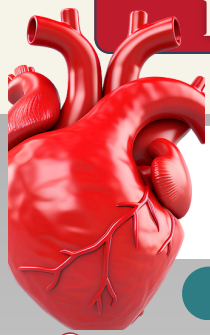
LC09 - Group 7

Heart Attack Risk Prediction:

Health Data Analysis for Early Detection



This dataset examines factors affecting the 10-year risk of coronary heart disease (CHD). CVDs are the leading cause of death globally, with 85% of deaths from heart attacks and strokes. Key risks include poor diet, inactivity, smoking, alcohol, and air pollution, leading to conditions like hypertension and obesity. Early detection and lifestyle changes are crucial for prevention.

Analysis Goal!

Risk prediction

Preventive Intervention

Trend Analysis

Feature Engineering

Feature Selection | Impute Null Values | Handle Outliers | Normalization | Label Enconding

Data Understanding

Dimension: 16 x 4238

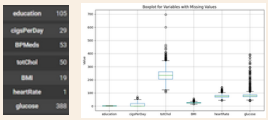
Continuous

- age
- cigsPerDay
- totChol
- sysBP
- diaBP
- BMI
- heartRate
- glucose

Categorical

- male
- education
- currentSmoker
- BPMeds
- prevalentStroke
- prevalentHyp
- diabetes
- TenYearCHD (TARGET)

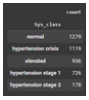
Missing Value

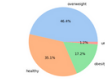


Handling missing values: mean (outliers), median (no outliers), mode (categorical).

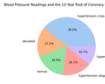
Make sure all data is in the correct type.

Classification





Categorizes BMI into 4 classes



Categorizes systolic blood pressure into five classes and adds a new column

Normalization

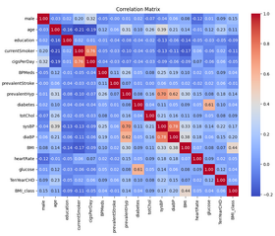
The column **cigsPerDay** is continuous, with a scale ranging from **0 to 70** and only a few outliers, making **Min-Max Scaling** appropriate. It transforms the data into a **[0, 1] range**. The same process repeated to reduce **outliers** for **glucose**, but it decrease the accuracy score so we decided not to use it.

Risk Factors

Variable-Target Correlation Comparison

age

blood pressure



glucose

hypertension

FIT IN MODEL

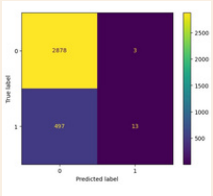
Stacking Classifier:
Logistic Regression, Random Forest, SVC

```
Final Accuracy: 0.8525508699498673
              precision    recall  f1-score   support

           0       1.00      0.85      0.92      3375
           1       0.03      0.81      0.05        16

    accuracy                           0.85      3391
   macro avg       0.51      0.83      0.48      3391
weighted avg       0.99      0.85      0.92      3391
```

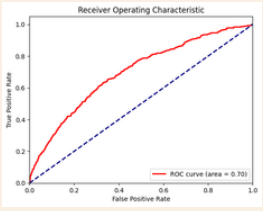
85.255 % Accuracy



There are 2878 true negatives (correctly predicted 0), 3 false positives (incorrectly predicted 1), 497 false negatives (incorrectly predicted 0), and 13 true positives (correctly predicted 1).

AUC: 0.70, the model is decent but not strong. Ideal is closer to 1.

ROC Rises above the diagonal, better than random, but not steep at the start.



CONCLUSION

The model evaluation reveals that while the overall accuracy is 85%. Linear models like Logistic Regression and SVM work well with simpler, linearly separable data, while Random Forest captures more complex patterns, indicating that the data's complexity can be handled by both linear and non-linear models. Logistic Regression was chosen for its interpretability, while Random Forest was selected for capturing non-linear relationships. Hyperparameter tuning optimized each model to balance accuracy, regularization, and prevent overfitting. Stacking is recommended to combine the strengths of each model, overcoming individual weaknesses and improving overall performance with a meta-learner.

Lifestyle Strategies

Eat **heart-friendly** foods

Watch your **blood pressure**

Set **exercise** goals

Eirene Michella (2702256630) - Lea Audrey (2702261914) - Tifanny Felita (2702265641)