Credit Hours System

SBEN454: Data Mining and Machine Learning in Healthcare

Cairo University

Faculty of Engineering

CARDIOVASCULAR DISEASE CLASSIFICATION PROJECT

(https://www.kaggle.com/sulianova/cardiovascular-disease-dataset).

WE HAVE USED CARDIOVASCULAR DISEASE DATASET. BASED ON SOME HEALTH INFORMATION OF AN INDIVIDUAL OUR MODEL WILL PREDICT WHETHER HE HAS ANY CARDIOVASCULAR DISEASE OR NOT.

Submitted to: Dr. Inas A. Yassine

Names	IDs
Amira Mahmoud	1170498
Alaa Ossama	1170185
Salma Hazem	1170425

DATA DESCRIPTION

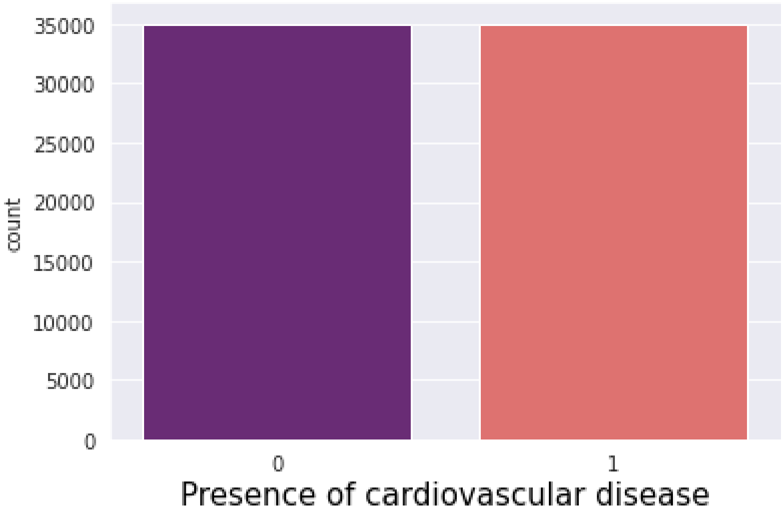
There are 3 types of input features:

1. Objective: factual information;
2. Examination: results of medical examination;
3. Subjective: information given by the patient.

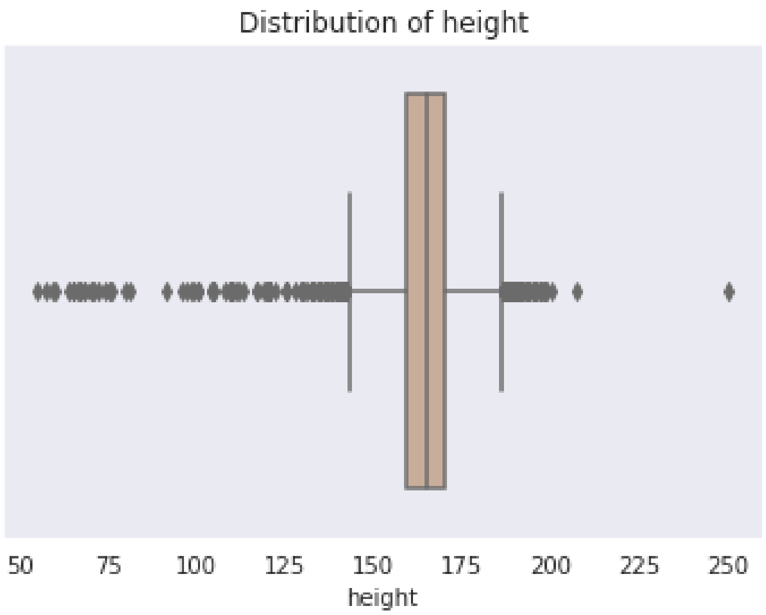
FEATURES:

- Age | Objective Feature | age | int (days)
- Height | Objective Feature | height | int (cm) |
- Weight | Objective Feature | weight | float (kg) |
- Gender | Objective Feature | gender | categorical code |
- Systolic blood pressure | Examination Feature | ap_hi | int |
- Diastolic blood pressure | Examination Feature | ap_lo | int |
- Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
- Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
- Smoking | Subjective Feature | smoke | binary |
- Alcohol intake | Subjective Feature | alco | binary |
- Physical activity | Subjective Feature | active | binary |
- Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

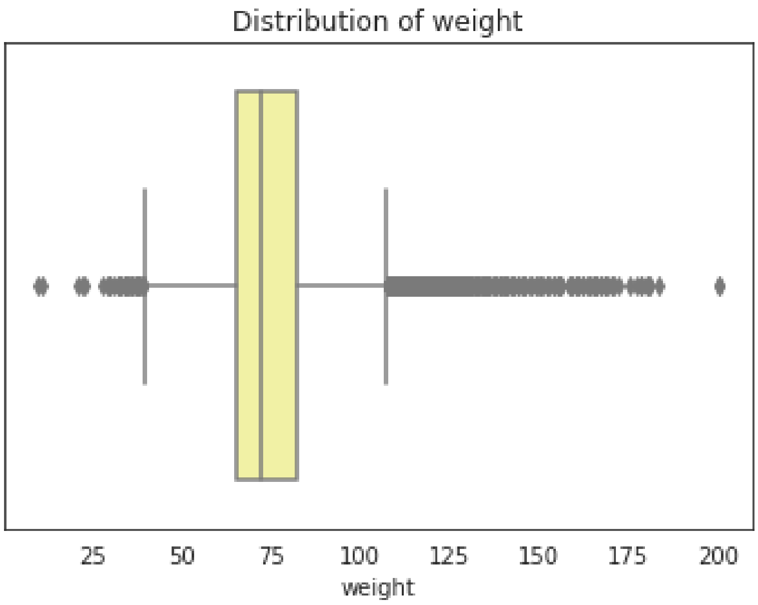
EXPLORATORY DATA ANALYSIS (VISUALISTION)



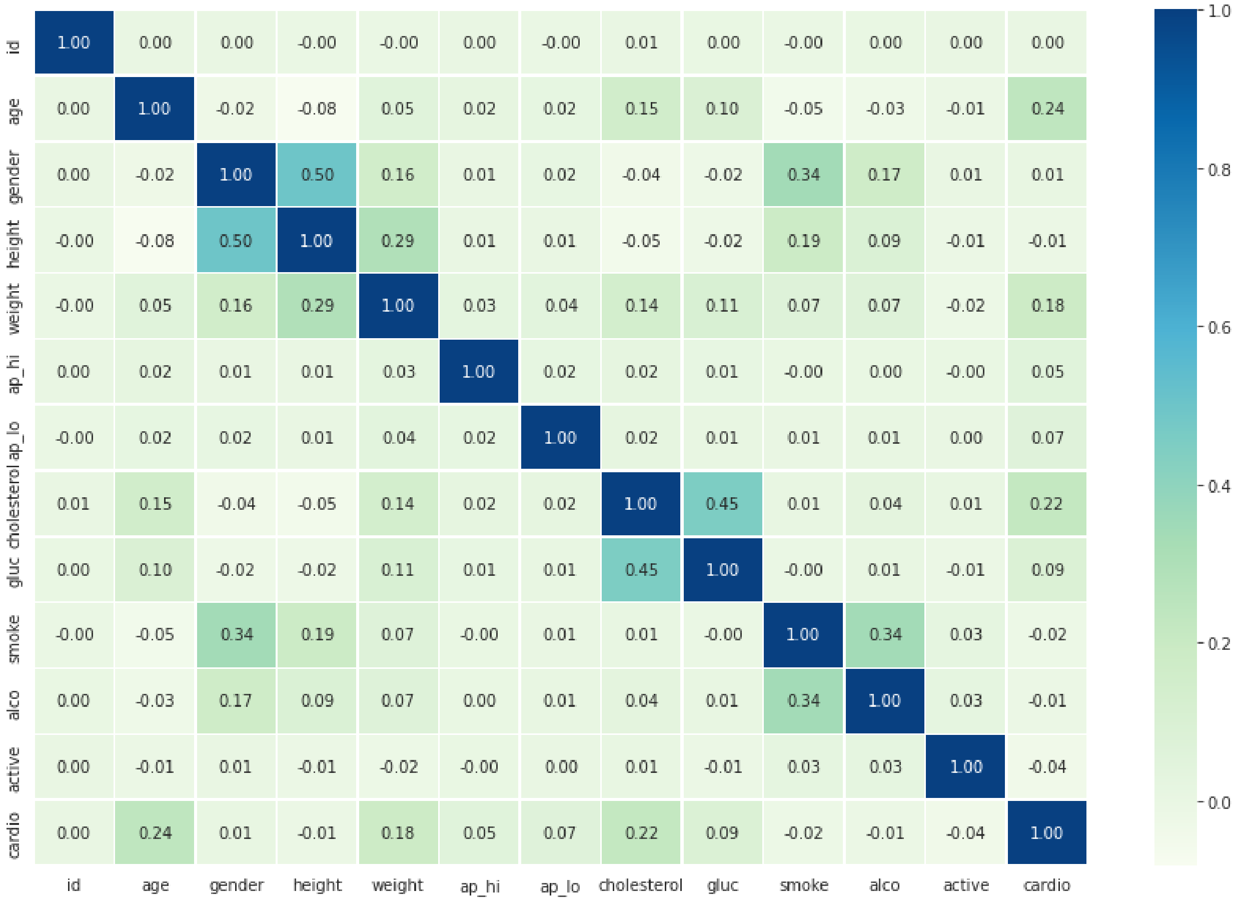
Data is almost balanced



250 cm height is extremely rare cases

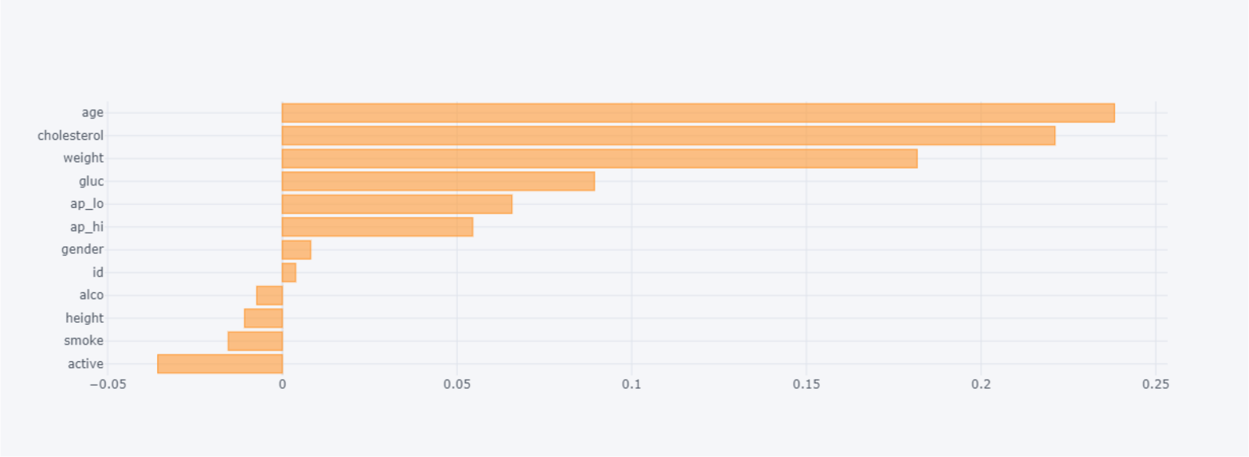


200 kg weight is extremely rare cases



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
id	1.00	0.00	0.00	-0.00	-0.00	0.00	-0.00	0.01	0.00	-0.00	0.00	0.00	0.00
age	0.00	1.00	-0.02	-0.08	0.05	0.02	0.02	0.15	0.10	-0.05	-0.03	-0.01	0.24
gender	0.00	-0.02	1.00	0.50	0.16	0.01	0.02	-0.04	-0.02	0.34	0.17	0.01	0.01
height	-0.00	-0.08	0.50	1.00	0.29	0.01	0.01	-0.05	-0.02	0.19	0.09	-0.01	-0.01
weight	-0.00	0.05	0.16	0.29	1.00	0.03	0.04	0.14	0.11	0.07	0.07	-0.02	0.18
ap_hi	0.00	0.02	0.01	0.01	0.03	1.00	0.02	0.02	0.01	-0.00	0.00	-0.00	0.05
ap_lo	-0.00	0.02	0.02	0.01	0.04	0.02	1.00	0.02	0.01	0.01	0.01	0.00	0.07
cholesterol	0.01	0.15	-0.04	-0.05	0.14	0.02	0.02	1.00	0.45	0.01	0.04	0.01	0.22
gluc	0.00	0.10	-0.02	-0.02	0.11	0.01	0.01	0.45	1.00	-0.00	0.01	-0.01	0.09
smoke	-0.00	-0.05	0.34	0.19	0.07	-0.00	0.01	0.01	-0.00	1.00	0.34	0.03	-0.02
alco	0.00	-0.03	0.17	0.09	0.07	0.00	0.01	0.04	0.01	0.34	1.00	0.03	-0.01
active	0.00	-0.01	0.01	-0.01	-0.02	-0.00	0.00	0.01	-0.01	0.03	0.03	1.00	-0.04
cardio	0.00	0.24	0.01	-0.01	0.18	0.05	0.07	0.22	0.09	-0.02	-0.01	-0.04	1.00

CORRELATION OF FEATURES WITH TARGET VARIABLE



The first 3 feature (Age, Cholesterol and weight) are most effective on cardiovascular disease (Age is the most effective)

PREPROCESSING NEEDED

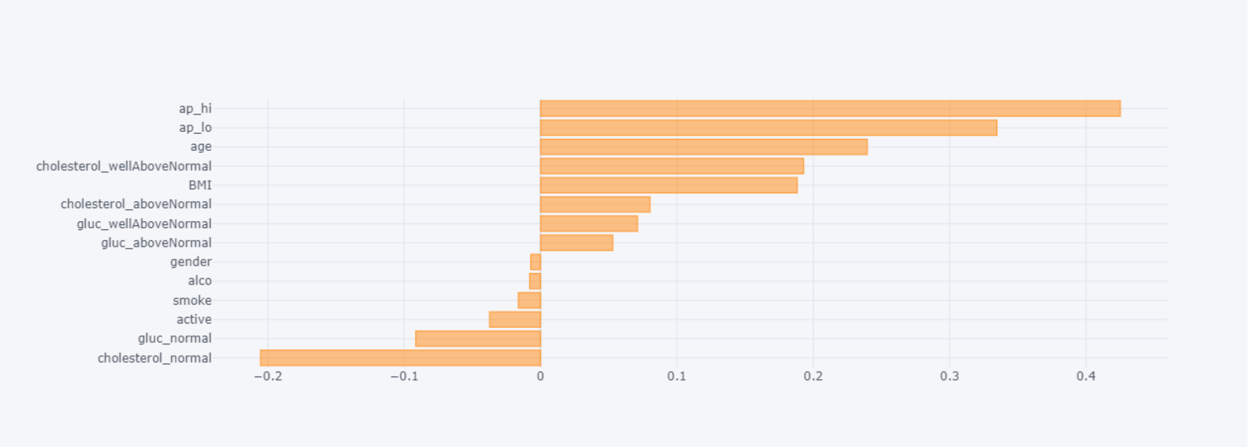
- ID needs to be dropped.
- Age provided is in days. We will convert it to years
- Gender can be converted to binary
- ap_hi and ap_lo has negative numbers. This means that we have outliers so we need to remove them.
- Gluc and Cholesterol need to be converted to dummies
- There are many rare cases in height and weight features, so we can combine them in BMI feature (get 1 feature from 2 features).

DATA CLEANING & PREPROCESSING

- Remove Outliers
  - BMI more than 100 or less than 10
  - ap_hi more than 250 or less than 20
  - ap_lo more than 200 or less than 20
- Convert categorical variable into indicator variables
- Scaling non-categorical data
- PCA

We removed 1,251 row that means 1.7 % of data which is not too high

DATA CORRELATION AFTER PREPROCESSING



TRAINING

Splitting data into 0.25% for testing and 0.75% for training

Note: Numbers may vary with every run

CLASSIFICATION

We used different classifiers and used accuracy and F1 score to evaluate the classifiers

CLASSIFICATION WITHOUT PCA OR DATA SCALING

Classifiers	**Accuracy (%)**	**F1-score**
Logistic Regression	71.92	0.70
Decision Tree	62.76	0.63
Random Forrest	69.83	0.70
Support Vector Machine	55.16	0.69
K-Nearest Neighbor	68.80	0.68
Naïve Bayes	66.92	0.63

CLASSIFICATION MODELS WITHOUT DATA SCALING TO CHOOSE PCA NUMBER OF COMPONENTS

The non-normalized data showed the highest accuracy in almost all models with PCA n_components = 5

Classifiers	**Accuracy (%)**	**F1-score**
Logistic Regression	71.50	0.70
Decision Tree	63.08	0.63
Random Forrest	69.43	0.69
Support Vector Machine	71.31	0.70
K-Nearest Neighbor	68.37	0.68
Naïve Bayes	70.00	0.67

SCALE DATA USING MINMAXSCALER, STANDARDSCALER, AND NORMALIZER WITH PCA COMPONENT = 5

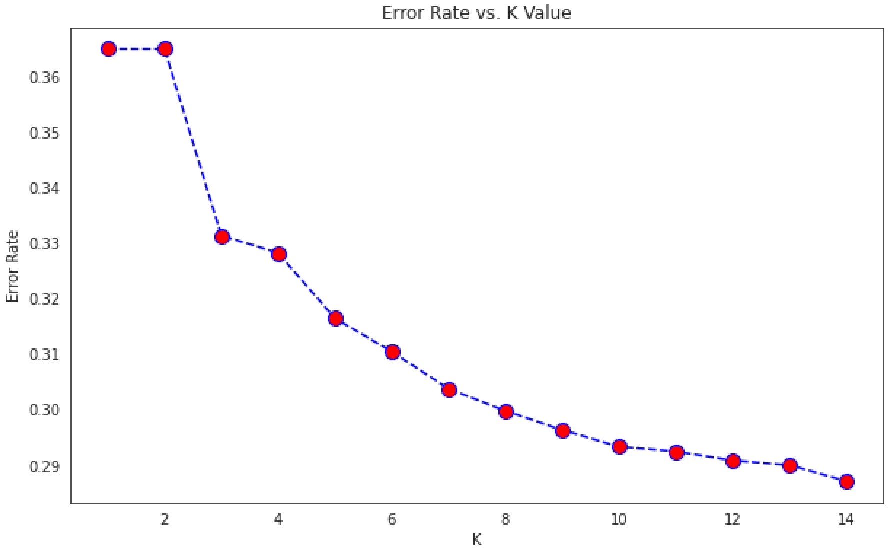
Accuracies of 'StandardScaler' and 'Normalization' are too close so we will choose 'StandardScaler'

ACCURACIES WITH PCA COMPONENTS = 5 AND SCALED DATA

Classifiers	**Accuracy (%)**	**F1-score**
Logistic Regression	71.43	0.70
Decision Tree	62.82	0.63
Random Forrest	69.03	0.69
Support Vector Machine	71.29	0.69
K-Nearest Neighbor	68.36	0.68
Naïve Bayes	69.47	0.67

ENHANCEMENTS BY CHANGING HYPER-PARAMETERS

KNN ENHANCEMENT



Accuracy went from 68.36% to accuracy 71.28%, with k = 14

TREE ENHANCEMENT

Accuracy went from 62.82% to accuracy 72.45%, max depth = 5

RANDOM FOREST ENHANCEMENT

Accuracy went from 69.03% to accuracy 71.11%, with best parameters:'max_depth': 90, 'max_features': 2, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 100

CONCLUSION

Classifiers	**Accuracy (%)**
Logistic Regression	71.43
Decision Tree	72.45
Random Forrest	71.11
Support Vector Machine	71.29
K-Nearest Neighbor	71.28
Naïve Bayes	69.47

Decision Tree is the highest accuracy 72.45%