Risk Prediction of Cardiovascular Disease

Course: BCA 2nd Year
Semester: IV
Section: E
Team: 4ET1
Program: Machine Learning
Project: Risk Prediction of Cardiovascular Disease
DCS Mentor: Nazmul Arefin

Sr. No.	Student ID	Roll No.	Student Name
1	BC2023194	2310201620	Vipin Mishra (Team Lead)
2	BC2023714	2310201679	Shivam Verma
3	BC2023086	2310201139	Arsala Mazhar
4	BC2023337	2310201114	Anshul Chauhan
5	BC2023078	2310201304	Minhaj Khan
6	BC2023273	2310201528	Shivam Diwakari
7	BC2023244	2310201267	Khushi Pandey
8	BC2023242	2310201265	Nikhil Pandey
9	BC2023184	2310201485	Sachin Singh
10	BC2023087	2310201474	Ramesh Kashyap
11	BC2023612	2310201183	Danish Khan
12	BC2023349	2310201164	Ayush Kumar
13	BC2023225	2310201545	Shreya Rathore
14	BC2023678	2310303160	Krishna Shrivastav
15	BC2023159	2310201374	Nawazish Ali Khan

Risk Prediction of Cardiovascular Disease using Machine Learning:

Table of Contents

- Introduction
- Role of IIoT in Cardiovascular Analytics
- Libraries and Setup
- Importing the Dataset
- Handling Null Values
- Data Cleaning
- Distribution Plots
- Data Visualization (Plots)
- Skewness
- Exploratory Data Analysis (EDA)
- Encoders
- Heat Map
- Standard Scaler
- Algorithms
- **Conclusion**
- **Future Scope**

Introduction

Cardiovascular diseases (CVDs) are the number one cause of death globally. The rise of data-driven technologies has revolutionized healthcare, particularly in cardiovascular medicine. **Cardiovascular analytics** involves applying advanced data analysis, machine learning, and artificial intelligence (AI) techniques to extract actionable insights from health-related data such as:

- Electronic Health Records (EHRs)
- Medical imaging (e.g., ECG, MRI)
- Wearable health monitors
- Genetic profiles

Key Objectives:

- Early detection of heart disease
- Risk stratification and prediction
- Personalized treatment plans
- Improved clinical decision-making

Role of IIoT in Cardiovascular Analytics

Industrial Internet of Things (IIoT) enables the collection and transmission of real-time health data from connected medical devices. Its role in cardiovascular analytics is critical:

1. **Real-Time Monitoring:** Devices like smartwatches continuously track heart rate, oxygen levels, and arrhythmias.
2. **Remote Patient Management:** Enables healthcare providers to monitor patients post-discharge or those in remote areas.
3. **Predictive Analytics:** Continuous data can help forecast potential cardiovascular events.
4. **Personalized Treatment:** Tailoring interventions based on real-time data.
5. **Improved Decision Support:** AI-driven insights support quicker and more accurate decisions.
6. **Reduced Hospital Readmissions:** Early warnings reduce emergency admissions.
7. **Enhanced Research:** IIoT fuels large-scale longitudinal studies and clinical trials.

Libraries and Setup

Data Analysis & Manipulation

- pandas: Data handling and manipulation
- numpy: Numerical computing

Visualization

- matplotlib.pyplot: General plotting
- seaborn: Advanced statistical visualizations

Machine Learning (scikit-learn)

- Classification Algorithms: LogisticRegression, KNeighborsClassifier, DecisionTreeClassifier, RandomForestClassifier
- Utilities: train_test_split, StandardScaler, LabelEncoder, OrdinalEncoder
- Evaluation: accuracy_score, classification_report, confusion_matrix

Statistical Tools

- Variance Inflation Factor (VIF) from statsmodels for multicollinearity detection

Importing the Dataset

Best Practices:

- Ensure data is in .csv, .xlsx, or database format
- Understand data structure, types, and meaning of each feature
- Conduct preliminary checks for consistency and completeness

Challenges:

1. **Data inconsistencies:** Date format mismatches, broken rows

2. **Missing values:** Often due to device errors or non-response
3. **Duplicate entries:** Common in merged medical datasets

Null Values

Null values (NaNs) represent missing or unrecorded data.

Handling Techniques:

- Drop rows or columns with nulls
- Fill missing values using:
  - Mean, median, or mode
  - Forward fill / backward fill
  - Model-based imputation (KNN imputer, etc.)

Data Cleaning

Data cleaning improves model accuracy and robustness.

Key Cleaning Steps:

1. Handle nulls and duplicates
2. Fix incorrect data types (e.g., date strings)
3. Rename columns for clarity
4. Remove irrelevant or redundant features
5. Standardize text (e.g., upper/lowercase)
6. Detect and treat outliers
7. Normalize or scale numerical data
8. Encode categorical data
9. Final quality check

IIoT-Specific Data Cleaning Issues:

- Time drift or unsynchronized timestamps
- Mixed units (e.g., BPM vs. Hz)
- Data spikes due to noise

Distribution Plots

Used to understand the shape of a variable.

- **Histogram:** Frequency of value ranges
- **KDE (Kernel Density Estimate):** Smooth distribution curve
- **Use:** Identify skewness, detect outliers, and assess normalization needs

Plots and Visualization

1. Scatter Plot:

Shows correlation between two numerical features.

2. Box Plot:

Visualizes spread and outliers in a single feature.

3. Bar Plot:

Useful for categorical comparisons.

4. Line Plot:

Ideal for time-series data such as continuous heart monitoring.

Skewness

Skewness indicates how asymmetric a distribution is:

- **Positive Skew:** Long tail to the right (e.g., age, income)
- **Negative Skew:** Long tail to the left
- **Importance:** Highly skewed data may need transformation (e.g., log, Box-Cox)

Exploratory Data Analysis (EDA)

1. Descriptive Analysis

- Central tendency: Mean, median, mode
- Spread: Range, IQR, standard deviation

2. Univariate Analysis

- Focus on one variable
- Helps in understanding individual features

3. Bivariate & Multivariate Analysis

- Relationships between features (correlation, interaction effects)
- Essential for feature engineering

4. Predictive Analysis

- Uses labeled data to build classification or regression models

Encoders

Machine learning models require numeric input.

1. Label Encoder

- Assigns numeric values to each category
- Best for target variables or binary features

2. Ordinal Encoder

- Maintains order among categories
- Useful for ranked features (e.g., low, medium, high)

Heat Map

A heat map uses color to represent correlation strength between features.

- **+1:** Perfect positive correlation
- **-1:** Perfect negative correlation
- **0:** No correlation

Use it to:

- Detect multicollinearity
- Reduce dimensionality by removing redundant features

Standard Scaler

StandardScaler normalizes features by removing the mean and scaling to unit variance.

- Ideal for models sensitive to feature scaling (e.g., KNN, SVM)
- Makes training faster and more stable

Algorithms

1. Logistic Regression

- Suitable for binary outcomes (e.g., heart disease: Yes/No)
- Provides probabilistic predictions and interpretable coefficients

2. K-Nearest Neighbors (KNN)

- Instance-based learner
- Sensitive to feature scaling
- Works well with smaller, balanced datasets

3. Decision Tree Classifier

- Tree-like decision-making
- Interpretable and handles both numerical and categorical data

4. Random Forest Classifier

- Ensemble of decision trees

- Robust to overfitting
- Excellent for complex cardiovascular datasets with nonlinear relationships

Conclusion

Cardiovascular analytics powered by machine learning is revolutionizing how we detect, understand, and treat heart disease. With proper data preprocessing, insightful EDA, and the right algorithms, we can build systems that:

- Predict heart disease with high accuracy
- Provide real-time monitoring and alerts
- Enable better, faster, and more personalized clinical decisions

Machine learning not only improves diagnostics but also supports preventive healthcare, helping to reduce the burden of cardiovascular diseases on society.

Future Scope

1. **Deep Learning & Neural Networks:** Can analyze ECG images and time-series data for more accurate predictions.
2. **Wearable Tech Integration:** Real-time data from smartwatches and fitness trackers can enable 24/7 health analytics.
3. **Federated Learning:** Enables privacy-preserving training across multiple hospitals without sharing patient data.
4. **Explainable AI (XAI):** To build trust in ML models by making them more interpretable for doctors.
5. **Genomics + ML:** Using genetic data to identify individuals at high risk of hereditary heart conditions.
6. **Edge AI for IoT:** Performing real-time analytics directly on wearable devices without cloud dependency.
7. **Automated Diagnosis Systems:** AI-powered diagnostic assistants for rural and under-resourced regions.