PRIOR IDENTIFICATION AND PREDICTION OF CARDIOVASCULAR EXPOSURE BY UTILIZING SUPERVISED MACHINE LEARNING MODEL USING K-NEAREST NEIGHBORS CLASSIFIER ALGORITHM

Overview:

According to World Health Organization, numerous people are being suffered due to cardiovascular disease, making one of the major causes of death. Early physiological markers frequently provide modest warnings of deteriorating cardiovascular health. Yet, these signals are frequently ignored due to the pressures of modern living. Prolonged neglect of physical well-being can impede daily functioning and lead to the development of significant illnesses such as arrhythmia, atherosclerosis, heart failure, and many more. In this context, the current work uses a cardiovascular dataset to raise awareness and aid in early risk prediction. A K-Nearest Neighbor classifier model was utilized to examine an individual’s chances of acquiring cardiovascular disease.

Problem Statement:

People in today’s culture frequently overlook long term health monitoring because they are burdened with numerous personal and professional obligations that are limited by tight deadlines and timetables. Due to lifestyle pressures and a lack of knowledge about preventive healthcare practices, basic physiological indicators such as blood pressure, stress related reactions, blood glucose level, cholesterol are often ignored despite being accurate early predictors of cardiovascular risk. As a result, many people are still ignorant of their possible risk of cardiovascular disease within next ten years, which leads to a inadequate risk reduction tactics and delayed therapeutic intervention. The creation of a quick, easily accessible, and user friendly predictive model using K-Nearest Neighbor algorithm that can evaluate cardiovascular risk in a brief amount of time could close this gap by enabling early detection, encouraging well-informed health choices.

Main Objective(s):

- To gather and examine each person's basic clinical parameters such as age, bp, heart rate, blood glucose level, smoking.
- Using the retrieved clinical parameters, train the K-Nearest Neighbor classification (KNN-C) model.
- To estimate a person's possibility towards cardiovascular defect exposure and their level.

DATASET DESCRIPTION:

A. **Source of data:** Kaggle: https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea?select=train.csv

B. **Structure of data:**

1. **Rows:** 3390 records.
2. **Columns:** 7 features (Age (in Years), Gender (Male – M, Female – F), Smoker (Yes / No), Cigarettes per day (stick count), Upper Blood Pressure (in mm/Hg), Lower Blood Pressure (in mm/Hg), Heart Rate (in beats per minute), TenYearCHD).

C. **Data columns:**

COLUMN NAME	DATATYPE	COLUMN DESCRIPTION
age	Integer	Age of the individual (in Years).
sex	String	Gender of the individual. If male indicate "M", female indicate "F".
Is_smoking	String	Information of the individual about smoking habit. Indicate as "Yes" if individual smoke, else "No".
cigsPerDay	Integer	Number of cigarette sticks the individual would smoke (stick count).
sysBP	Integer	The top/first/upper value obtained from the blood pressure machine (systolic, in mmHg).
diaBP	Integer	The bottom/second/lower value obtained from the blood pressure machine. (diastolic, in mmHg).
heartRate	Integer	Number of heart beats within 60 seconds. Obtained through ecg machine or manual wrist counting. (in bpm).
TenYearCHD	String	Individual's exposure towards cardiovascular disease. (0 - indicates may not be exposed, 1 - indicates to be exposed).

D. Key Variables:

1 **X – Independent variables:** age, sex, is_smoking, cigsPerday, sysBP, diaBP, heartRate.

2 **Y – Dependent variables:** TenYearCHD.

Y DATA PROFILE:

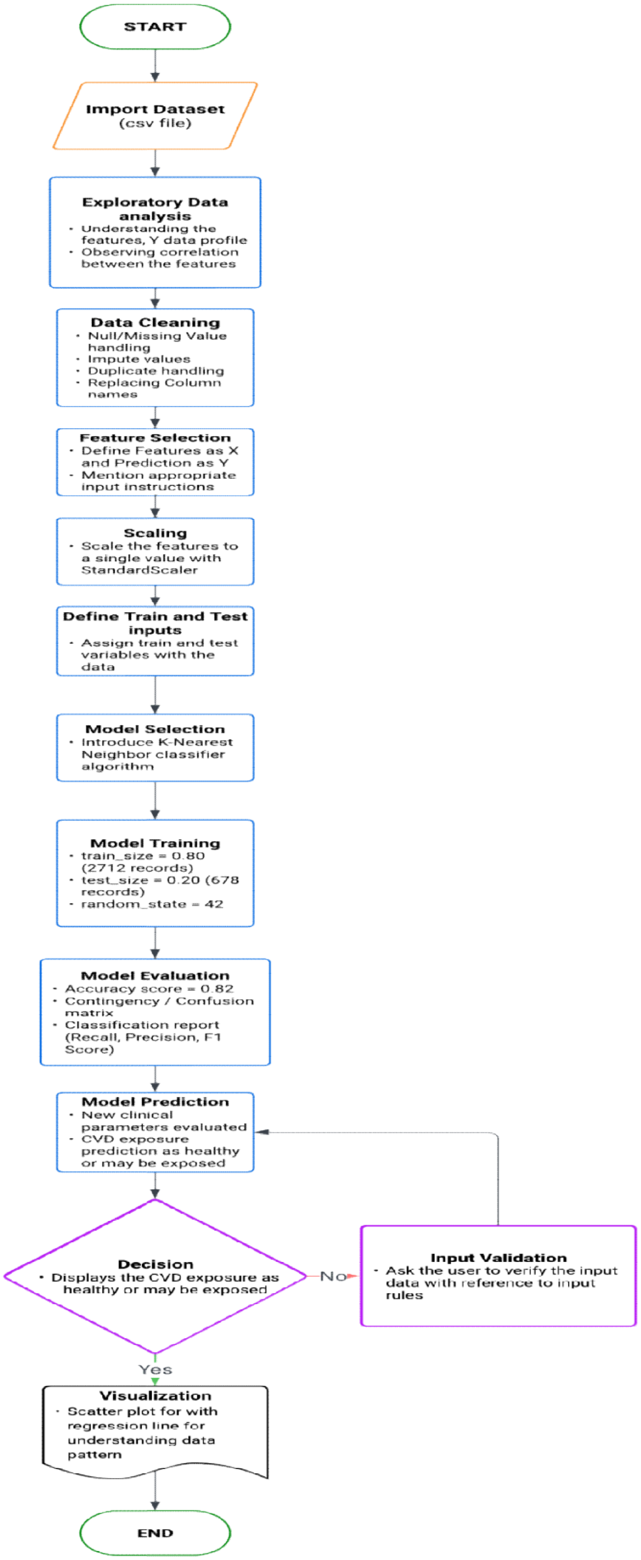
- Profile data:

Cardiovascular-Study-train-yprofile.html

- Statistical description of the data

FEATURES	COUNT	MEAN	MEDIAN	MODE	STD	MIN	25%	50%	75%	MAX
age	3390	49.5422	49.0000	40.0000	8.5929	32.0000	42.0000	49.0000	56.0000	70.0000
education	3303	1.9709	2.0000	1.0000	1.0191	1.0000	1.0000	2.0000	3.0000	4.0000
Gender	3390	0.4327	0.0000	0.0000	0.4955	0.0000	0.0000	0.0000	1.0000	1.0000
Smoker	3390	0.4976	0.0000	0.0000	0.5001	0.0000	0.0000	0.0000	1.0000	1.0000
Cigarette(s) per day	3390	9.1297	0.0000	0.0000	11.8639	0.0000	0.0000	0.0000	20.0000	70.0000
BPMeds	3346	0.0299	0.0000	0.0000	0.1703	0.0000	0.0000	0.0000	0.0000	1.0000
prevalentStroke	3390	0.0065	0.0000	0.0000	0.0803	0.0000	0.0000	0.0000	0.0000	1.0000
prevalentHyp	3390	0.3153	0.0000	0.0000	0.4647	0.0000	0.0000	0.0000	1.0000	1.0000
diabetes	3390	0.0257	0.0000	0.0000	0.1582	0.0000	0.0000	0.0000	0.0000	1.0000
Recent cholestrol level	3390	237.0743	235.0000	240.0000	44.9930	107.0000	206.0000	235.0000	264.0000	696.0000
Upper BP	3390	132.6012	128.5000	110.0000	22.2920	83.5000	117.0000	128.5000	144.0000	295.0000
Lower BP	3390	82.8830	82.0000	80.0000	12.0236	48.0000	74.5000	82.0000	90.0000	142.5000
BMI	3390	25.7950	25.4000	22.9100	4.1069	15.9600	23.0300	25.4000	27.9975	56.8000
heartRate	3390	75.9773	75.0000	75.0000	11.9701	45.0000	68.0000	75.0000	83.0000	143.0000
glucose	3390	82.0865	80.0000	82.0865	23.1318	40.0000	72.0000	80.0000	85.0000	394.0000
Cardiovascular disease exposure	3390	0.1507	0.0000	0.0000	0.3578	0.0000	0.0000	0.0000	0.0000	1.0000

METHODOLOGY:


START
Import Dataset
(csv file)
Exploratory Data analysis
• Understanding the features, Y data profile
• Observing correlation between the features
Data Cleaning
• Null/Missing Value handling
• Impute values
• Duplicate handling
• Replacing Column names
Feature Selection
• Define Features as X and Prediction as Y
• Mention appropriate input instructions
Scaling
• Scale the features to a single value with StandardScaler
Define Train and Test inputs
• Assign train and test variables with the data
Model Selection
• Introduce K-Nearest Neighbor classifier algorithm
Model Training
• train_size = 0.80 (2712 records)
• test_size = 0.20 (678 records)
• random_state = 42
Model Evaluation
• Accuracy score = 0.82
• Contingency / Confusion matrix
• Classification report (Recall, Precision, F1 Score)
Model Prediction
• New clinical parameters evaluated
• CVD exposure prediction as healthy or may be exposed
Decision
• Displays the CVD exposure as healthy or may be exposed
No
Input Validation
• Ask the user to verify the input data with reference to input rules
Yes
Visualization
• Scatter plot for with regression line for understanding data pattern
END

TRAINING – TESTING SPLIT:

Dataset training and testing size split details are as follows:

SPLIT	SIZE (Total = 1 (3390 records)
Training	0.80 (2,712 records)
Testing	0.20 (678 records)

Split Description:

Training: Set of records with multiple combinations utilized by the model to understand the pattern.

Testing: Balance records are utilized to assess the model's performance following the training before a fresh input

MODEL SELECTION:

Model: K-Nearest Neighbor Classification (Supervised Model)

Description:

- The K-Nearest Neighbor classifier was selected because cardiovascular disease exposure results in categorical groups.
- It works on idea of proximity based learning, in which new individual's risk category is calculated comparing their clinical profile. Thus supporting for detecting trends in health data
- Cardiovascular risk assessment is influenced by several features. Hence, distance based KNN classification is appropriate for capturing trends between clinical features.

MODEL PARAMETERS:

Parameters Reference:

PARAMETERS	Values
test_size	0.20 (Lesser the better model learns).
random_state	42

Parameters Description:

- test_size : To diagnose the model's performance from a part of the original data.
- random_state: By shuffling the dataset, the model can be trained in many combinations. This ensures the model's repeatability by learning data using random data combination.

MODEL TRAINING:

- Total Dataset used = 3390 records
- Test size were listed and used test_size = 0.20. Hence, 678 records were adopted as test case and 2712 records were assigned in training the model.
- Scaler used: StandardScaler. Utilized to normalize several features into a single value.
- To assist the model with learning multiple possibilities, random_state 42 was implemented.

MODEL EVALUATION:

Manual conversion of factors: 'sex' : 'Gender', 'is_smoking' : 'Smoker', 'cigsPerDay' : 'Cigarette(s) per day', 'totChol' : 'Recent cholestrol level', 'sysBP' : 'Upper BP', 'diaBP' : 'Lower BP', 'TenYearCHD' : 'Cardiovascular disease exposure'.

Sample data values obtained:

TRY_X PARAMETERS							**FORECAST_Y**
AGE	**GENDER**	**SMOKER**	**CIGARETTE(S) PER DAY**	**UPPER BP**	**LOWER BP**	**HEART RATE**	**CARDIOVASCULAR DISEASE EXPOSURE**
64	0	1	9	128	71	60	0
61	1	1	10	130	70	67	0
58	1	0	0	139	96	75	0
46	1	0	0	136.5	92	68	0
65	0	0	0	171	89	82	0
43	0	1	15	101	68.5	79	0
58	0	1	3	120	80	78	0
50	1	0	0	148.5	100	80	0
42	0	1	5	127.5	80	75	0
52	1	0	0	126	80	104	0

EVALUATION METRICS:

Metrics used:

- **Confusion matrix**:
  - Confusion matrix represents a table visual that highlights number of predictions made correct and incorrect for each class by the model
  - The values are count based representation and used for other evaluation metrics.

		ACTUAL VALUES	
		POSITIVE (1)	NEGATIVE (0)
PREDICTED VALUES	POSITIVE (1)	TP	FP
	NEGATIVE (0)	FN	TN

Where:

$$TP = True\ positive\ = correct\ prediction\ of\ positive\ classes$$
$$TN\ = True\ negative\ = correct\ prediction\ of\ negative\ classes$$
$$FP\ = False\ positive\ = incorrect\ prediction\ of\ positive\ classes$$
$$FN\ = False\ negative = \ incorrect\ prediction\ of\ negative\ classes$$

  - True positive (TP) and true negatives (TN) provide the perfect right as right and perfect wrong as wrong prediction classes whereas false positive (FP) and false negative (FN) values show where the model predicts the right ones as wrong and vice versa.

- **Accuracy score**:
  - Used to measure the closeness of the measured value to the standard value. A single value that summarizes the whole model's performance
  - The value is calculated by using the below formula:

$$Accuracy\ score = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

$$TP = True\ positive\ = correct\ prediction\ of\ positive\ classes$$
$$TN\ = True\ negative\ = correct\ prediction\ of\ negative\ classes$$
$$FP\ = False\ positive\ = incorrect\ prediction\ of\ positive\ classes$$
$$FN\ = False\ negative = \ incorrect\ prediction\ of\ negative\ classes$$

    - The higher value of accuracy score marks the model’s performance at its best. Low accuracy score shows the model struggles to differentiate between classes.
- **Classification report**:
    - Provides comprehensive performance analysis such as recall, precision, F1 score and shows the behavior of each classes.
    - Recall: Ratio of correct predicted positive among all actual positives
        - Formula used: $Recall = \frac{TP}{TP+FN}$
        - Helps to capture exact positive cases
    - Precision: Ratio of correct predicted positives among all predicted positives
        - Formula used: $Precision = \frac{TP}{TP+}$
        - Useful to mitigate false positives
    - F1-Score: Utilizes precision and recall and calculates their mean harmonically to balance the false positives and negatives into a single value.
        - Formula used: $F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$
        - Reduces false positives and false negatives

Evaluated Metric Values from the model:

- Confusion matrix:

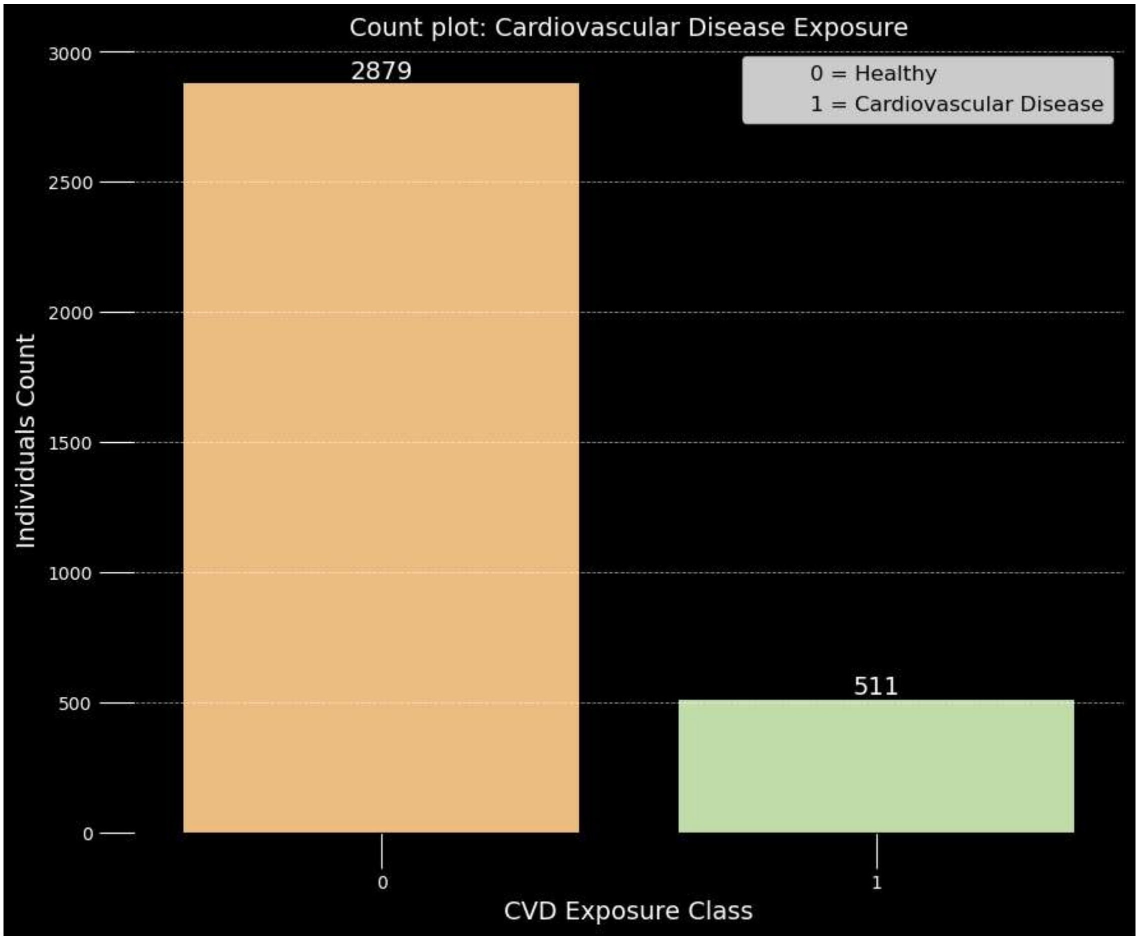
		ACTUAL VALUES	
		POSITIVE (1)	NEGATIVE (0)
PREDICTED VALUES	POSITIVE (1)	10	25
	NEGATIVE (0)	94	549

    - The contingency table utilized 678 test case record
    - 549 records were correctly predicted as negative (healthy) case, 25 false positive (unhealthy), 94 false negative (healthy), and 10 as correctly predicted as positive case (unhealthy)
    - Hence, the model should be optimized to reduce false positive and false negative
- Accuracy score = 0.82

- Classification report:
  - Recall: 0.96 (for healthy), 0.10 (for cardiovascular disease exposure)
  - Precision: 0.85 (for healthy), 0.29 (for cardiovascular disease exposure)
  - F1 score: 0.90 (for healthy), 0.14 (for cardiovascular disease exposure)

PLOTS:

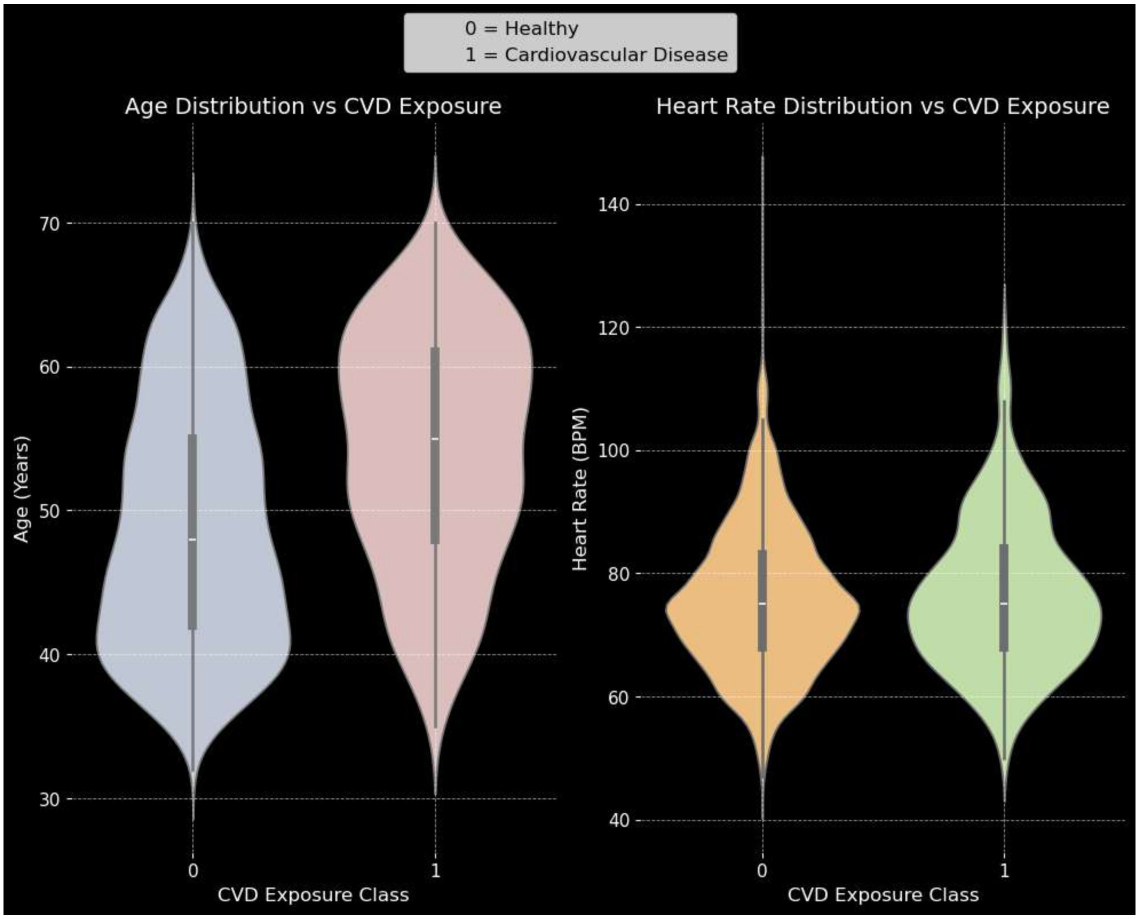
1) Count plot: CVD class wise individual count



PLOT INTERPRETATION:

- The count plot shows a clear imbalance, with healthy persons making up the majority of the dataset
- As per the given data, 3390 records reveal that 2879 people are healthy and 511 could be exposed to CVD.
- The inclusion of a large number of CVD exposure cases provides meaningful supervised learning.

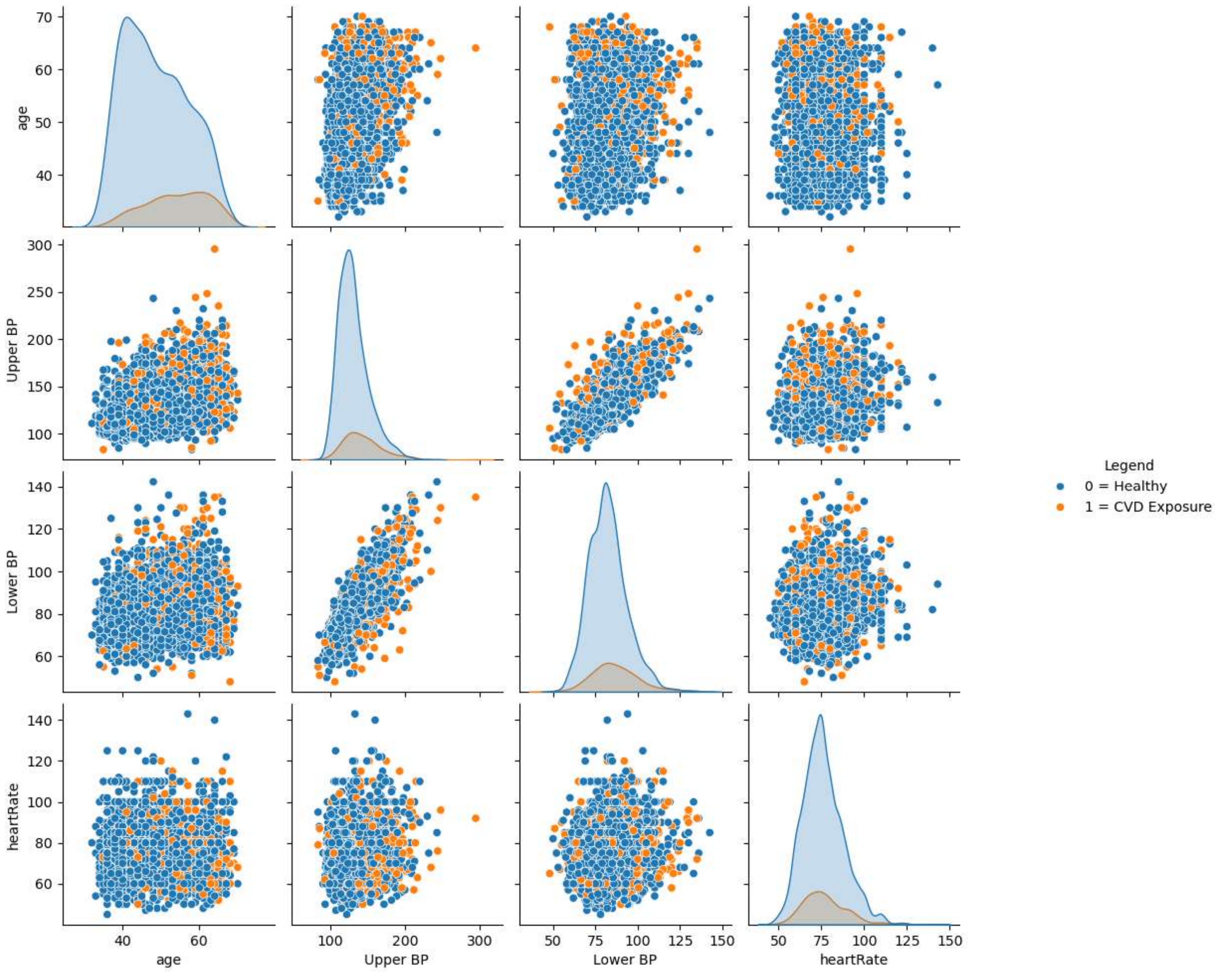
2) Violin plot: Age and CVD exposure count, Heart rate and CVD exposure



PLOT INTERPRETATION:

- On the left violin plot, the CVD group show greater median age and wide spread than healthy individuals between the age 50 to 65, exhibiting age as a substantial risk factor for disease exposure.
- On the right we have the individuals with CVD exposure, slightly raised heart rate distributions, which implies the anomaly contribution. Thus heart rate is another complimentary feature when paired with other clinical factors
- The above plots also prove the adoption of distance based classifier model for CVD exposure prediction.

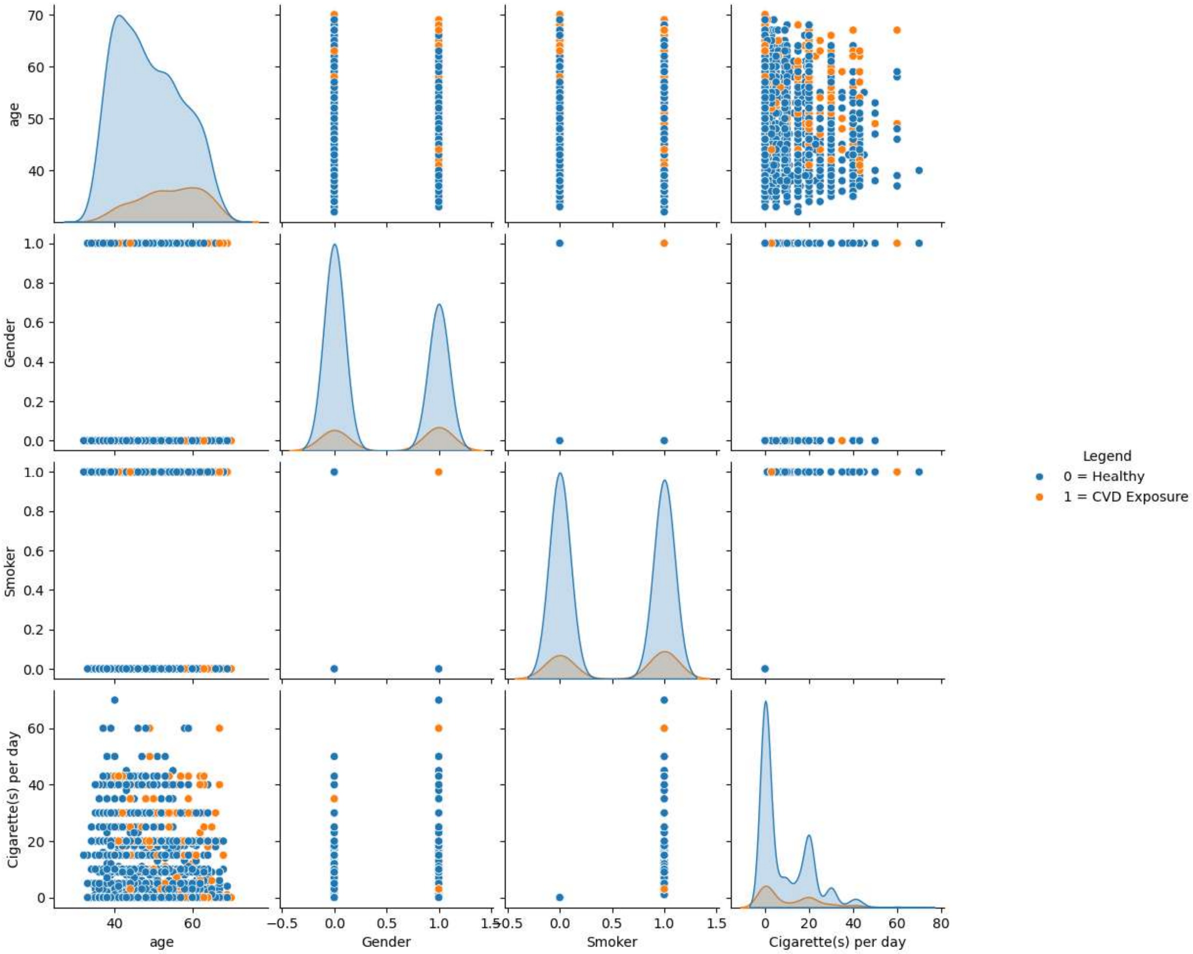
3) Pair plot: Relation between age, upper bp, lower bp, heart rate with CVD exposure



PLOT INTERPRETATION:

- The pair plot portrays the clustering of CVD exposure at higher age and blood pressure level.
- Features comparing with age and bp, heart rate and upper bp demonstrate the CVD exposure, which the KNN classifier model efficiently captures via feature similarity based learning.

4) Pair plot: Relation between age, gender, smoker, cigarette(s) per day with CVD exposure



PLOT INTERPRETATION:

- The above visual demonstrates minimal higher exposure among smokers and those more cigarette consumers at old ages.
- On the other hand, non-smokers are affected by smoker and have a higher risk of developing CVD.

Overall, the visualizations indicate the presence of clinically relevant patterns and multivariate correlations, justifying the KNN classifier model’s 82% accuracy.

NEW DATA PREDICTION:

Features	Input Values			Accuracy
	Set 1	Set 2	Set 3	
AGE	35	50	62	0.82
GENDER	female	male	male	
SMOKE	no	yes	yes	
CIGARETTE(S) PER DAY	0	10	25	
UPPER BLOOD PRESSURE	115	135	160	
LOWER BLOOD PRESSURE	75	88	100	
HEART RATE	72	82	95	
	Cardiovascular Disease Exposure Status			
	Healthy	May have CVD	May have CVD	

CONCLUSION:

From an individual's new clinical parameter data, the trained K-Nearest Neighbor classifier has successfully predicted cardiovascular disease. The model had the highest performance with a promising classification report and the best projected accuracy of 0.82. Additionally, when significantly more data is encountered, this would show that the model might be used in a real-time setting for further model research.