Prediction of Heart Disease with Emphasis on Factors Impacting it

Krishna Hemant . Yanming Liu

Problem Setting

- Numerous people have lost their lives because of not taking preventive measures with respect to their cardiovascular health before anything serious occurs
- As technology has increased multi fold over the years, data science has proved to play an important role in improving healthcare systems.

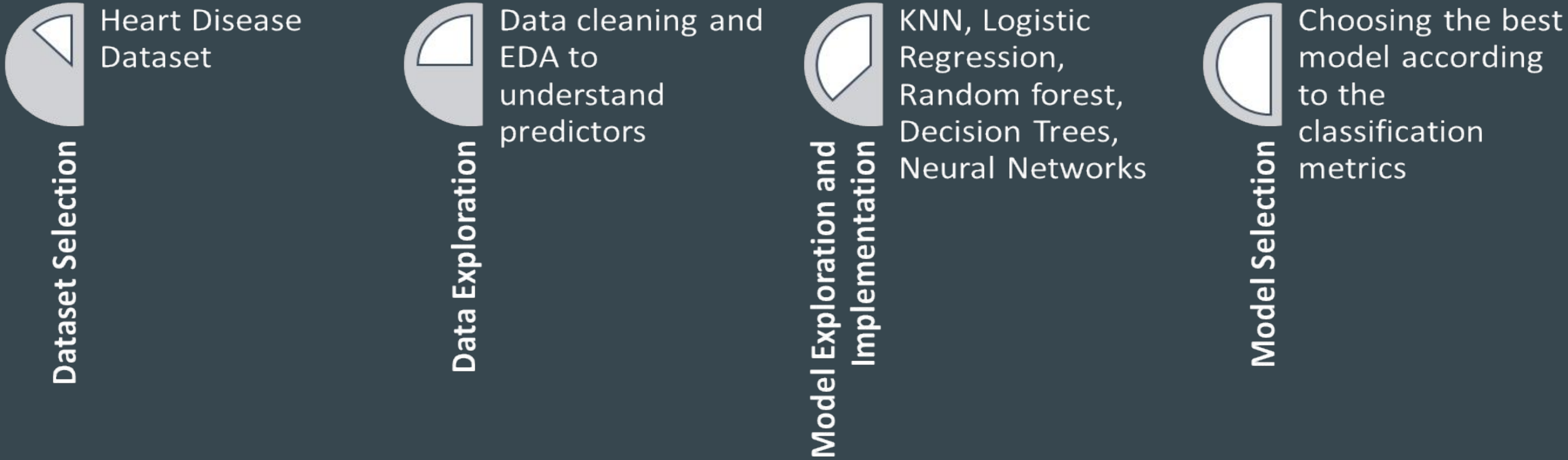
Research Objective

❏ Build a Machine Learning Model help predict if a subject has Heart Disease or not

❏ Find significant predictors which impact the prediction of Heart Disease

❏ General public can use the healthcare app supported by our Machine Learning Model to improve quality of life and treat heart disease in the earlier stages

Blueprint at a Glimpse



Dataset Selection

Heart Disease Dataset

Data Exploration

Data cleaning and EDA to understand predictors

Model Exploration and Implementation

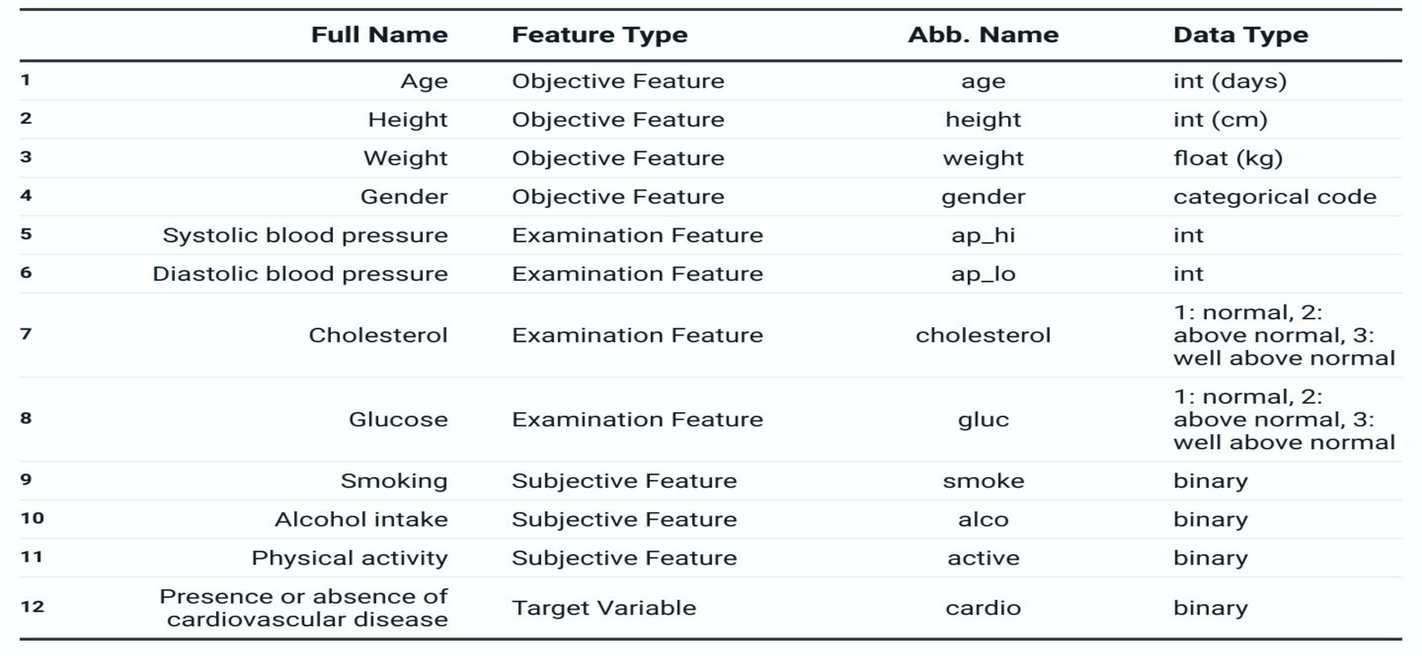
KNN, Logistic Regression, Random forest, Decision Trees, Neural Networks

Model Selection

Choosing the best model according to the classification metrics

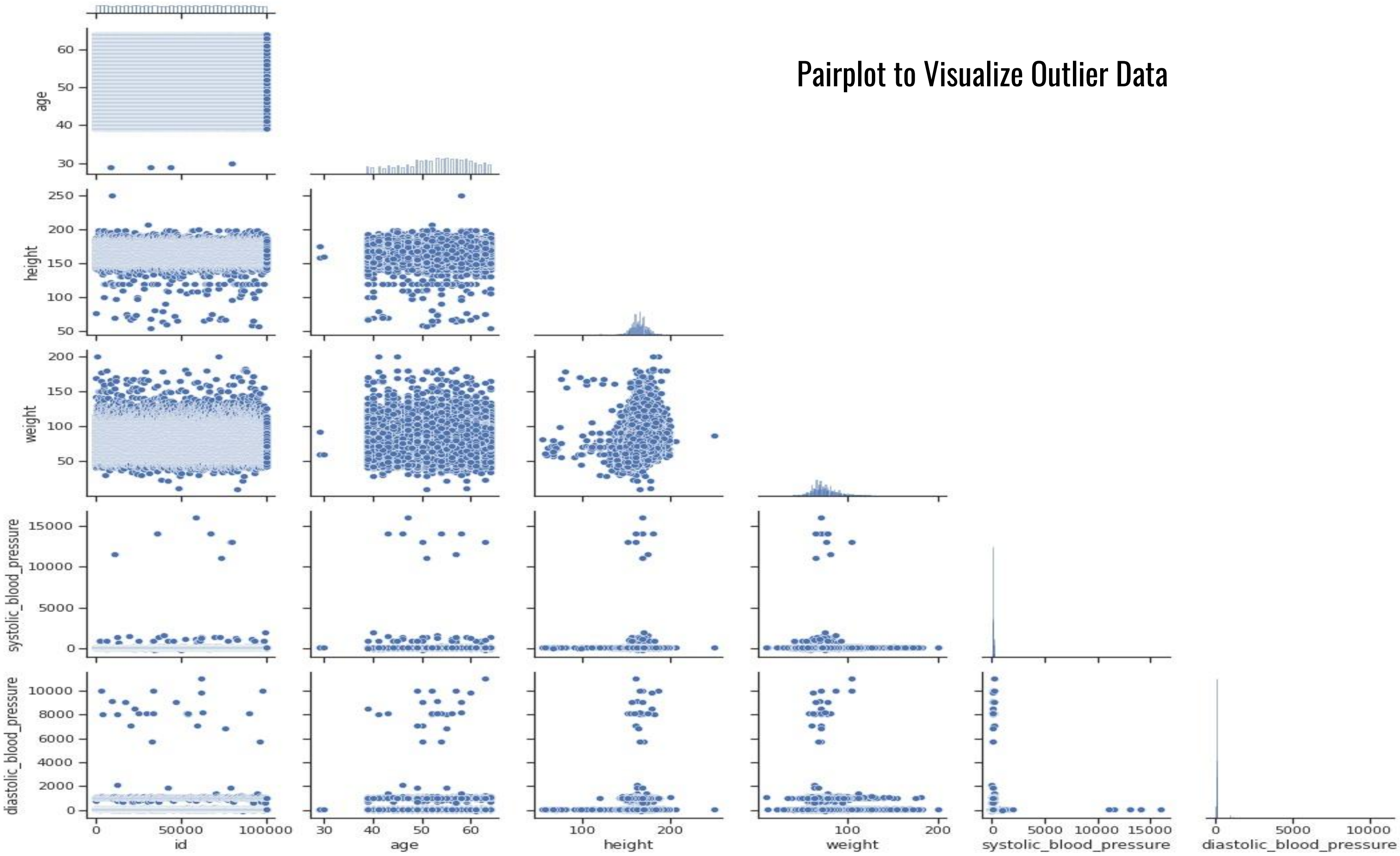
Data Source & Explanation

- https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
- 70000 records

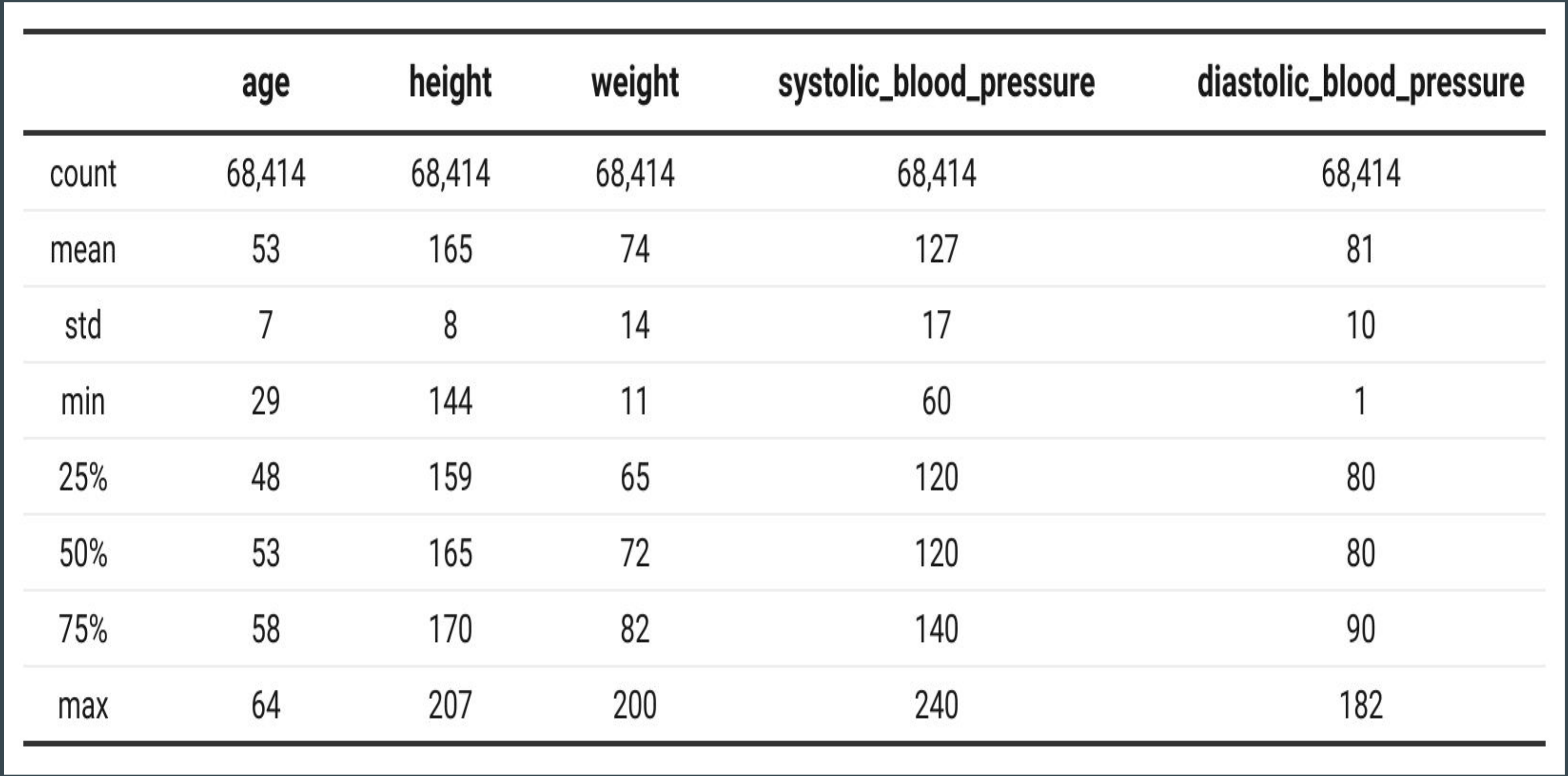


	Full Name	Feature Type	Abb. Name	Data Type
1	Age	Objective Feature	age	int (days)
2	Height	Objective Feature	height	int (cm)
3	Weight	Objective Feature	weight	float (kg)
4	Gender	Objective Feature	gender	categorical code
5	Systolic blood pressure	Examination Feature	ap_hi	int
6	Diastolic blood pressure	Examination Feature	ap_lo	int
7	Cholesterol	Examination Feature	cholesterol	1: normal, 2: above normal, 3: well above normal
8	Glucose	Examination Feature	gluc	1: normal, 2: above normal, 3: well above normal
9	Smoking	Subjective Feature	smoke	binary
10	Alcohol intake	Subjective Feature	alco	binary
11	Physical activity	Subjective Feature	active	binary
12	Presence or absence of cardiovascular disease	Target Variable	cardio	binary

Pairplot to Visualize Outlier Data

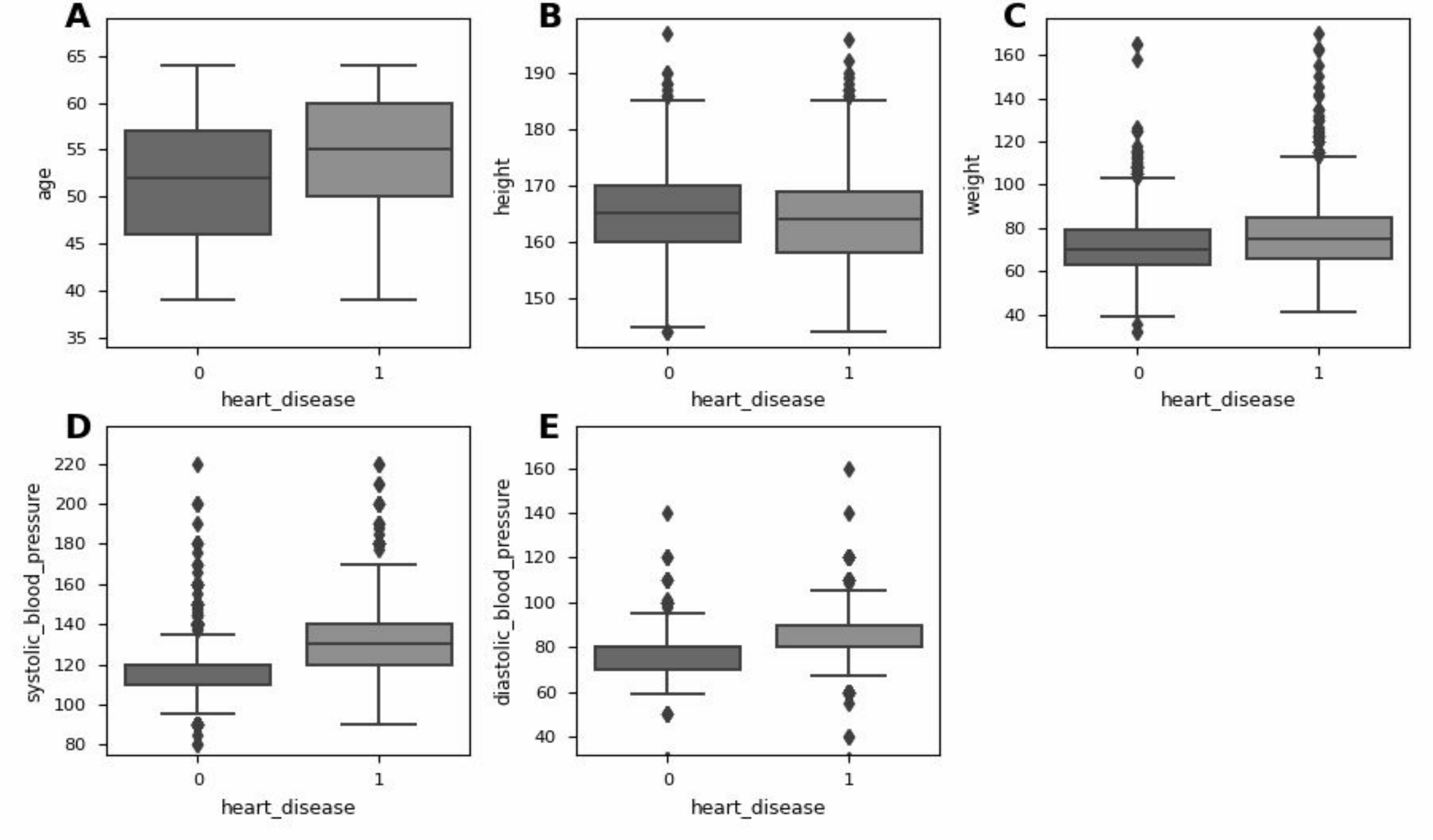


Data Pre-Processing for Numerical Variables

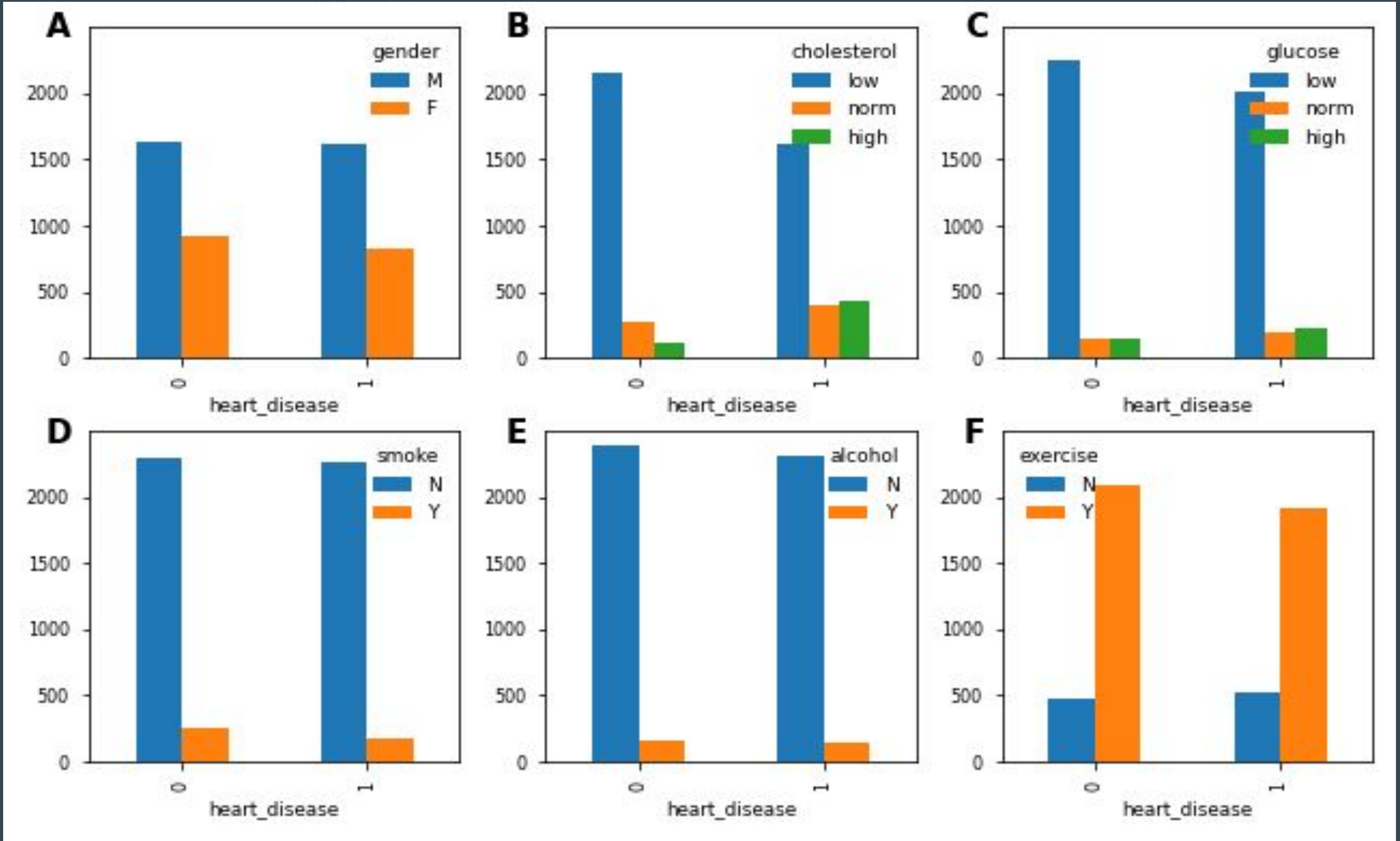


	age	height	weight	systolic_blood_pressure	diastolic_blood_pressure
count	68,414	68,414	68,414	68,414	68,414
mean	53	165	74	127	81
std	7	8	14	17	10
min	29	144	11	60	1
25%	48	159	65	120	80
50%	53	165	72	120	80
75%	58	170	82	140	90
max	64	207	200	240	182

Exploratory Data Analysis

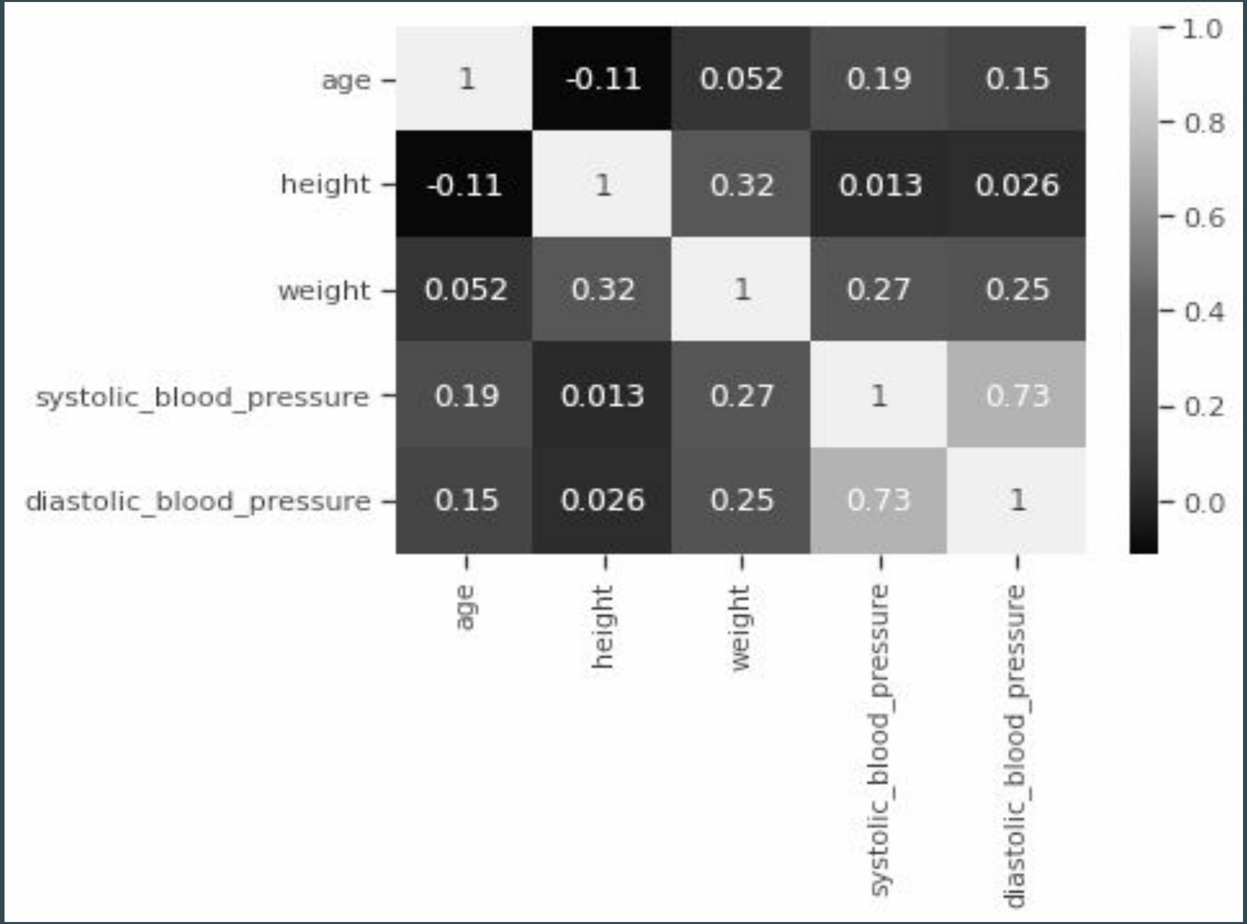


Exploratory Data Analysis



Correlation Analysis

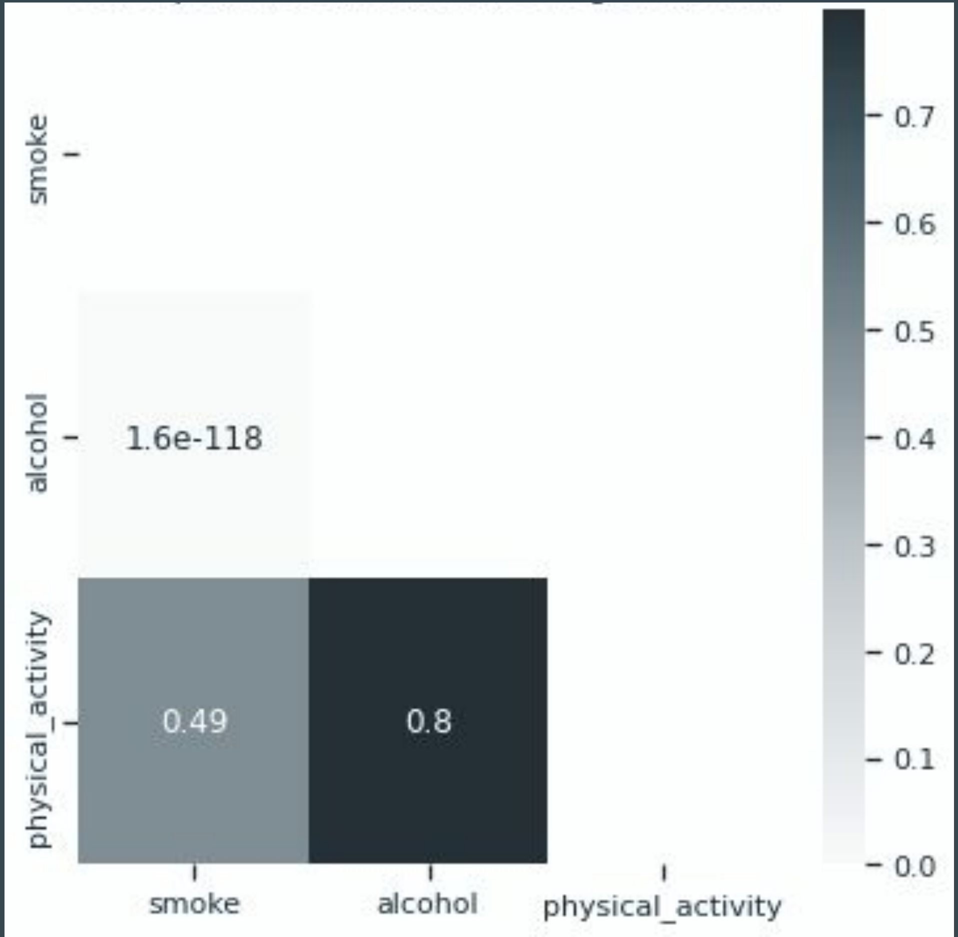
NUMERICAL VARIABLES ->



	age	height	weight	systolic_blood_pressure	diastolic_blood_pressure
age	1	-0.11	0.052	0.19	0.15
height	-0.11	1	0.32	0.013	0.026
weight	0.052	0.32	1	0.27	0.25
systolic_blood_pressure	0.19	0.013	0.27	1	0.73
diastolic_blood_pressure	0.15	0.026	0.25	0.73	1

Correlation Analysis - Chi-square

CATEGORICAL VARIABLES ->

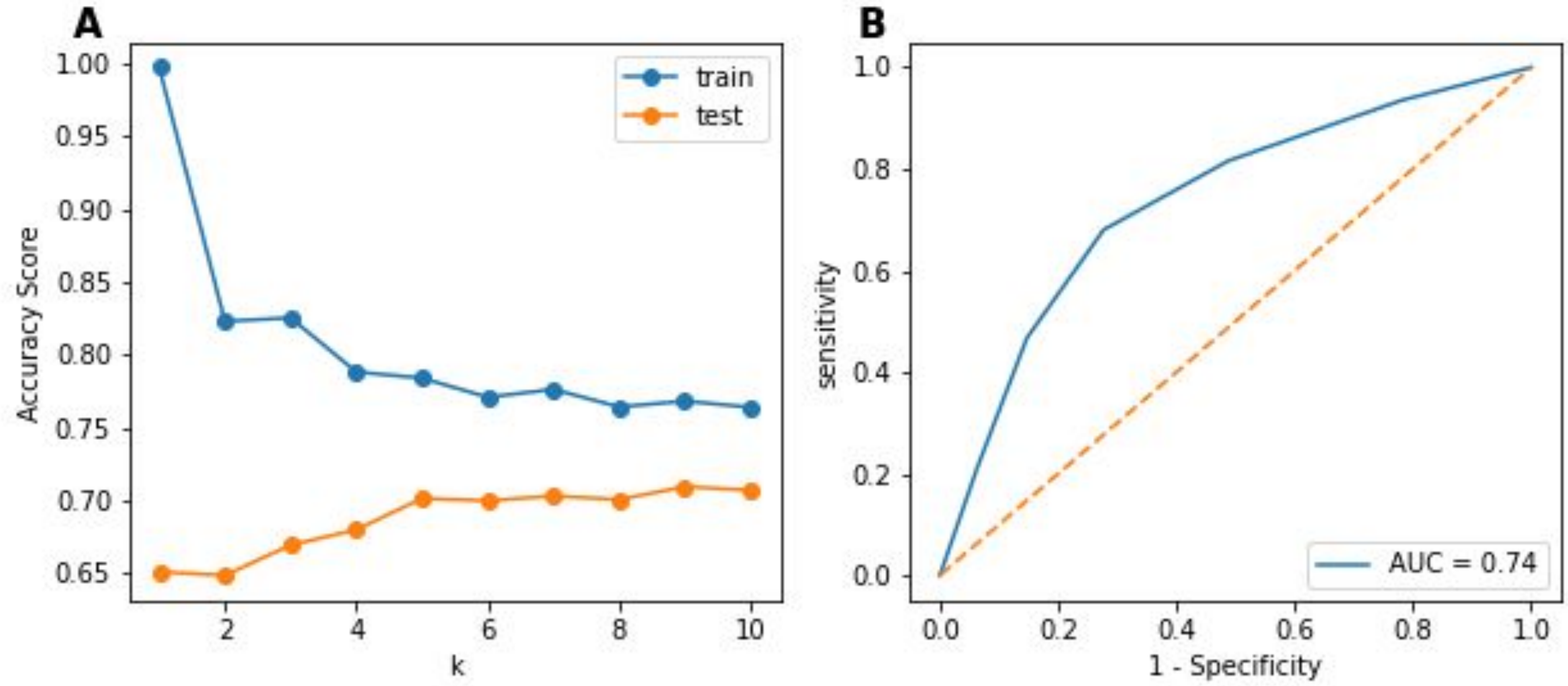


DATA MINING MODELS

KNN - K Nearest Neighbors

- ❏ Sampled the Dataset into 5000 records as KNN struggles with large Data
- ❏ Checked for K values in range (1,10) and decided to go with K = 5 to avoid overfitting
- ❏ Accuracy of 70.2% when K = 5

KNN - K Nearest Neighbors



TREES -> Decision Trees, Random Forest and Boosted Trees

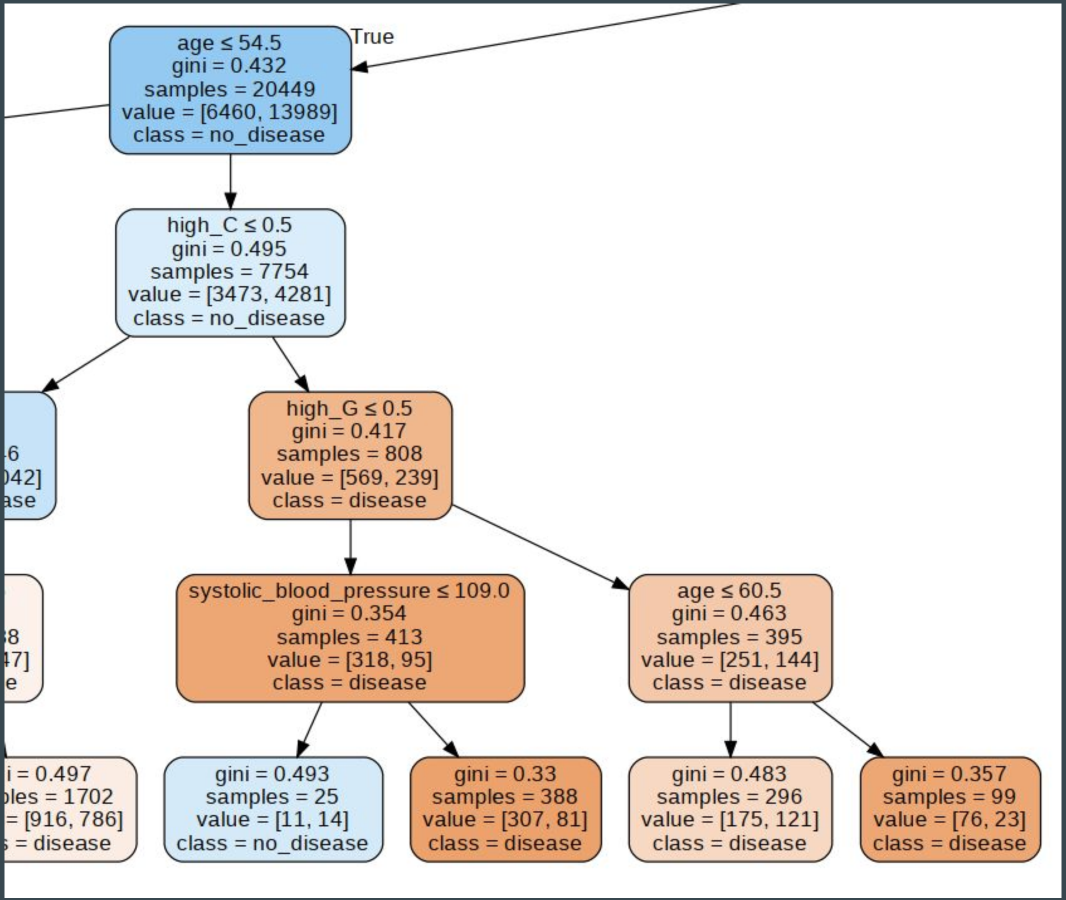
Decision Tree

- It was performed on the full cleaned dataset with 68k rows and the baseline accuracy for decision tree was 62.7%.
- Cross validation is applied for stable output on Decision Tree and Grid search was conducted to get the best parameters for a tree with the highest accuracy.
- The Best Parameters were:
  - criterion = "gini",
  - random_state = 1,
  - max_depth = 6,
  - min_impurity_decrease = 0.0004,
  - min_samples_split = 10

And it yielded an accuracy of 73.18 %

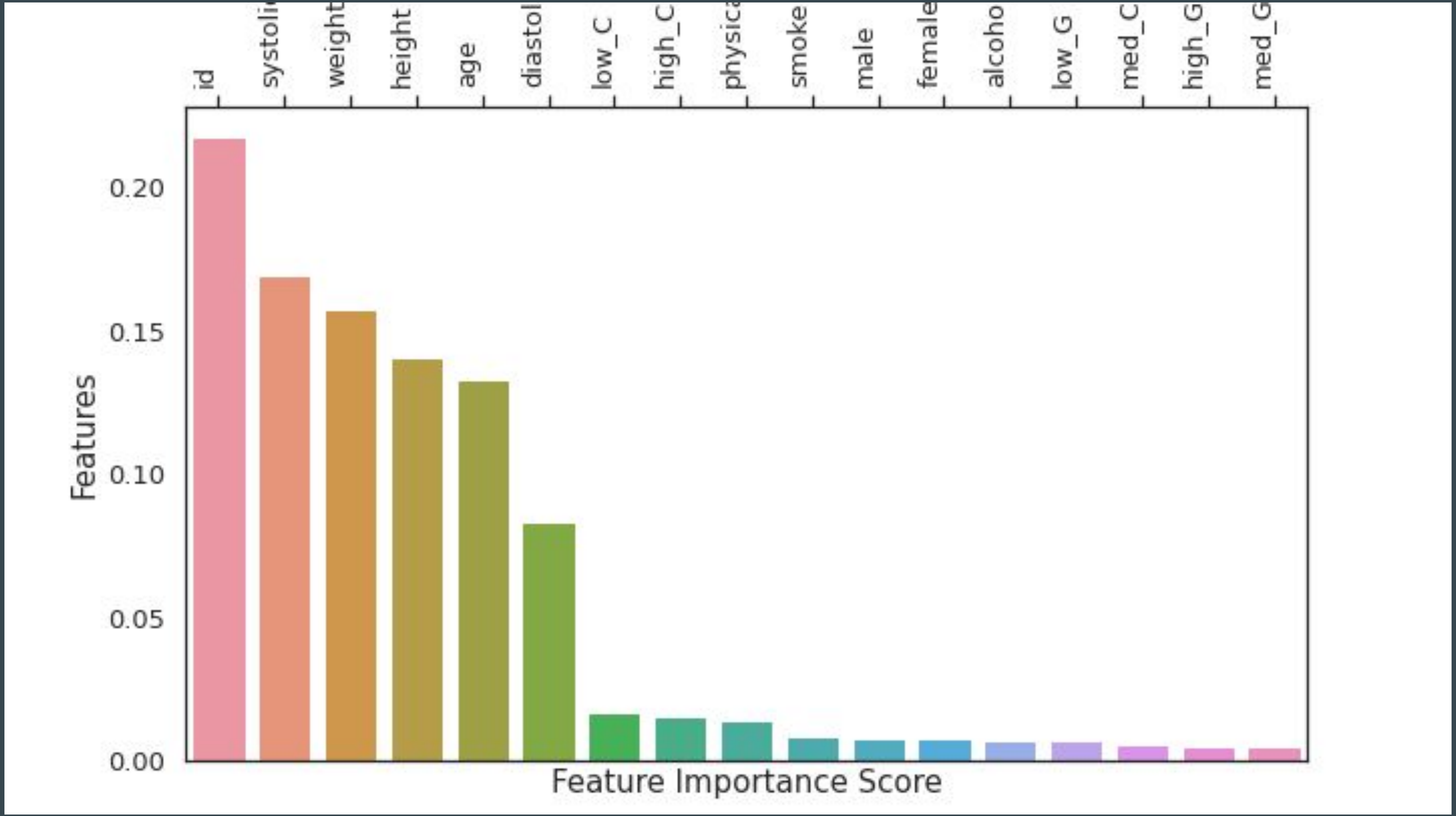
Decision Tree

Cross section of Best tree ->



Random Forest - Feature Importances

Feature Importance ->

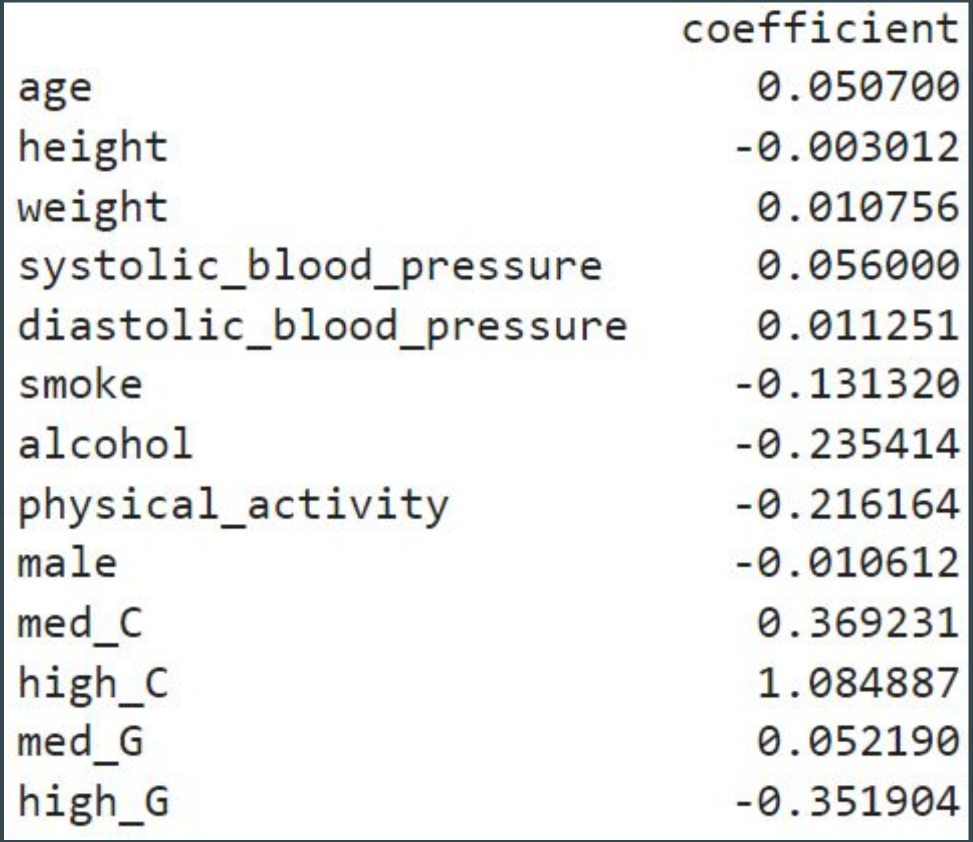


Logistic Regression

- Finds the propensity of a record belonging to a new class and classifies it
- We created m-1 dummy variables to avoid multicollinearity
- The standard model has a classification accuracy of 70.2%
- Lasso Regression Model - 72.97% accuracy
- Ridge Regression Model - 72.95% accuracy
- The above models show the benefits of penalising the predictors

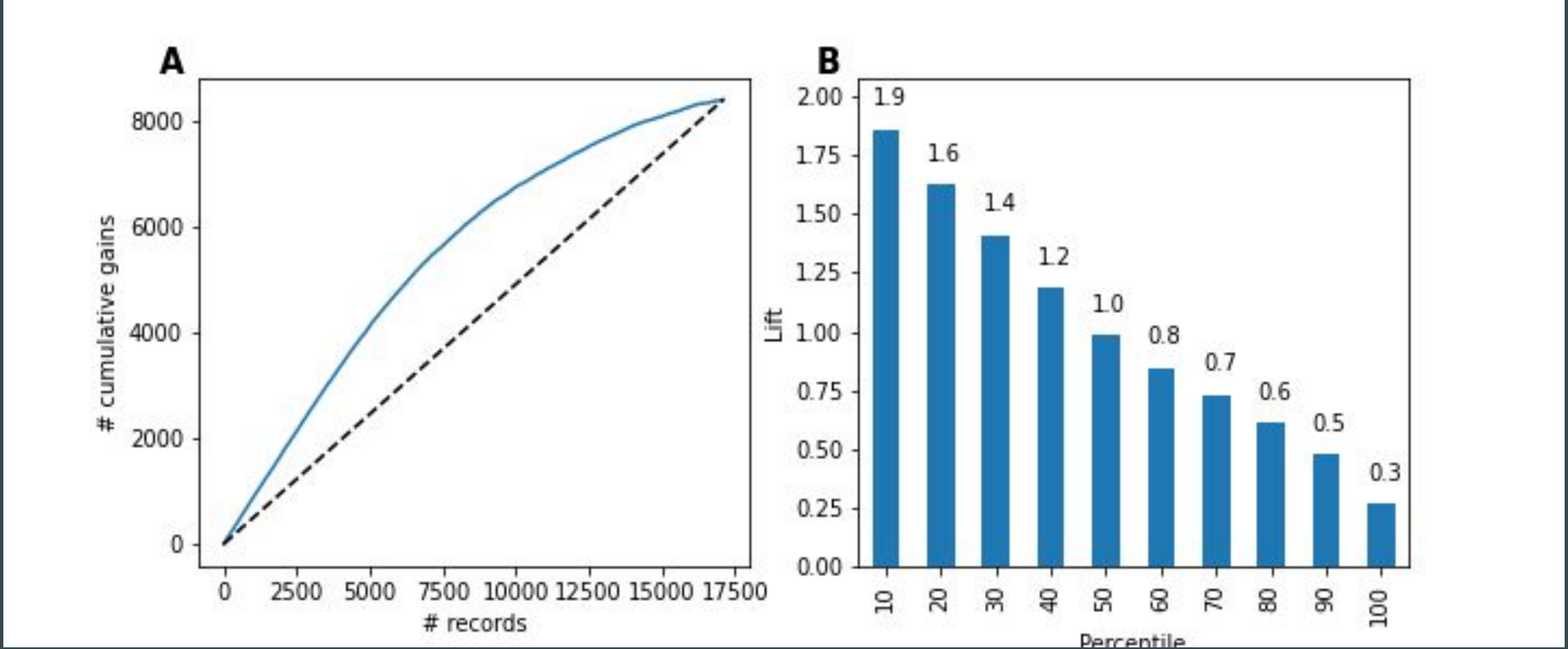
Logistic Regression

Coefficient values ->



```
                          coefficient
age                          0.050700
height                      -0.003012
weight                       0.010756
systolic_blood_pressure      0.056000
diastolic_blood_pressure     0.011251
smoke                       -0.131320
alcohol                     -0.235414
physical_activity           -0.216164
male                        -0.010612
med_C                        0.369231
high_C                       1.084887
med_G                        0.052190
high_G                      -0.351904
```

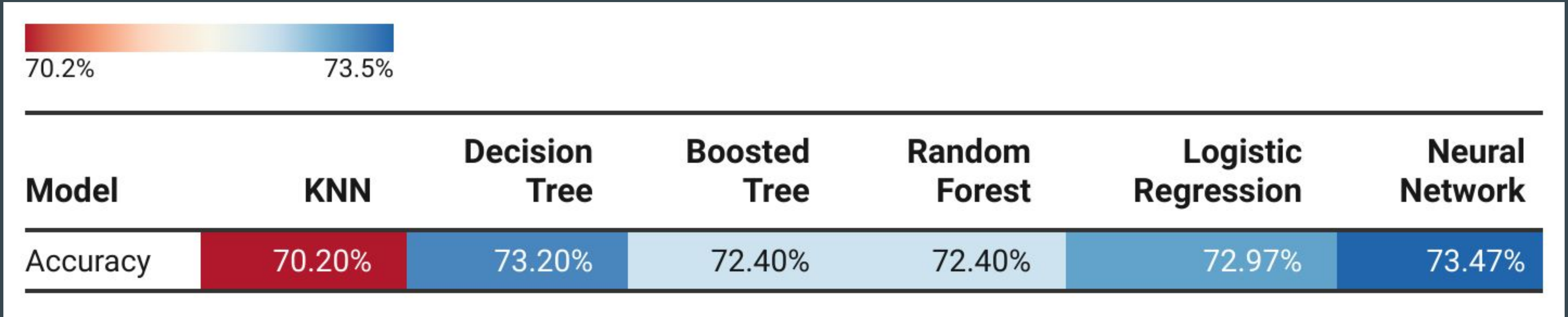
Logistic Regression - Gains and Lift chart



Neural Network

- Imitate brain property to learn underlying patterns in sets of data
- Two hidden layers with five nodes each and other default hyperparameters
- Achieve 73.47% overall accuracy (highest so far)
- Other hyper parameters combination test for a superior predictive performance will be tested in future under the balance of overfitting and underfitting.
- Activation function used was Logistic and solver was lbfgs

Model Performance Comparison and Selection



70.2% 73.5%

Model	KNN	Decision Tree	Boosted Tree	Random Forest	Logistic Regression	Neural Network
Accuracy	70.20%	73.20%	72.40%	72.40%	72.97%	73.47%

Summary

Compared to the KNN, Tree-based method and Logistic Regression, the optimal accuracy of 73.47% is obtained when training a Neural Network under the best parameters, where the parameters used were 2 hidden layers with 5 nodes each and activation function being logistic and solver being lbfgs. The second highest was the Decision tree with an accuracy of 73.2%. Glucose, Cholesterol and Blood pressure have a significant impact on heart disease. Overall, Neural Networks prove to be the best among all which would be used as our standard model for the application that the general population can use.

Acknowledgement

- Thanks to the concept illustration by Prof. Sagar Kamarthi in IE7275
- Thanks to our teaching assistant Sachini for Guiding us
- This was only possible by the good cooperation for both teammates, Hemant and Yanming

References

- Shmueli, Galit, et al. Data mining for business analytics: concepts, techniques, and applications in R. John Wiley & Sons, 2017.
- https://wisdomplexus.com/blogs/data-mining-algorithms-classification
- Stack overflow