

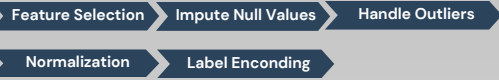


# Heart Attack Risk Prediction:

## 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!



### Data Understanding

**Continuous**

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

**Dimension: 16 x 4238**

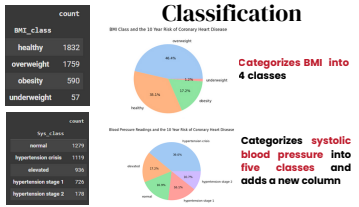
**Categorical**

- 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**

BMI Class and the 10 Year Risk of Coronary Heart Disease

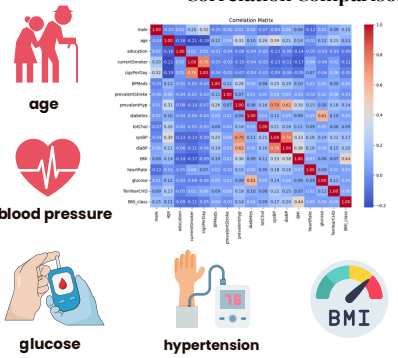
Categorizes BMI into 4 classes

**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



### Stacking Classifier: Logistic Regression, Random Forest, SVC

Final Accuracy: 0.8525508699498673

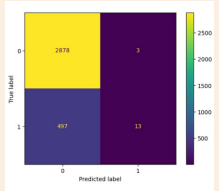
	precision	recall	f1-score	support
0	1.00	0.85	0.92	3375
1	0.83	0.81	0.85	16

	accuracy
macro avg	0.51
micro avg	0.83
weighted avg	0.99

**FIT IN MODEL**

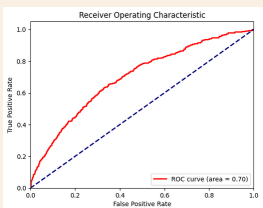
**85.255 % Accuracy**



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

