Classification
Project: Cardiovascular Risk Prediction

Project Description:

Business Context
The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

Main Libraries to be Used:

- Pandas for data manipulation, aggregation
- Matplotlib and Seaborn for visualisation and behaviour with respect to the target variable
- NumPy for computationally efficient operations
- Scikit Learn for model training, model optimization, and metrics calculation

Project should include:

1. **Problem Statement**
2. **Import libraries**
3. **Load dataset**
4. **Data cleaning**
    - Handle missing values
    - Convert data types
    - Remove duplicates
5. **Exploratory Data Analysis (EDA)**
    - Visualize distributions
        i. Univariate analysis
        ii. Bivariate analysis
        iii. Multivariate analysis
    - Correlation analysis
    - Feature-target relationships
6. **Outlier treatment**
    - Boxplot
7. **Check distributions & apply transformations (if needed)**
    - Skewness/Kurtosis
    - Log Transformation, sqrt
8. **Feature engineering**
    - Create new features
    - One-hot encoding (for categorical)
9. **Split data into train/test sets**

10. **Train Logistic Regression model**
11. **Feature Scaling**
12. **Prediction using the algorithm**
13. **Evaluate model performance using Confusion Metrics**
    - Accuracy
    - Precision
    - Recall
    - F1 Score
14. **Perform the same steps for**
    - Decision tree Classifier
    - Random Forest Classifier
    - Support Vector Machine Classifier
    - K Nearest Neighbor Classifier
15. **Perform the Cross Validation using Cross_val_score for all the algorithms**
16. **Print the final Conclusion**

Link to Dataset: https://github.com/rahulinchal/SPPU

Data Description:

Fields	**Description**
Sex	Gender
Education	Education level (1 - Low, 5 - High)
Age	Age (in years)
is_smoking	Whether smoking currently or not
Cigs_per_day	Cigarettes smoked per day
BP_meds	Whether taking BP meds or not (0 = No, 1 = Yes)
Prevalent stroke	If the person has a history of strokes (0 = No, 1 = Yes)
Prevalent hyp	If the patient has a history of hypertension (0 = No, 1 = Yes)
Diabetes	Patient has diabetes or not (0 = No, 1 = Yes)
Tot chol	Cholesterol measure
Sys BP	BP measure
Dia BP	BP measure
BMI	Body mass index
Heart Rate	Heart Rate measure
Glucose	Glucose level
TenYearCHD	Target Variable: 10-Year risk of coronary heart disease (0 = No, 1 = Yes)