STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

CS513 Knowledge Discovery and Data Mining

Cardiovascular Disease Prediction

Team Members

Prashant Mall	Mrunal Salunke	Pallavi Jaiswal	Preet Dabhi
10459371	10467935	10478910	10459151

Problem Statement

- To determine whether or not a person has Cardiovascular Disease:
- An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.
- The provided dataset has a cardio label which indicates 1 for the cardiovascular disease present and 0 for the false report of cardiovascular disease
- After EDA, we pass it through various algorithms for the final prediction of the person suspected to cardiovascular disease on the dataset
- Data Set : https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?select=cardio_train.csv

Data set Description

The Data set (cardio_train.csv) has 70K rows and 13 columns.

The data set has 70K records of patients and 11 features and target.

There are 3 types of input features:

- Objective: factual information
- Examination: results of medical examination
- Subjective: information given by the patient

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Data fields

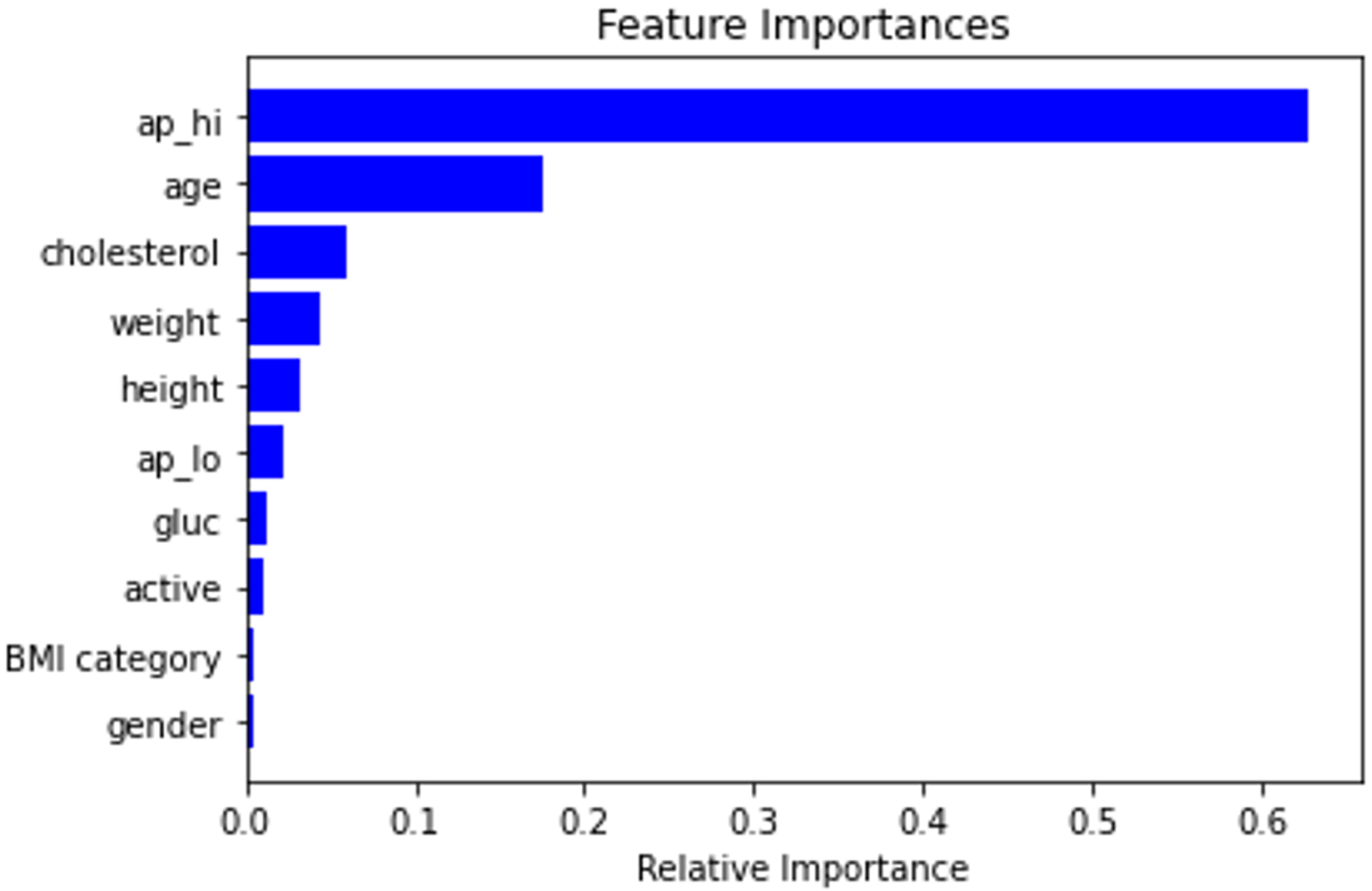
Variable (column name)	Description	Type
age	Age	Objective
height	Height	Objective
weight	Weight	Objective
gender	Gender	Objective
ap_hi	Systolic blood pressure	Examination
ap_lo	Diastolic blood pressure	Examination
cholesterol	Cholesterol 1: normal, 2: above normal, 3: well above normal	Examination
gluc	Glucose 1: normal, 2: above normal, 3: well above normal	Examination
smoke	Smoking	Subjective
alco	Alcohol intake	Subjective
active	Physical activity	Subjective
cardio	Presence or absence of cardiovascular disease	Target

Exploratory Data Analysis

Dataset was clear from all null values

```
id             0.0
age            0.0
gender         0.0
height         0.0
weight         0.0
ap_hi          0.0
ap_lo          0.0
cholesterol    0.0
gluc           0.0
smoke          0.0
alco           0.0
active         0.0
cardio         0.0
dtype: float64
```

Feature Importance

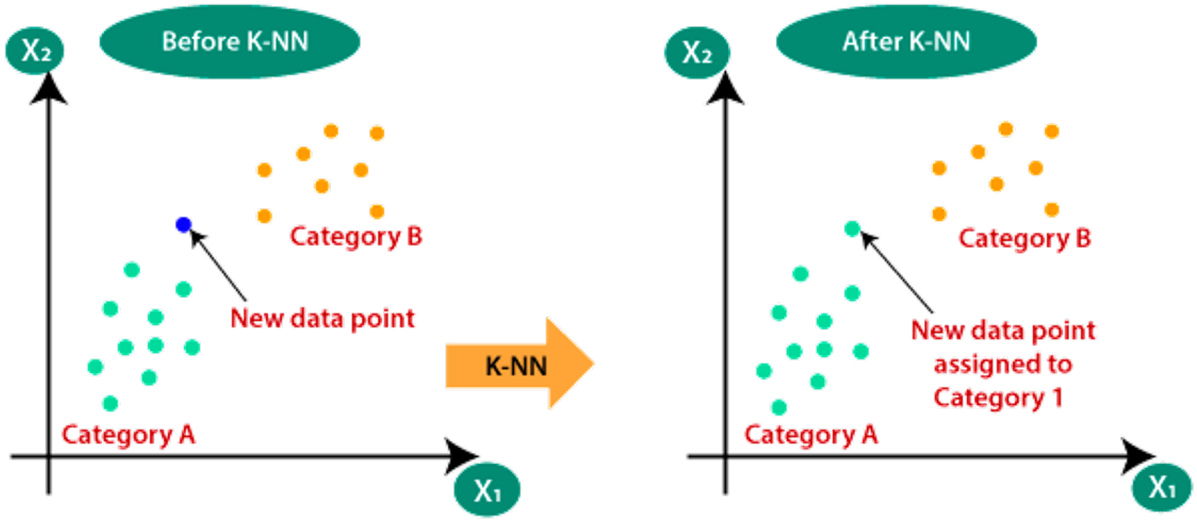


Models Used

- KNN
- CART
- Naïve Bayes
- Logistic Regression
- Random Forest

K nearest neighbors (KNN)

- Supervised machine learning algorithm
- Assumes that similar things exist in close proximity
- To select the value of k, we run algorithm several times with different values of k and choose the k that reduces the number of errors
- Relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data.
- Used for classification but can be used to estimation and prediction tasks

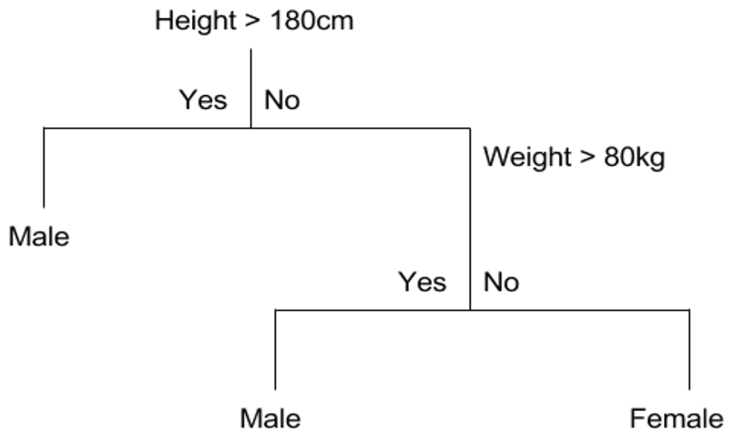


K nearest neighbors (KNN)

	precision	recall	f1-score	support
0	0.66	0.71	0.68	5514
1	0.69	0.64	0.66	5588
accuracy			0.67	11102
macro avg	0.67	0.67	0.67	11102
weighted avg	0.67	0.67	0.67	11102

Classification and Regression Trees (CART)

- CART is a tree-based classification and predictive method.
- It uses recursive partitioning to split the training records into segments with similar output field values.
- Each root node represents a single input variable (x) and a split point on that variable
- The tree's leaf nodes contain an output variable (y) which is used to make the prediction.



Classification and Regression Trees (CART)

	precision	recall	f1-score	support
0	0.62	0.63	0.62	5514
1	0.63	0.62	0.62	5588
accuracy			0.62	11102
macro avg	0.62	0.62	0.62	11102
weighted avg	0.62	0.62	0.62	11102

Naïve Bayes

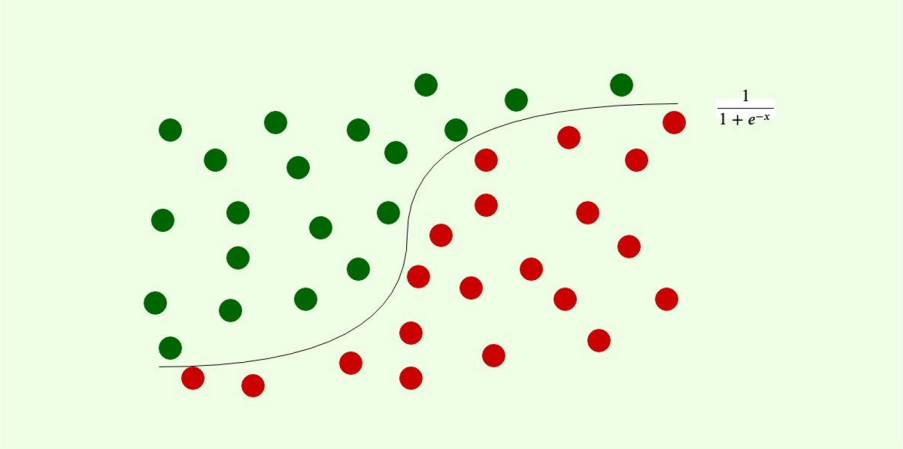
- Supervised machine learning algorithm
- Primarily used for classification
- Bayes Theorem is used to solve classification problems by following a probabilistic approach, it finds the probability of an event occurring given the probability of another event that has already occurred.
- All variables are independent of each other

Naïve Bayes

	precision	recall	f1-score	support
0	0.55	0.88	0.68	5514
1	0.72	0.29	0.42	5588
accuracy			0.59	11102
macro avg	0.63	0.59	0.55	11102
weighted avg	0.64	0.59	0.55	11102

Logistic regression

- Logistic Regression is a process of modelling the probability of a discrete outcome given an input variable.
- It predicts by analyzing the relationship between one or more existing variables.

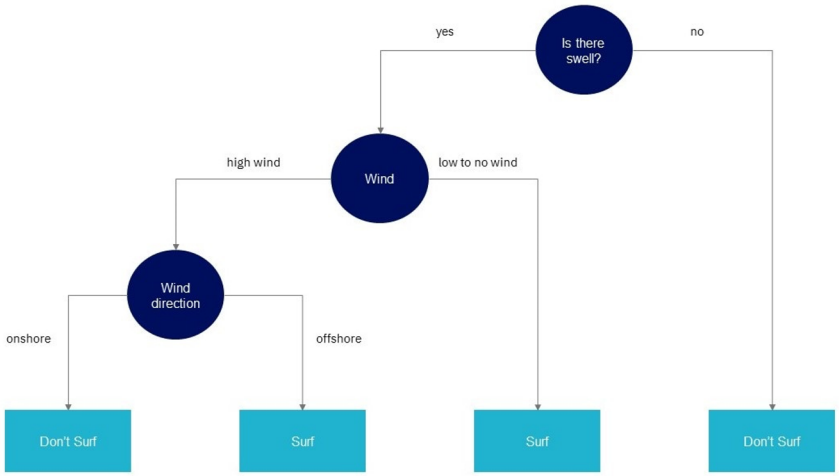


Logistic regression

	precision	recall	f1-score	support
0	0.67	0.75	0.71	5514
1	0.72	0.64	0.68	5588
accuracy			0.70	11102
macro avg	0.70	0.70	0.69	11102
weighted avg	0.70	0.70	0.69	11102

Random Forest Classifier

- Random Forest is a supervised Machine Learning Algorithm that is used in Classification and Regression problems. It is one of the most used algorithms due to its simplicity and diversity.
- Random Forest is made of multiple decisions trees. These questions make up the decision nodes in the tree, acting as a means to split the data.
- Each question helps an individual to arrive at the final decisions, which would be denoted as the leaf node.



Random Forest Classifier

	precision	recall	f1-score	support
0	0.70	0.73	0.71	5514
1	0.72	0.69	0.70	5588
accuracy			0.71	11102
macro avg	0.71	0.71	0.71	11102
weighted avg	0.71	0.71	0.71	11102

Comparison and Conclusion

Here we are comparing the different values of precision , recall and F1 score of various algorithms to determine which of the them gives the best accuracy.

Algorithms	Precision	Recall	F1-Score
KNN	0.69	0.64	0.66
CART	0.63	0.62	0.62
Naive Bayes	0.72	0.29	0.42
Logistic Regression	0.72	0.64	0.68
Random Forest	0.72	0.69	0.70

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Thank You