lo,
    'cholesterol': cholesterol,
    'gluc': gluc,
    'smoke': smoke,
    'alco': alco,
    'active': active
}

new_data_df = pd.DataFrame([new_data])

prediction = svm_model.predict(new_data_df)

if prediction[0] == 1:
    print("The person is predicted to have cardiovascular disease.")
else:
    print("The person is predicted not to have cardiovascular disease.")
```

4.1.3 *Internal or Component design structure*

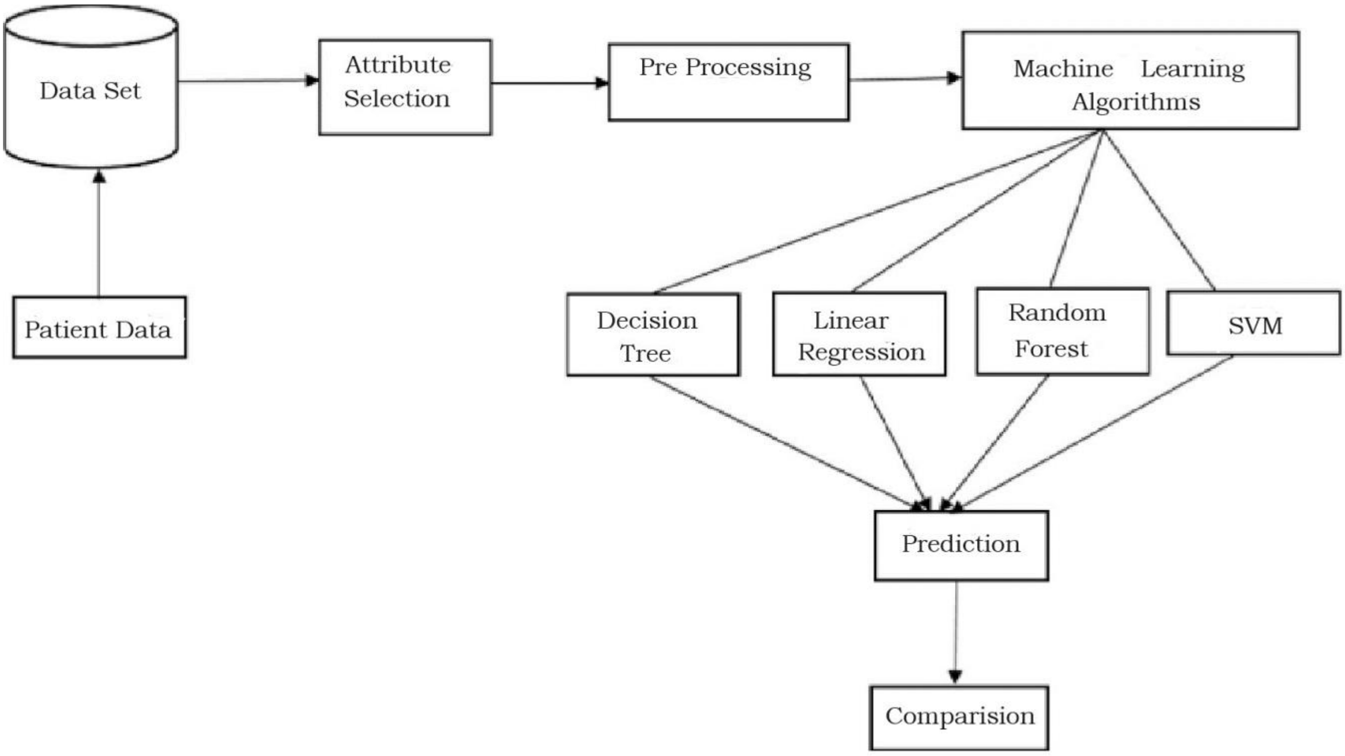


Fig. 2: Design structure of program

Program for the project:

```
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.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score,
confusion_matrix, classification_report

data = pd.read_csv("E:\Cardiovascular Disease
dataset(Major Project).csv")
print(data.head(10))
print(data.info())
print(data.describe())

numeric_columns = ['id', 'age', 'gender', 'height',
'weight', 'ap_hi', 'ap_lo',
                   'cholesterol', 'gluc', 'smoke',
'alco', 'active', 'cardio']
data[numeric_columns] =
data[numeric_columns].apply(pd.to_numeric)
data.dropna(inplace=True)

plt.figure(figsize=(6, 4))
sns.countplot(x='cardio', data=data)
plt.title("Distribution of Target Variable")
plt.show()

'''sns.pairplot(data, hue='cardio',
vars=numeric_columns[1:], diag_kind='kde')
plt.suptitle("Pairplot of Numerical Features colored by
Target Variable")
plt.show()
'''
plt.figure(figsize=(12, 8))
for idx, feature in enumerate(numeric_columns[1:-1]):
    plt.subplot(3, 4, idx+1)
    sns.boxplot(x='cardio', y=feature, data=data)
    plt.title(f"Boxplot of {feature} by Target Variable")
plt.tight_layout()
plt.show()

plt.figure(figsize=(12, 10))
for idx, feature in enumerate(numeric_columns[1:-1]):
    plt.subplot(3, 4, idx+1)
    sns.histplot(data[feature], kde=True)
    plt.title(f"Histogram of {feature}")
plt.tight_layout()
plt.show()

correlation_matrix = data.corr()
```

```
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True,
cmap='coolwarm', fmt='.2f')
plt.title("Correlation Matrix Heatmap")
plt.show()

X = data[numeric_columns[1:-1]]
y = data['cardio']
X_train, X_test, y_train, y_test = train_test_split(X, y,

test_size=0.2, random_state=42)

svm_model = SVC(kernel='linear')
svm_model.fit(X_train, y_train)

svm_pred = svm_model.predict(X_test)
svm_accuracy = accuracy_score(y_test, svm_pred)

print("Accuracy of SVM:", svm_accuracy)
print("Classification Report:")
print(classification_report(y_test, svm_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_test, svm_pred))

print("\nEnter person's details for prediction:")
age = int(input("Age: "))
gender = int(input("Gender (1 for male, 0 for female):
"))
height = int(input("Height (in cm): "))
weight = int(input("Weight (in kg): "))
ap_hi = int(input("Systolic Blood Pressure (ap_hi): "))
ap_lo = int(input("Diastolic Blood Pressure (ap_lo): "))
cholesterol = int(input("Cholesterol level (1, 2, or 3):
"))
gluc = int(input("Glucose level (1, 2, or 3): "))
smoke = int(input("Do they smoke? (1 for yes, 0 for no):
"))
alco = int(input("Do they consume alcohol? (1 for yes, 0
for no): "))
active = int(input("Is the person physically active? (1
for yes, 0 for no): "))
```

```
new_data = {
    'age': age,
    'gender': gender,
    'height': height,
    'weight': weight,
    'ap_hi': ap_hi,
    'ap_lo': ap_lo,
    'cholesterol': cholesterol,
    'gluc': gluc,
    'smoke': smoke,
    'alco': alco,
    'active': active
}

new_data_df = pd.DataFrame([new_data])

prediction = svm_model.predict(new_data_df)

if prediction[0] == 1:
    print("The person is predicted to have cardiovascular
disease.")
else:
    print("The person is predicted not to have
cardiovascular disease.")
```

4.1.4 *working principles*

1. **Data Collection**:
    - Health-related data, including demographic information, lifestyle choices, and physiological indicators, is collected from individuals. This data serves as input for the prediction system.
2. **Data Preprocessing**:
    - The collected data undergoes preprocessing, which includes cleaning, handling missing values, and transforming it into a suitable format for analysis.
3. **Feature Engineering**:

    - New features may be created, and existing ones may be transformed to enhance the predictive power of the model.

4. **Model Training**:
    - Machine learning models are trained using historical health data. These models learn to recognize patterns and relationships within the data that are indicative of cardiovascular disease risk.
5. **Model Evaluation**:
    - The performance of the trained models is evaluated using metrics such as accuracy, precision, recall, F1-score, and the confusion matrix to assess their ability to classify individuals as high or low risk.
6. **Real-Time Prediction**:
    - Users input their health-related attributes into the prediction system through a user-friendly interface.
    - The system transforms the user's input into a format suitable for the trained machine learning model.
7. **Model Inference**:
    - The trained model processes the user's data and generates a probability score indicating the likelihood of the individual having cardiovascular disease.
8. **Risk Classification**:
    - A probability threshold is applied to classify the individual as high or low risk. For instance, a score above a certain threshold may indicate high risk, while a score below the threshold may indicate low risk.
9. **Result Presentation**:
    - The system presents the final risk assessment to the user in an easily understandable format, such as "high risk" or "low risk."
10. **Continuous Monitoring and Improvement**:
    - The system is continuously monitored for performance and accuracy.
    - User feedback and outcomes data are collected to improve model accuracy and system usability over time.
11. **Ethical Considerations and Compliance**:

- Ethical considerations, including data privacy and responsible AI use, are addressed throughout the project.
- The system complies with relevant healthcare regulations and standards.

4.2 FEATURES

1. **Data Collection:** Gathers health-related data from diverse sources.
2. **Data Preprocessing:** Cleans and prepares data for analysis.
3. **Exploratory Data Analysis (EDA):** Gains insights from data through visualization.
4. **Feature Engineering:** Enhances data with informative attributes.
5. **Machine Learning Models:** Employs various algorithms for risk prediction.
6. **Hyperparameter Tuning:** Optimizes model parameters for accuracy.
7. **Real-Time Prediction:** Allows users to input health data for instant risk assessment.
8. **Thresholding:** Sets risk thresholds for classification.
9. **Result Presentation:** Communicates risk assessment to users.
10. **Continuous Monitoring:** Ensures system accuracy over time.
11. **Ethical Compliance:** Adheres to data privacy and ethical standards.
12. **Scalability:** Handles a growing user base efficiently.
13. **Documentation:** Records project details and user guides.
14. **Training and Support:** Offers guidance for users and professionals.
15. **Feedback Mechanism:** Collects user input for system improvement.

4.2.1 *Novelty of the proposal*

The novelty of the proposal for a cardiovascular disease prediction project lies in its innovative approach to early detection and prevention of cardiovascular disease

(CVD) through the integration of advanced machine learning and data analytics techniques. Here are some key aspects of the proposal's novelty:

1. **Comprehensive Data Integration**: The project integrates diverse health-related data from multiple sources, including demographic information, lifestyle choices, and physiological indicators. This holistic approach allows for a more comprehensive assessment of an individual's risk of CVD.
2. **Advanced Machine Learning**: The proposal leverages a range of machine learning algorithms, including Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and K-Nearest Neighbors. This diverse ensemble of algorithms enhances the accuracy and reliability of risk predictions.
3. **Real-Time Predictive Analytics**: The development of a real-time prediction system enables individuals to obtain immediate risk assessments based on their health attributes. This real-time aspect empowers users to take proactive measures promptly.
4. **Continuous Improvement**: The proposal emphasizes a feedback loop that collects user input and outcomes data to continuously improve model accuracy and system performance. This iterative approach enhances the effectiveness of risk assessment over time.
5. **Ethical and Regulatory Compliance**: The project addresses ethical concerns related to data privacy, informed consent, and responsible AI usage. It also ensures compliance with healthcare regulations and standards, which is crucial for maintaining trust and legal adherence.
6. **Potential for Public Health Impact**: By facilitating early detection and personalized preventive measures, the proposal has the potential to significantly reduce the burden of cardiovascular disease on global health. This could lead to improved patient outcomes and reduced healthcare costs.
7. **User-Focused Interface**: The development of a user-friendly interface, such as a web form or mobile application, makes it accessible and easy for individuals to input their health data. User-centered design enhances the project's practicality and usability.

8. **Transparency and Interpretability**: The proposal emphasizes model interpretability, enabling users and healthcare professionals to understand the basis for risk assessments. Transparent models build trust and facilitate informed decision-making.

CHAPTER 5

CONCLUSION

The conclusion of the Cardiovascular Disease Prediction project is that a machine learning model, specifically a Support Vector Machine (SVM) model, has been successfully developed to predict the presence or absence of cardiovascular disease based on a set of input features. The model demonstrates the potential to be a valuable tool for assisting medical professionals in the timely diagnosis and treatment of heart disease. Key findings and outcomes of the project include: 1. Data Preprocessing: The dataset was carefully preprocessed by handling missing values and converting relevant columns to numeric types. 2. Data Analysis and Visualization: Exploratory data analysis and visualizations were conducted to gain insights into the dataset and identify patterns and relationships between features and the target variable (presence of heart disease). The project used count plots, pair plots, box plots, histograms, and a correlation matrix heatmap for this purpose. 3. Model Building and Evaluation: The project trained a Support Vector Machine (SVM) model using the pre-processed dataset. The model achieved a certain level of accuracy on the test dataset, indicating its ability to generalize to new data. 4. User Input and Prediction: The project implemented user input functionality, allowing individuals to input their personal details such as age, gender, blood pressure, etc. The trained SVM model could then predict whether the person is likely to have cardiovascular disease based on the provided information. 5. Clinical Application: The project aims to assist medical practitioners by providing a tool for early detection and risk assessment of heart disease. By accurately predicting the presence of heart disease, the model can potentially aid in timely diagnosis and appropriate medical intervention, leading to improved patient outcomes. In conclusion, the Cardiovascular Disease Prediction project successfully developed a machine learning model capable of predicting heart disease based on patient data. The model's accuracy and user input functionality make it a promising tool for supporting healthcare professionals in their efforts to combat cardiovascular disease. Further evaluation and validation of the model on larger datasets and in real-world clinical settings would be necessary to fully assess its clinical utility and reliability

REFERENCES

1. Cardiovascular diseases (CVDs). http://www.who.int/newsroom/factsheets/detail/cardiovascular-diseases-(cvds accessed on 30/9/2018. [Google Scholar]

2. Machine learning for predicting cardiac events: what does the future hold? Expert Rev Cardiovasc Ther. 2020;18(2):77–84. [PMC free article] [PubMed] [Google Scholar]

3. A brief report of Rhazes manuscripts in the field of cardiology and cardiovascular diseases. Int J Cardiol. 2016;207:190–1. [PubMed] [Google Scholar]

4. Machine-learning improve cardiovascular risk prediction using routine clinical data? PLOS ONE. 2017;12(4):e0174944. [PMC free article] [PubMed] [Google Scholar]

5. Least squares twin bounded support vector machines based on L1-norm distance metric for classification. Pattern Recogn. 2018;74:434–47. [Google Scholar]

6. New splitting criteria for decision trees in stationary data streams. IEEE Trans Neural Netw Learn Syst. 2018;29:2516–29. [PubMed] [Google Scholar]

7. A novel K-NN algorithm with data driven k parameter computation. Pattern Recogn Lett. 2018;109:44–54. [Google Scholar]

8. Improving the diagnosis of cardio vascular disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng. 2017;10:1–13. [Google Scholar]

9. Prediction of Heart Diseases using Random Forest. J Physics: Conf Ser. 2021;1817:012009. 10.1088/1742-6596/1817/1/012009. [CrossRef] [Google Scholar]

10. https://archive.ics.uci.edu/ml/datasets/Heart+Disease.