AMRITA VISHWA VIDYAPEETHAM

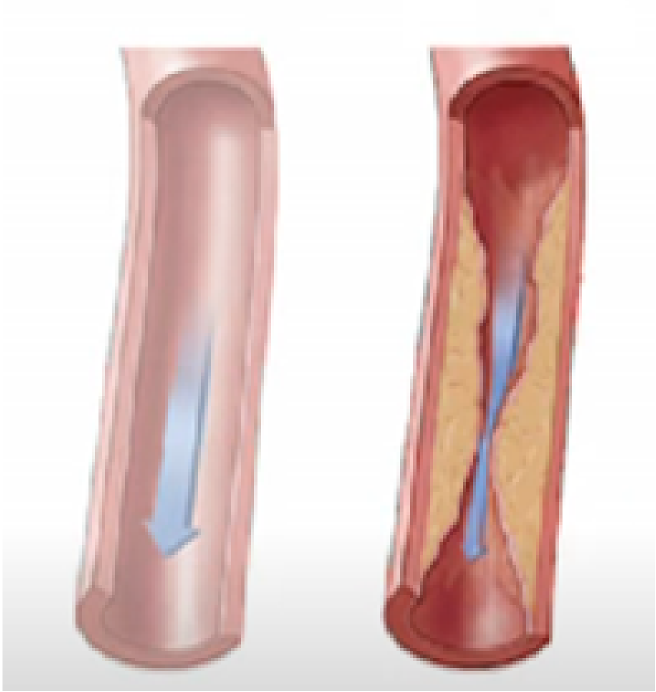
19AIE205 - PML
End Semester Project
Cardiovascular Disease Prediction

Mentor: Dr. Kumaran U
(Department of Computer Science and Engineering)

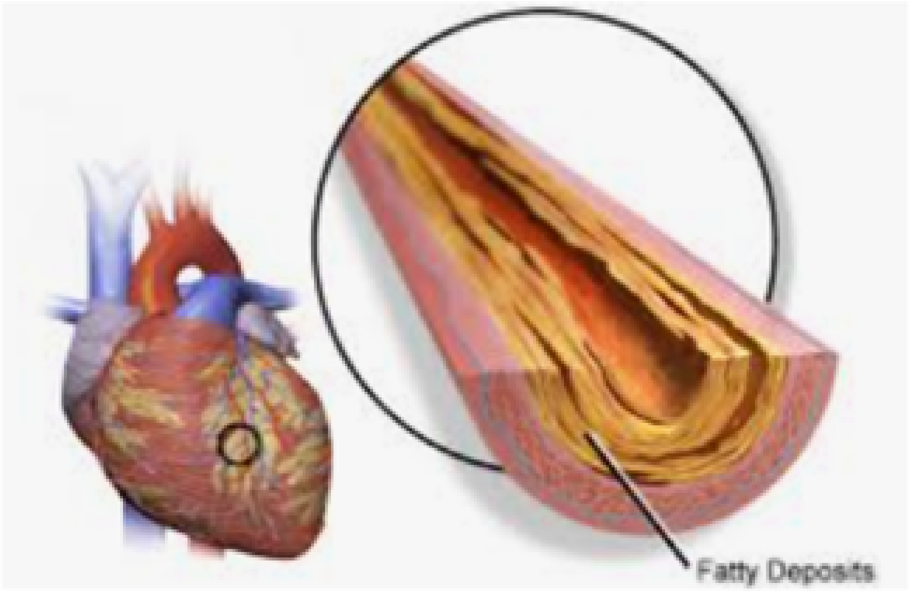
BY: TEAM OUTLIERS
Apoorva M
Bhuvanashree Murugadoss
Sai Nikhilesh Reddy

Introduction to Cardiovascular Diseases:

Cardiovascular disease involves abnormalities of the heart and the blood vessels. The lifetime risk of developing significant cardiovascular disease is greater than the likelihood of developing cancer



Our heart is a muscular pump that requires heart arteries to supply oxygen rich blood to keep it going. Coronary heart disease occurs when these blood vessels become narrowed due to a buildup of plaque



Objective of our Project:

What we aim to do is create an efficient classification model that best predicts if a given patient is likely to develop cardio vascular disease or not.

Overview of our approach towards the problem:

1. Perform Exploratory Data Analysis to gain insights on the Data
   a. Data Cleaning
   b. Data Summarization: Describe the data and its distributions
   c. Data Visualization: Create graphical summaries of the data

2. Transform the data for training the different classification algorithms

3. Apply the different Machine Learning Algorithms
   a. Train, Cross validate, perform appropriate hyperparameter tuning
   b. Comparative analysis of the accuracy of the models

4. Gain insights from the results obtained

Dataset Description:

The dataset consists of 70,000 records of patients data, with 11 features and 1 target variable

There are 3 types of input features:

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

Features:

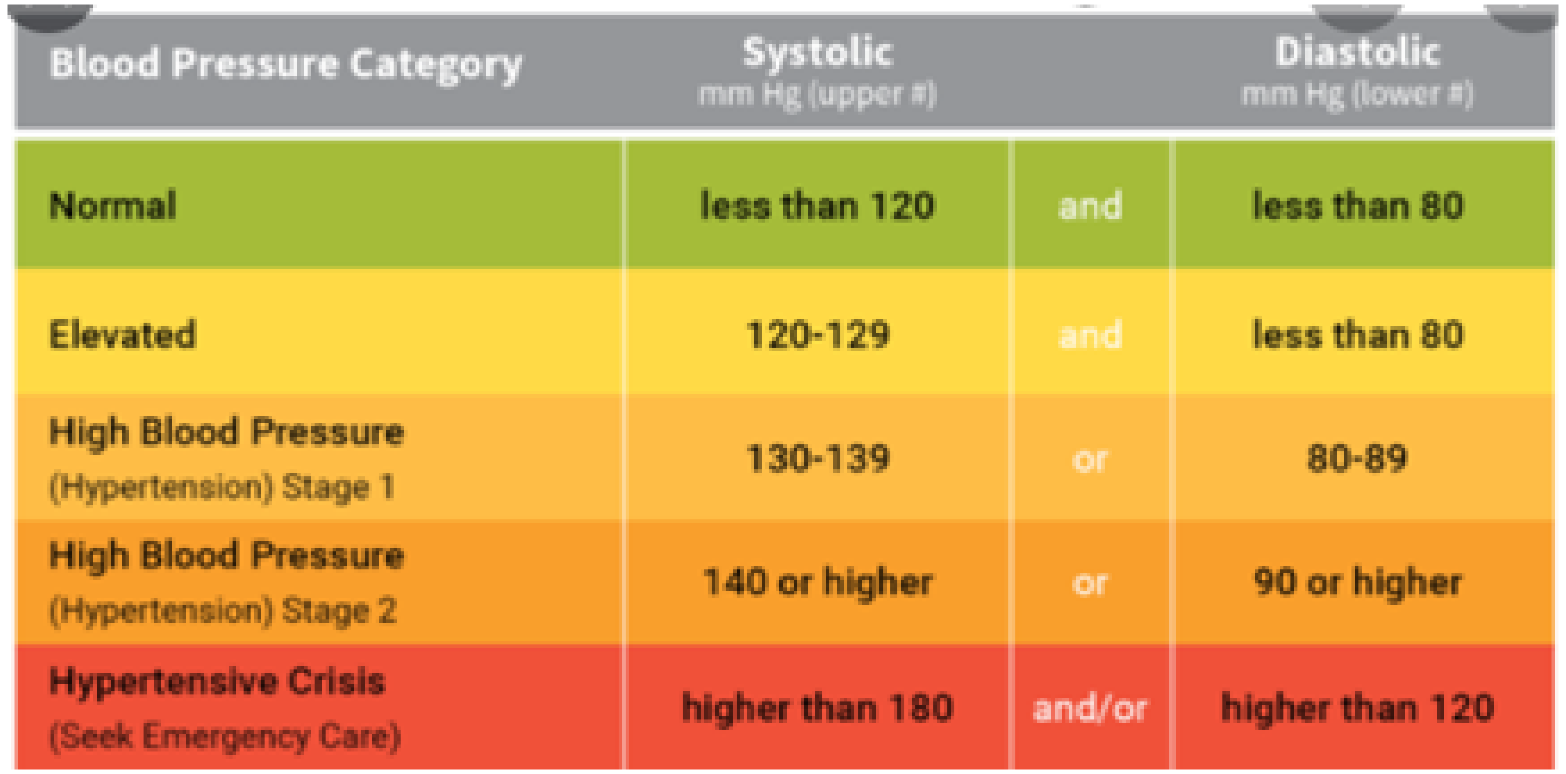
1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |

-Systolic blood pressure, the top number, measures the force your heart exerts on the walls of your arteries each time it beats

6. Diastolic blood pressure | Examination Feature | ap_lo | int

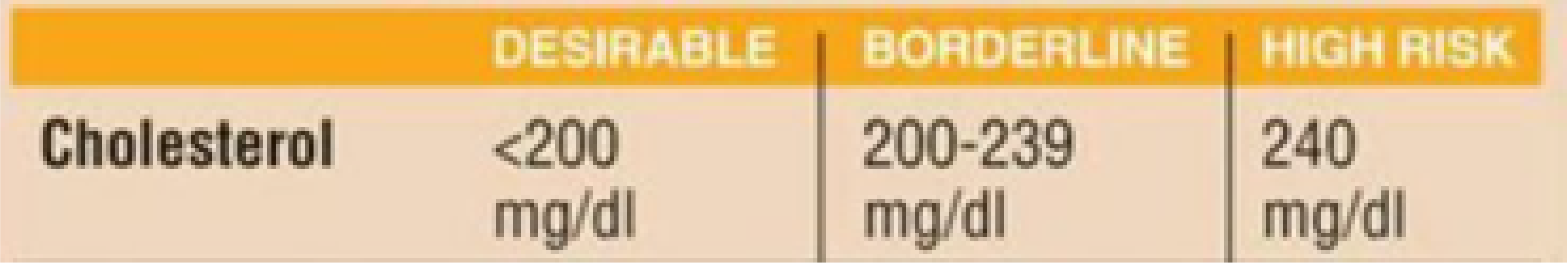
-Diastolic blood pressure, the bottom number, measures the force your heart exerts on the walls of your arteries in between beats.

Dataset Description:



Blood Pressure Category	Systolic mm Hg (upper #)		Diastolic mm Hg (lower #)
Normal	less than 120	and	less than 80
Elevated	120-129	and	less than 80
High Blood Pressure (Hypertension) Stage 1	130-139	or	80-89
High Blood Pressure (Hypertension) Stage 2	140 or higher	or	90 or higher
Hypertensive Crisis (Seek Emergency Care)	higher than 180	and/or	higher than 120

7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |

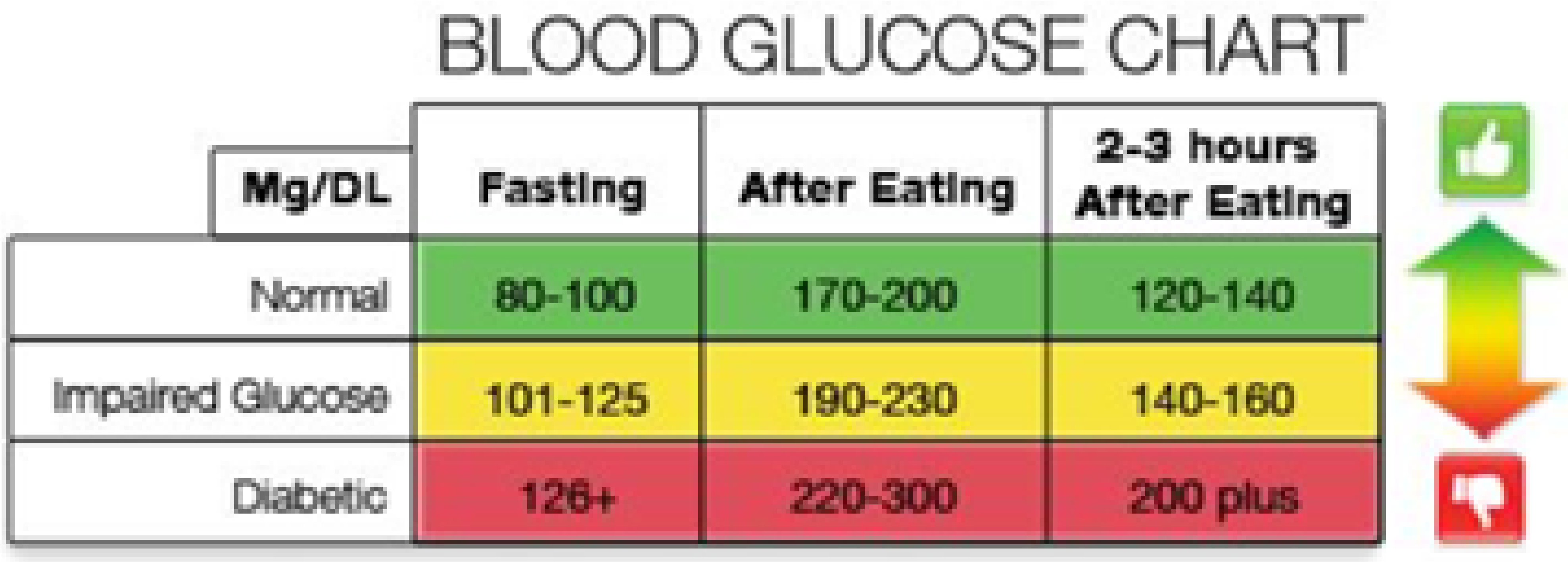


	DESIRABLE	BORDERLINE	HIGH RISK
Cholesterol	<200 mg/dl	200-239 mg/dl	240 mg/dl

Dataset Description:

8. Glucose | Examination Feature | gluc | 1 normal, 2: above normal, 3: well above normal

The blood glucose level is the amount of glucose in the blood. Glucose is a sugar that comes from the foods we eat, and it's also formed and stored inside the body. It's the main source of energy for the cells of our body, and it's carried to each cell through the bloodstream.



BLOOD GLUCOSE CHART

Mg/DL	Fasting	After Eating	2-3 hours After Eating
Normal	80-100	170-200	120-140
Impaired Glucose	101-125	190-230	140-160
Diabetic	126+	220-300	200 plus

Dataset Description:

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 |

    All of the dataset values were collected at the moment of medical examination.

Exploratory Data Analysis:

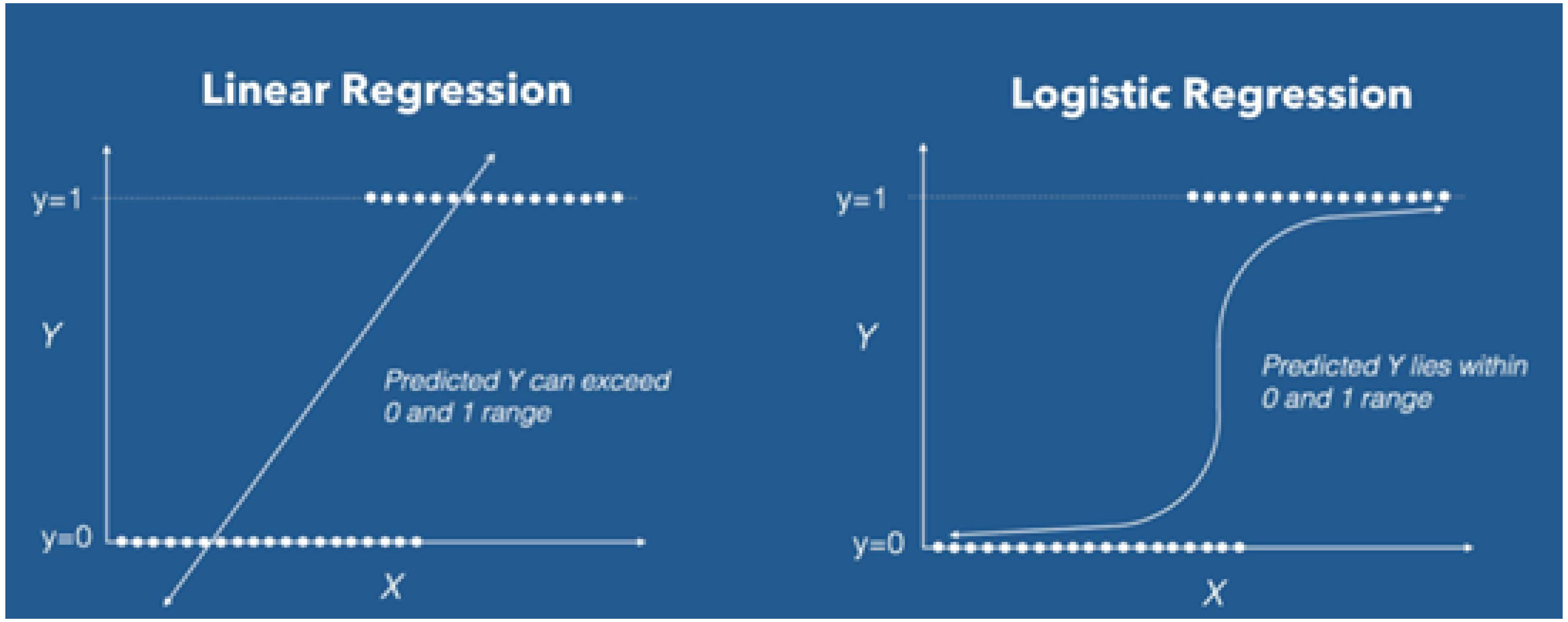
a. Data Cleaning
b. Data Summarization: Describe the data and its distributions
c. Data Visualization: Create graphical summaries of the data

Transform the data for training the different classification algorithms

Moving to the Jupyter Notebook for further explanation

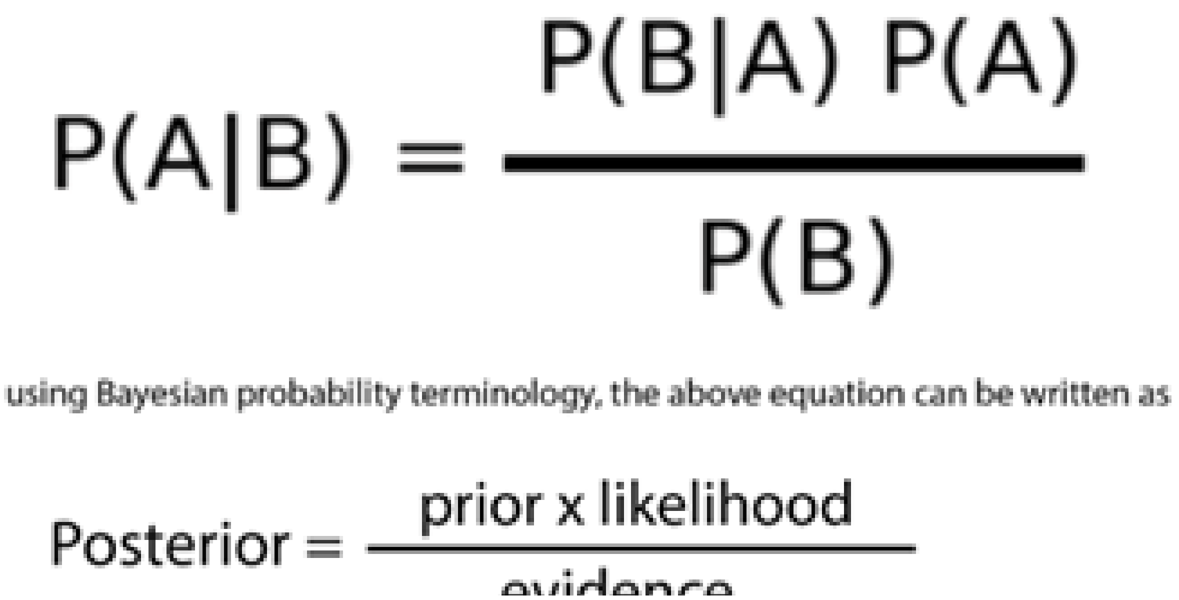
Logistic Regression

Unlike linear regression the response variables can be categorical or continuous, as the model does not strictly require continuous data. To predict group membership, LR uses the log odds ratio rather than probabilities and an iterative maximum likelihood method rather than a least squares to fit the final model.


Linear Regression
y=1
Y
y=0
X
Predicted Y can exceed 0 and 1 range
Logistic Regression
y=1
Y
y=0
X
Predicted Y lies within 0 and 1 range

Naive Bayes Classifier

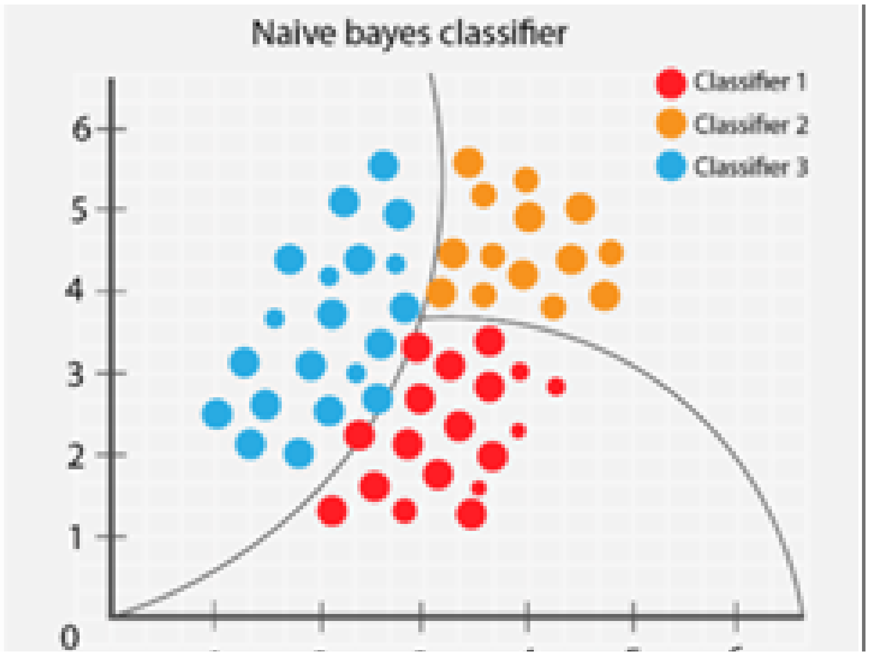
Naive Bayes classifiers calculate the probability of a sample to be of a certain category, based on prior knowledge. They use the Naive Bayes Theorem, that assumes that the effect of a certain feature of a sample is independent of the other features. That means that each character of a sample contributes independently to determine the probability of the classification of that sample, outputtingthe category of the highest probability of the sample.



$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

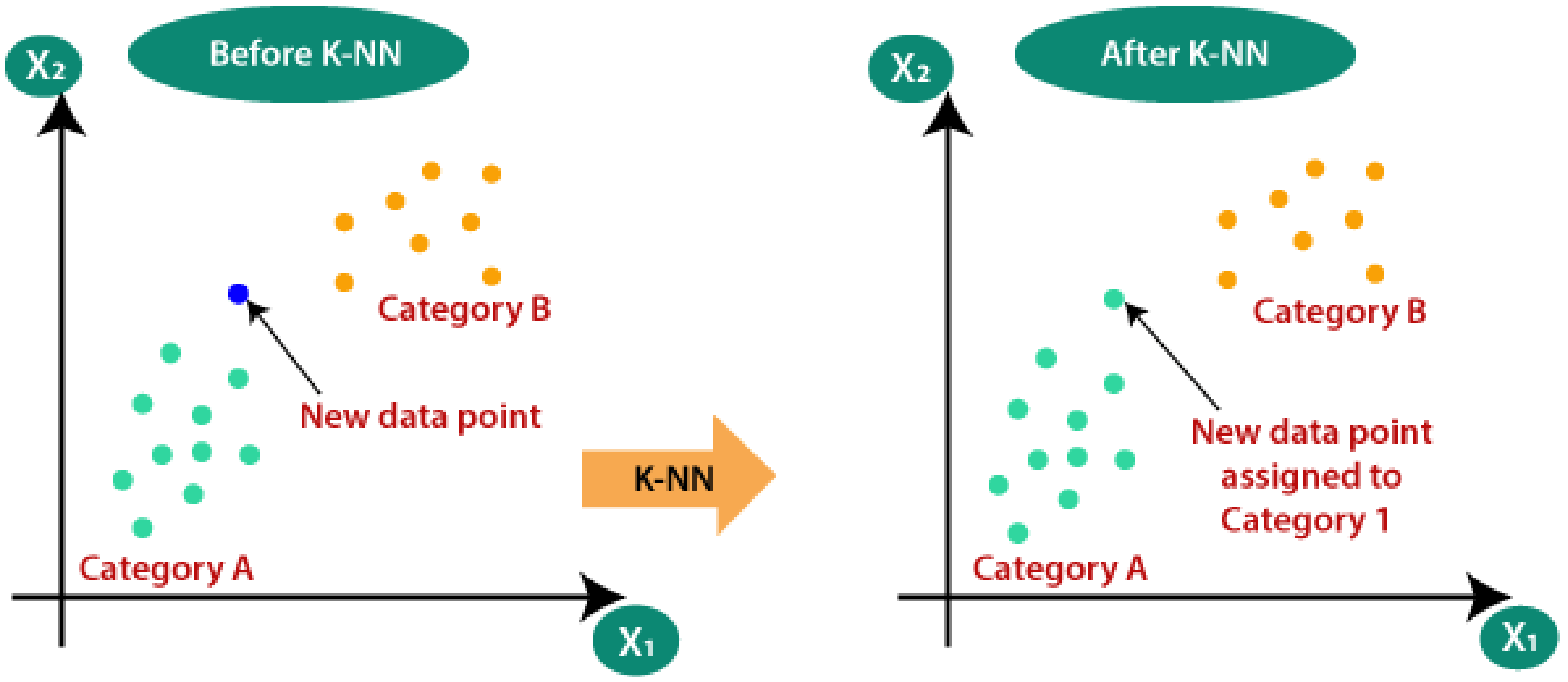
using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$



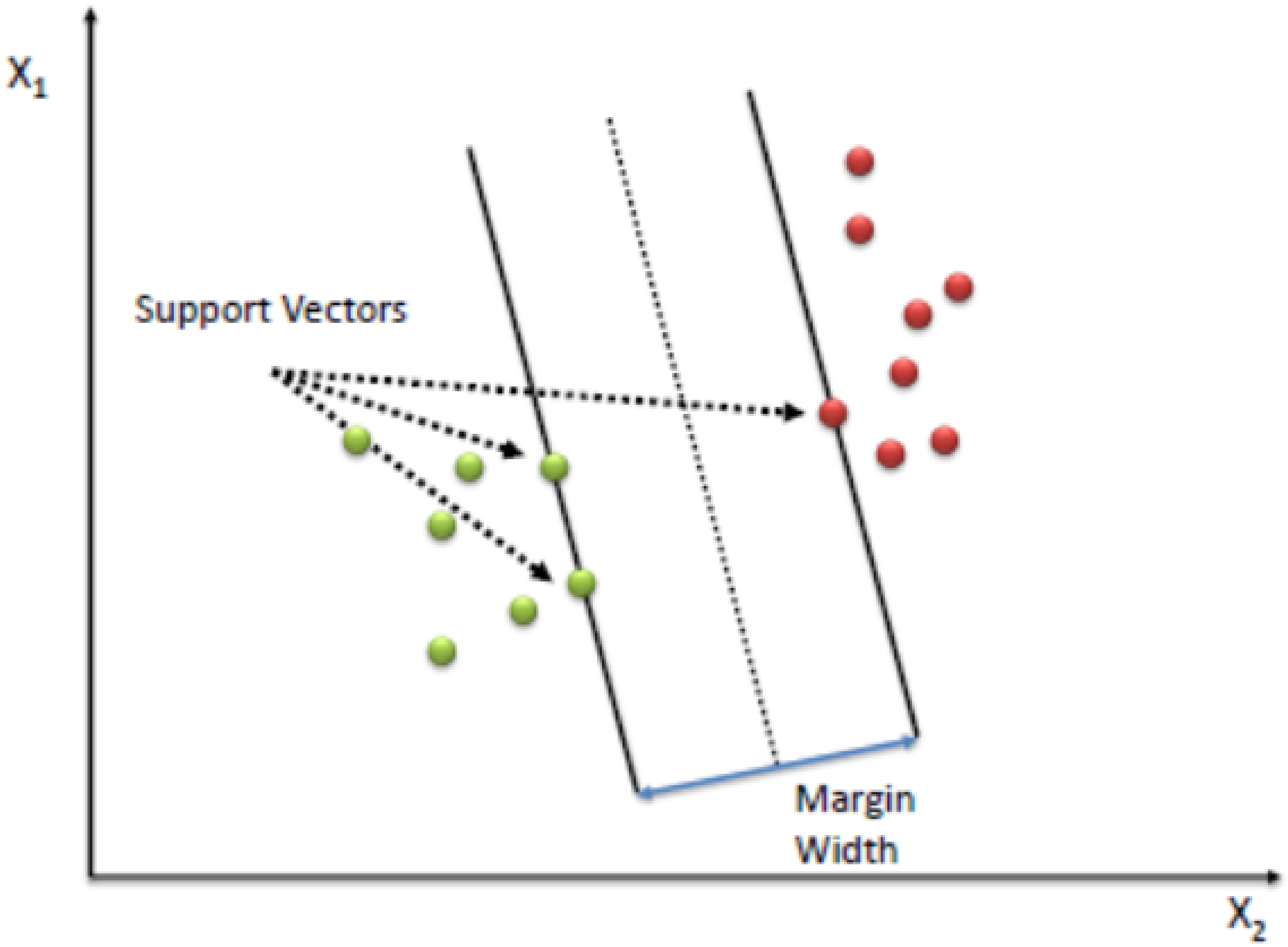
K Nearest Neighbors Algorithm

the k-nearest neighbors algorithm (k-NN) is a non-parametric method proposed by Thomas Cover used for classification and regression



Support Vector Machines

A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.



Decision Tree Classifier

Decision Tree algorithm is a supervised learning algorithm that can be used for solving regression and classification problems. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

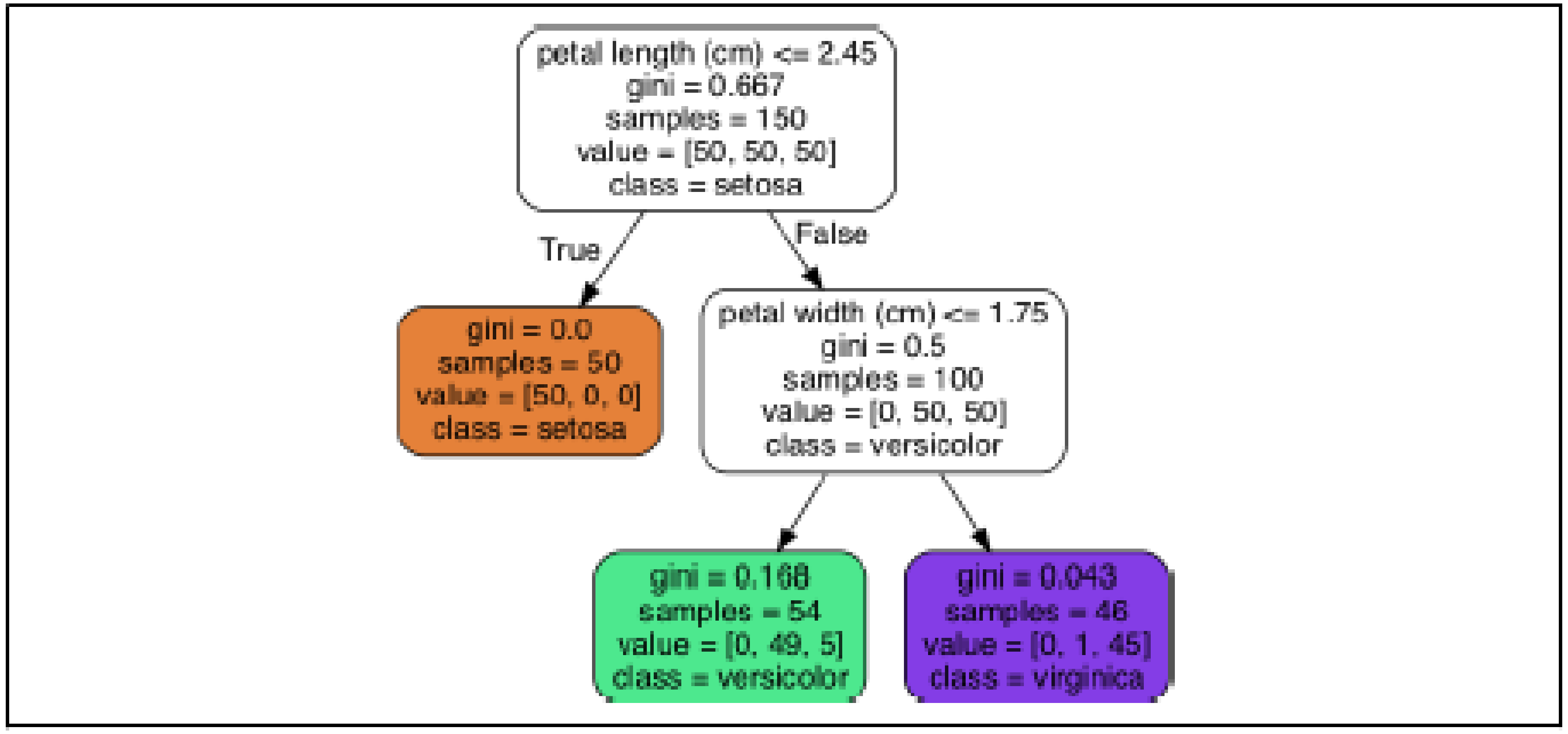
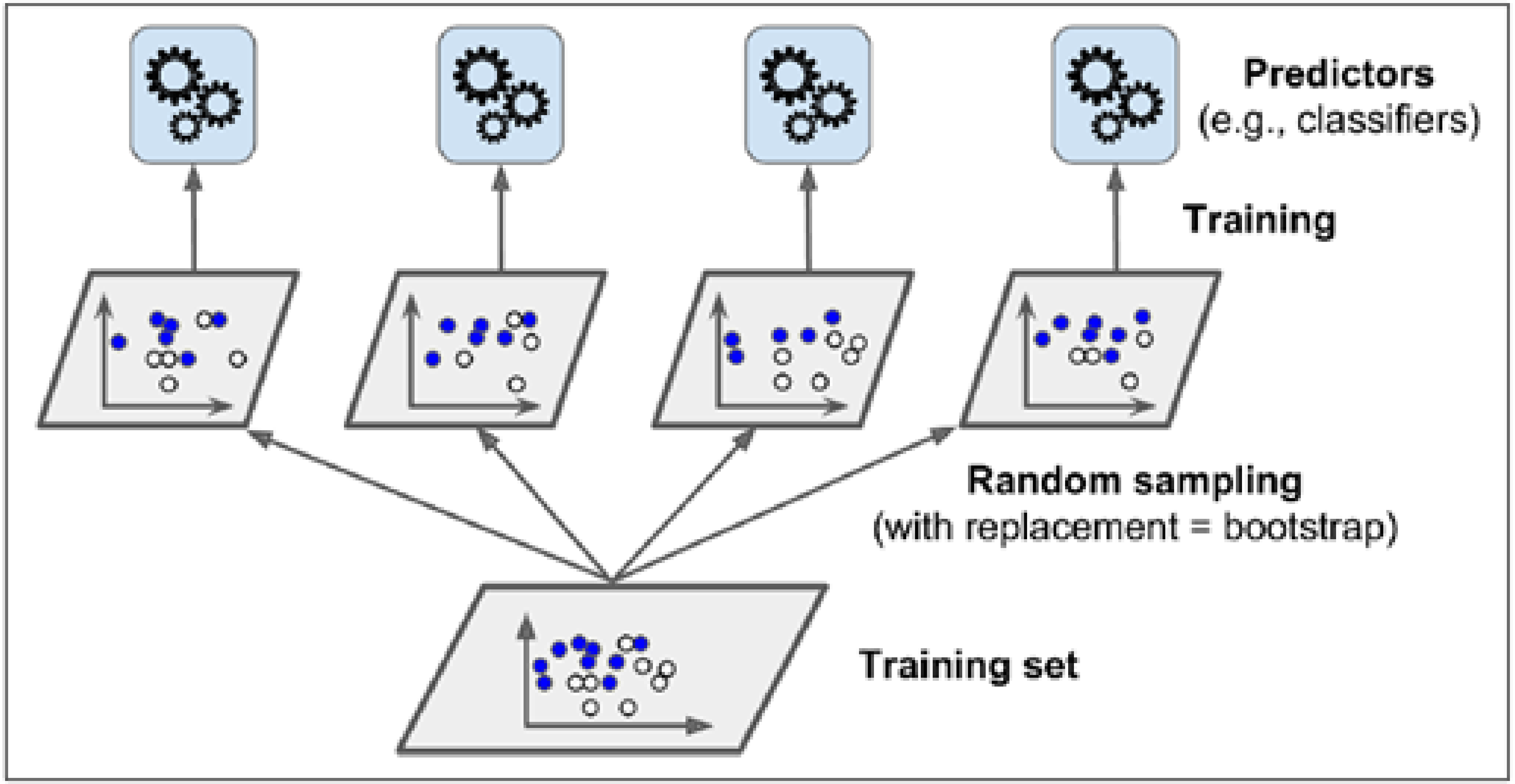
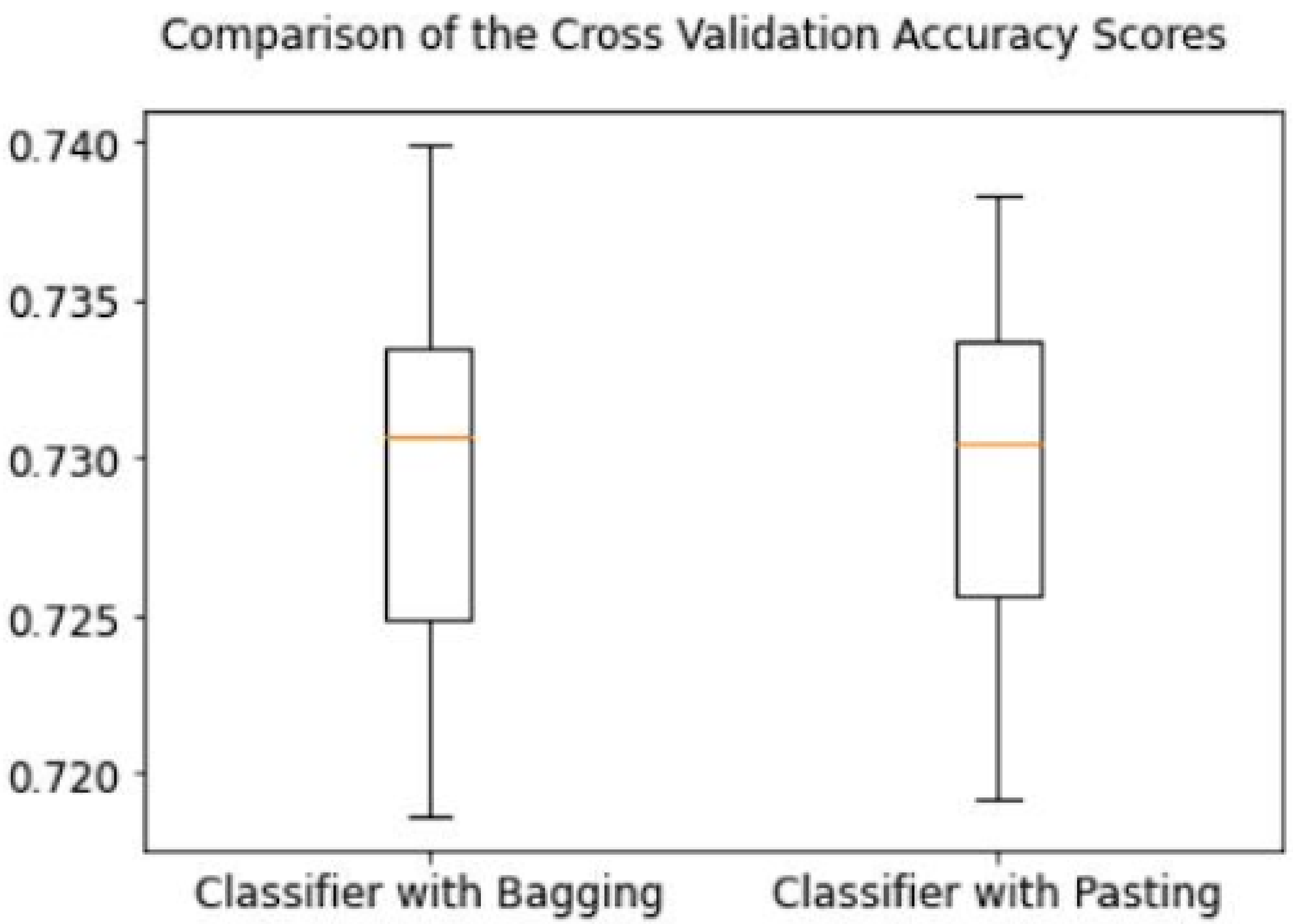


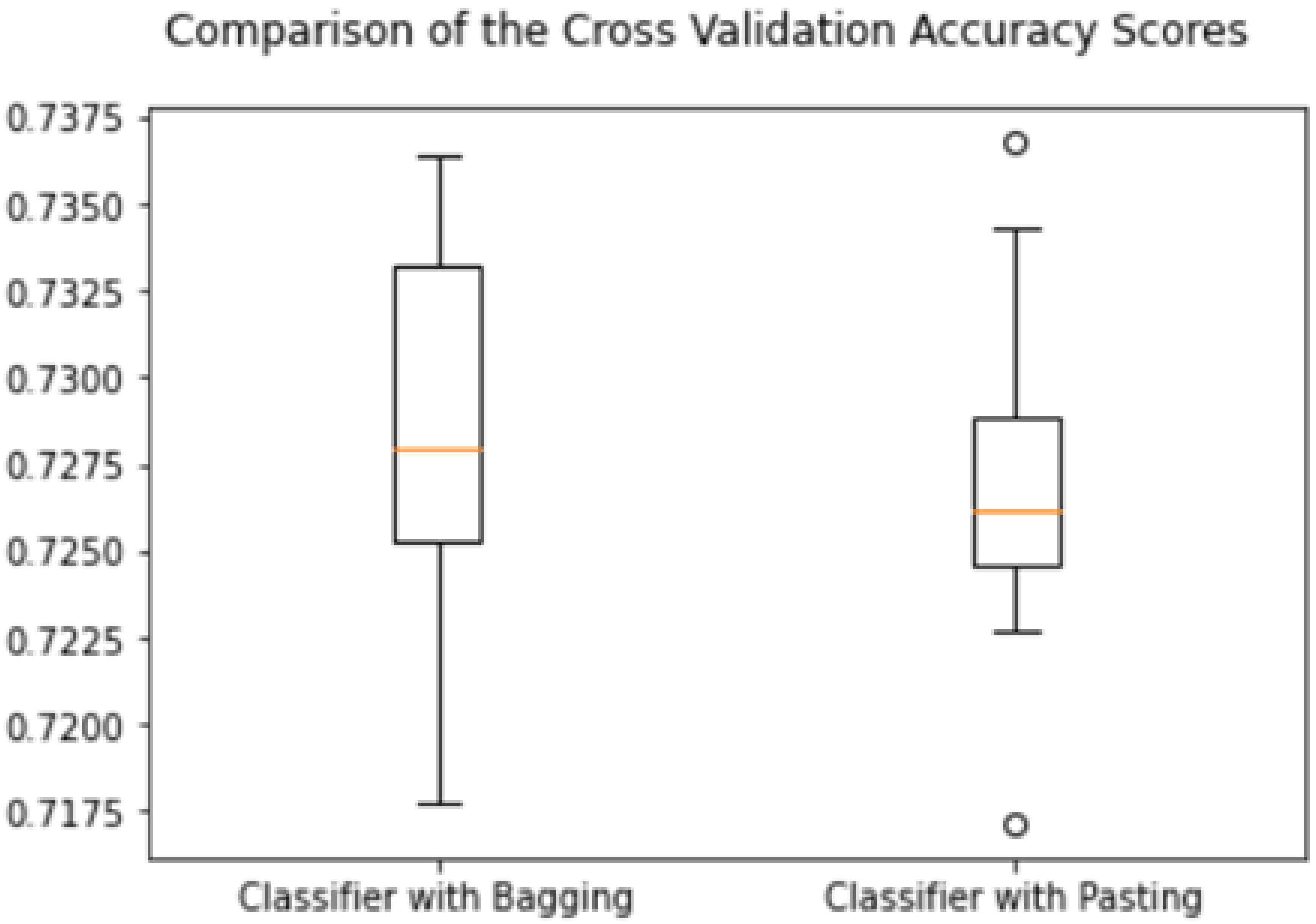
Figure 6-1. Iris Decision Tree

Ensemble Methods for Classification:



Comparison of Bagging and Pasting:





Comparison of Hard and Soft Voting:

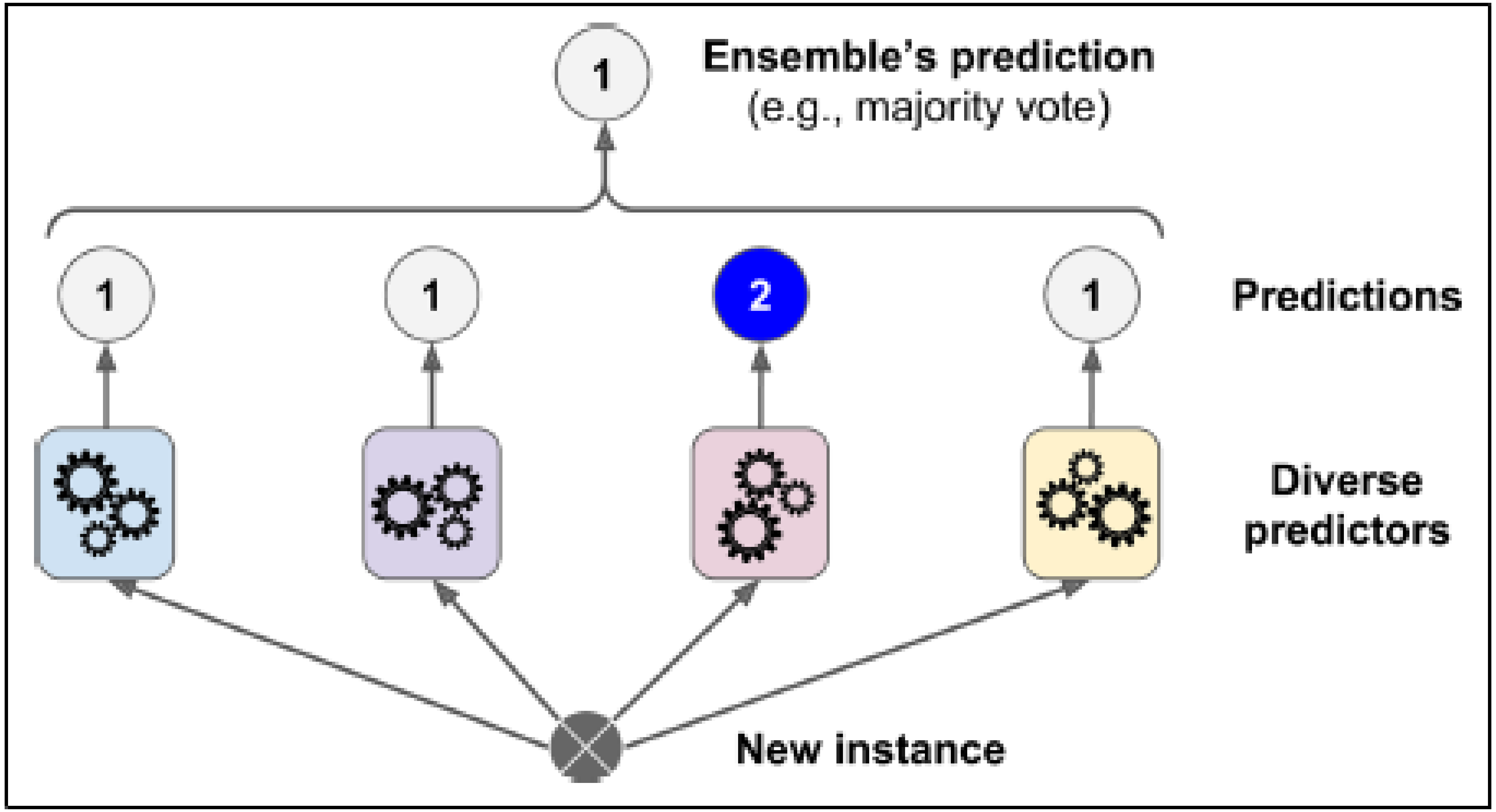
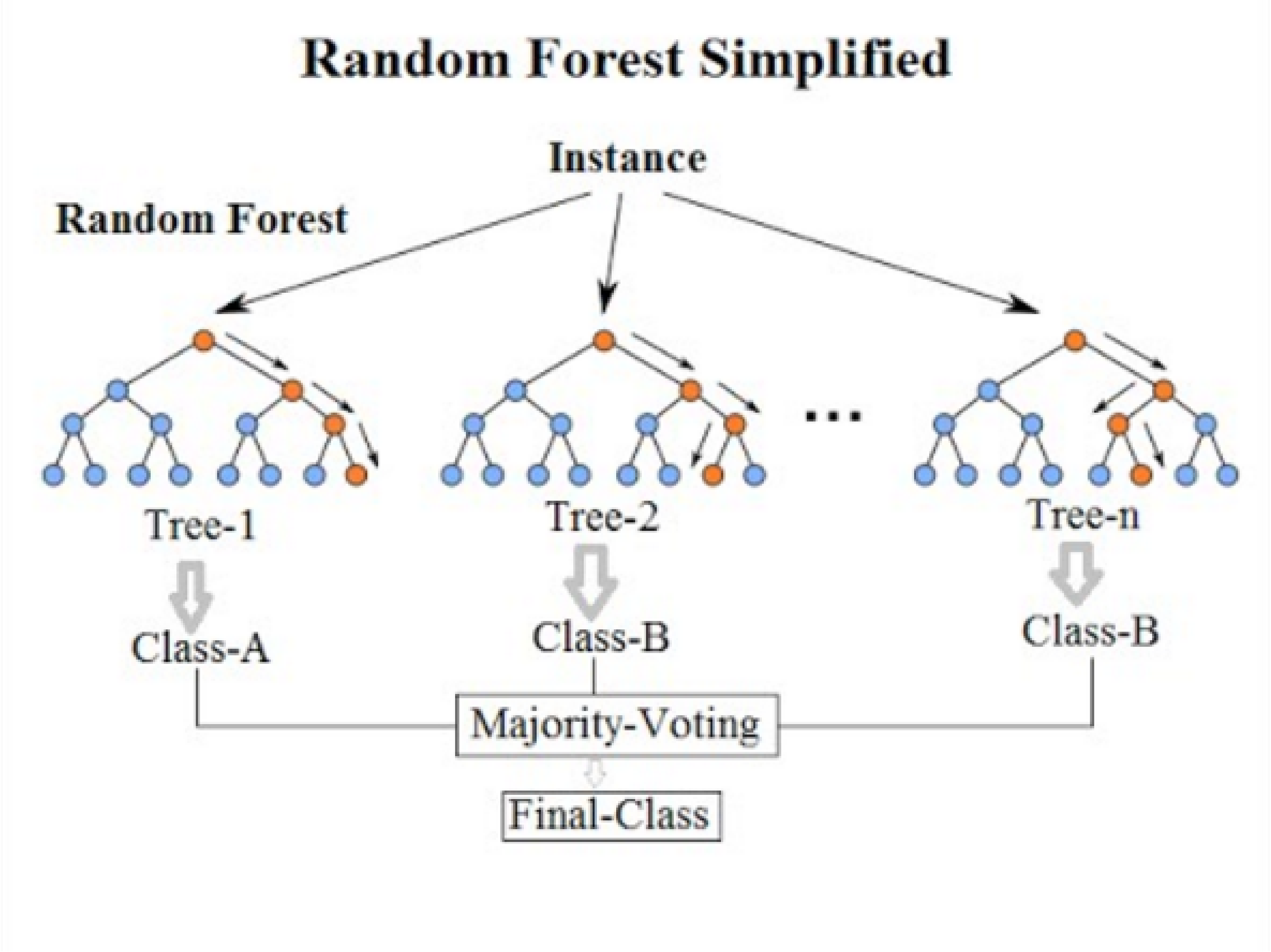


Figure 7-2. Hard voting classifier predictions

Random Forest Classifier:


Random Forest Simplified
Instance
Random Forest
...
Tree-1
Tree-2
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class