SDAIA
الهيئة السعودية للبيانات
والذكاء الاصطناعي
Saudi Data & AI Authority

أكاديمية طويق
TUWAIQ ACADEMY

PROJECT 4

Iris Dataset

Mohammed Alageel
Mazen Alamri

TABLE OF CONTENT

01. DATA DESCRIPTION

02. DATA PREPROCESSING

03. UNSUPERVISED LEARNING

04. SUPERVISED LEARNING

05. MODEL COMPARISON

06. MODEL TUNING

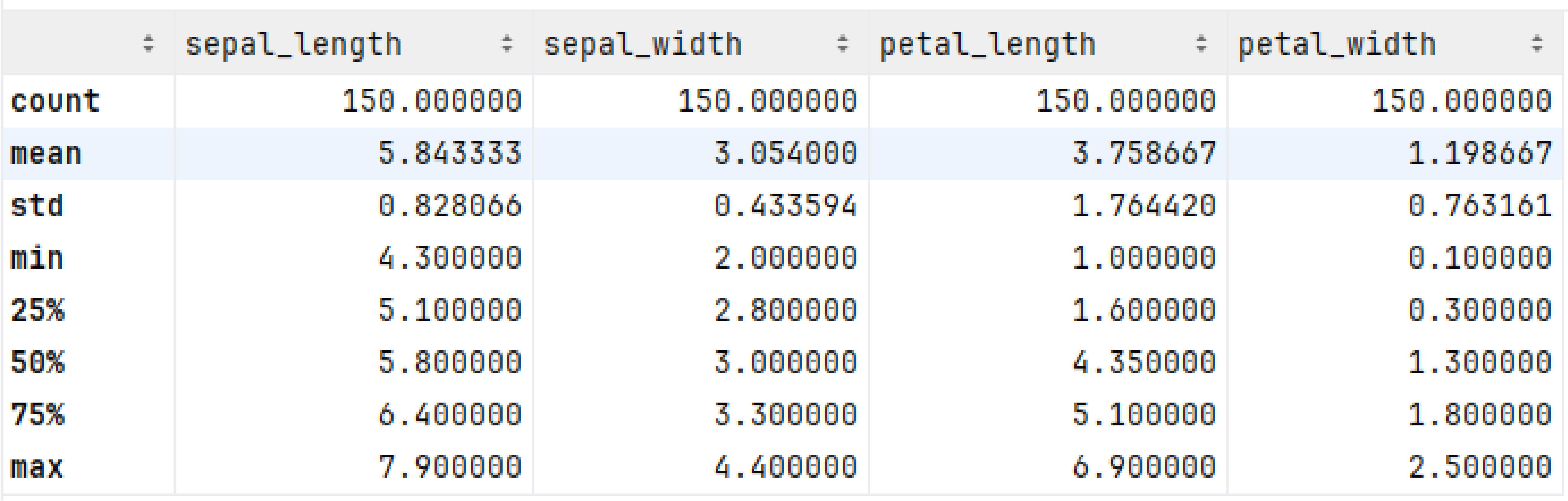
DATA DESCRIPTION

The Iris dataset is a popular dataset in data science, it consists of 4 Features: Sepal length, Sepal width, Petal length and Petal width (all in cm), and it has one target variable: the Species

DATA EXPLORATION

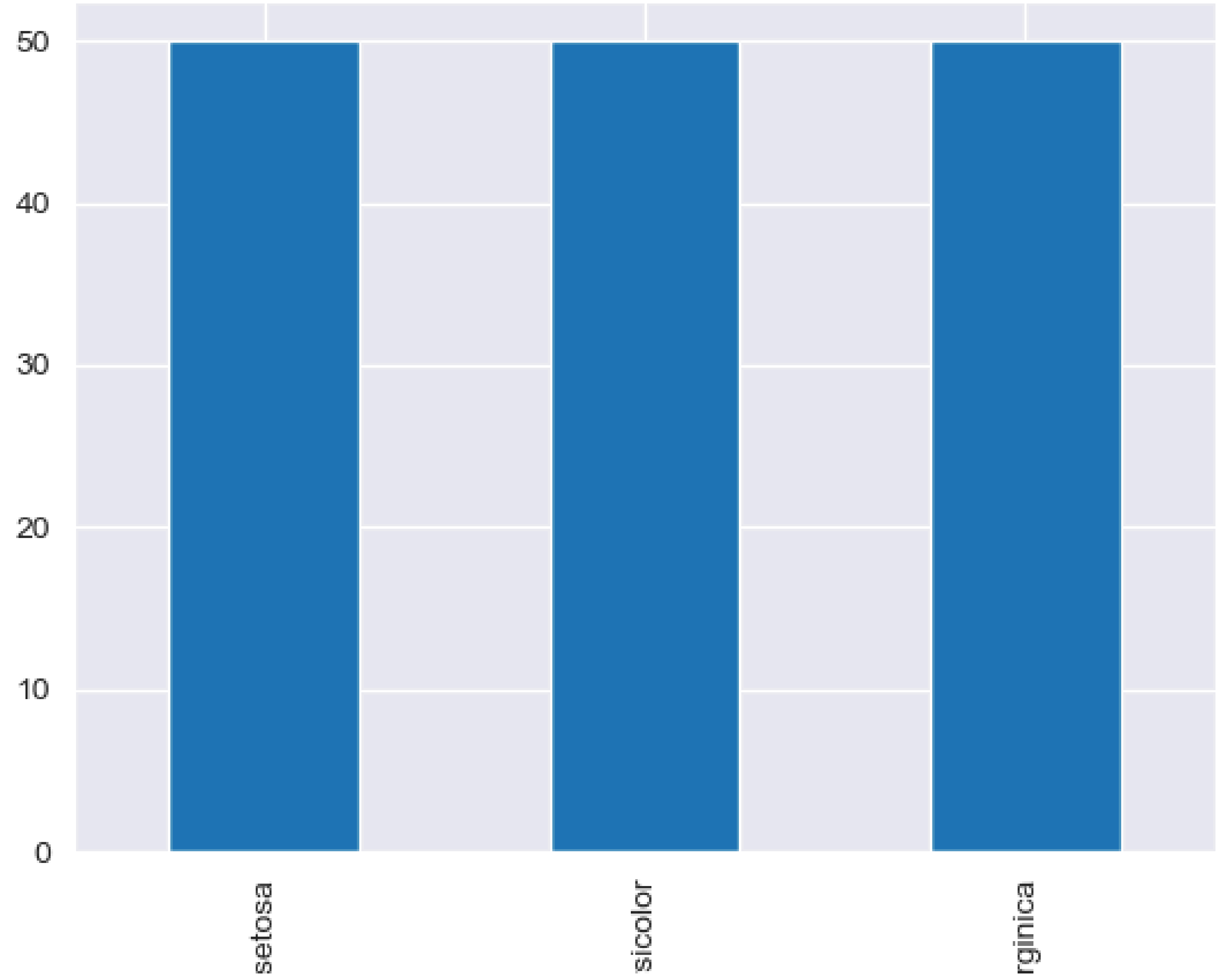
- The dataset containts 150 rows and 5 columns.
- The dataset had no null values.
- We explored descriptive statistics of each numeric feature.
- We split the data to x (4 features) and y (the target).

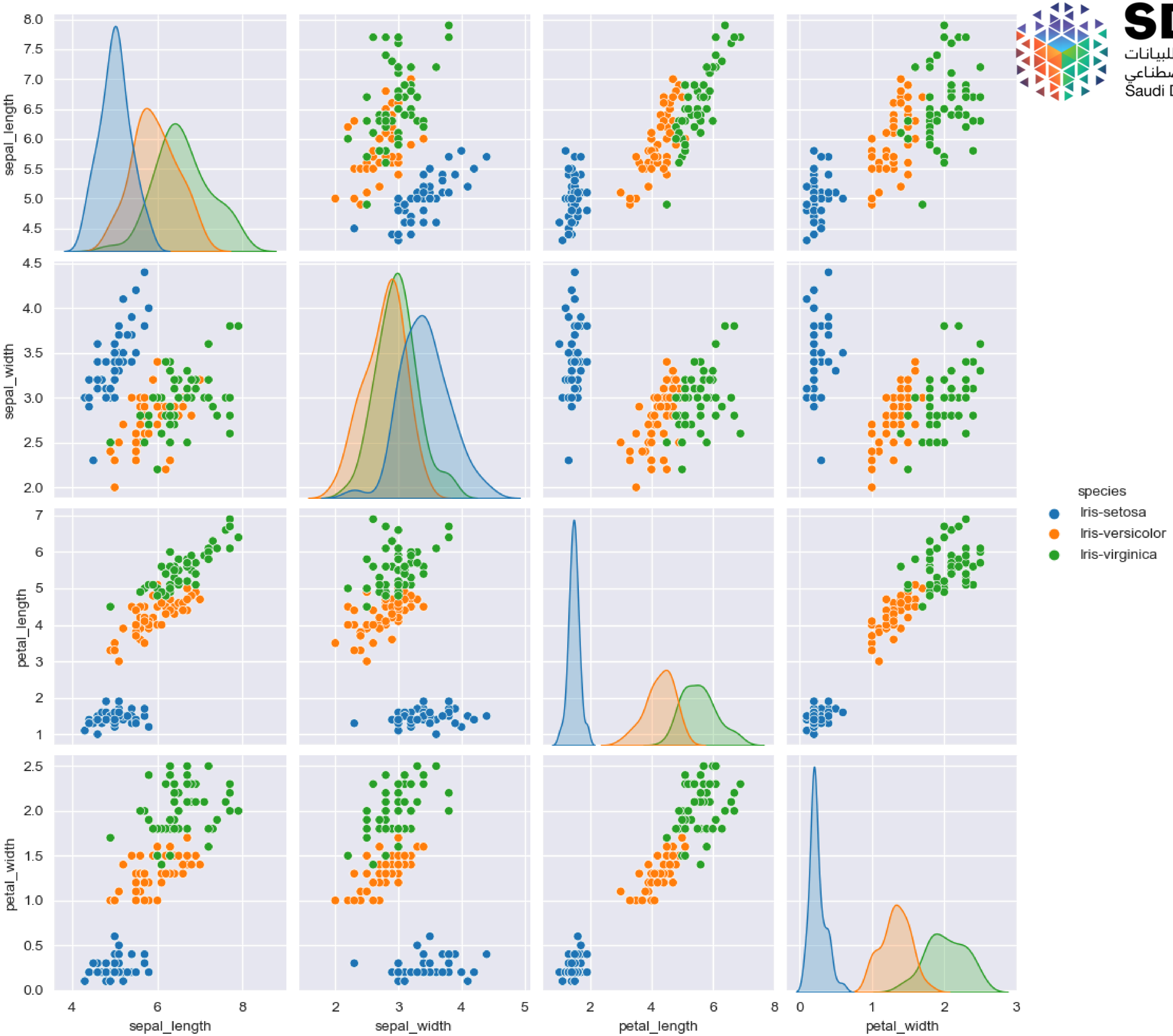
DESCRIPTIVE STATISTICS



	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

THE DISTRIBUTION OF THE 3 SPECIES



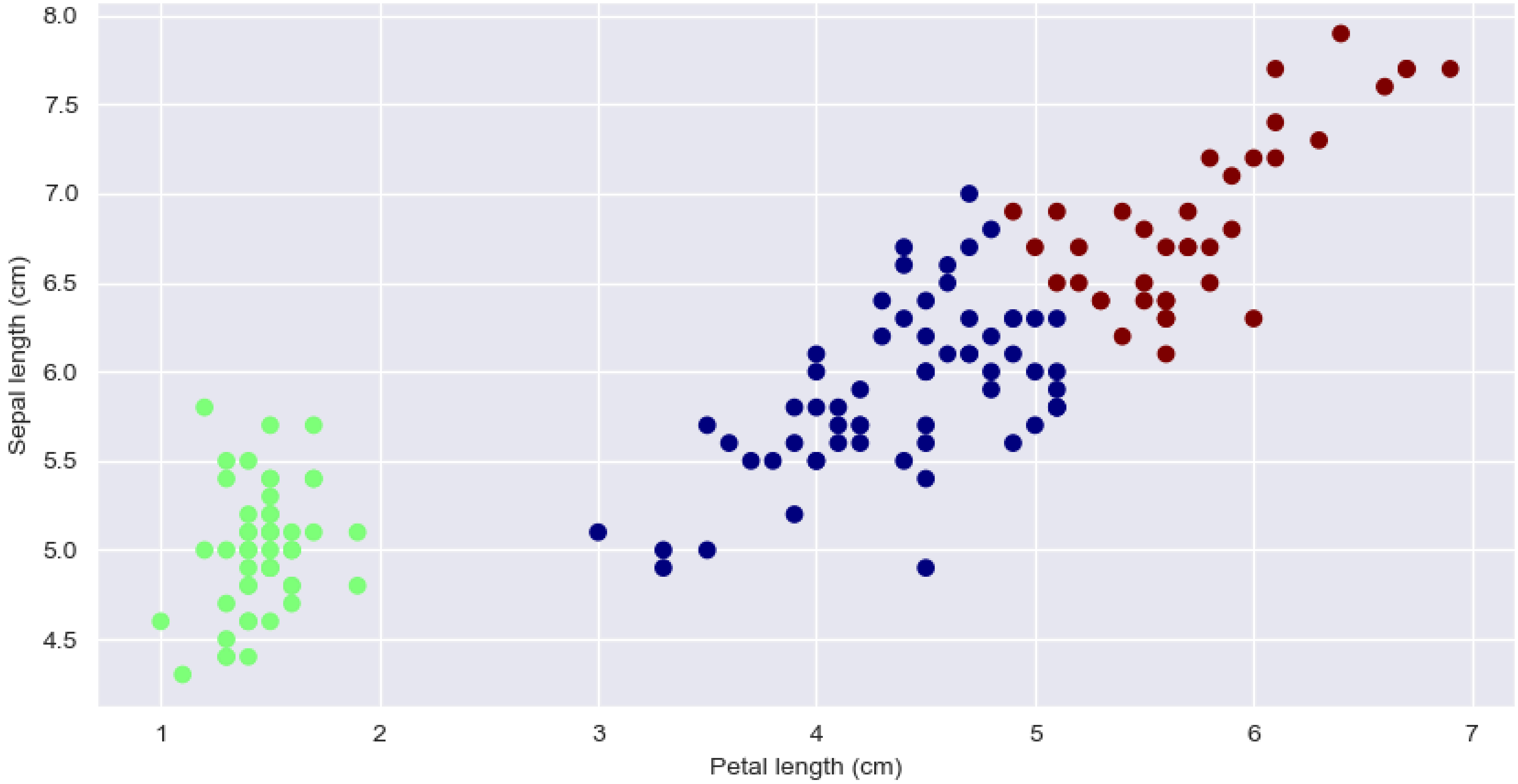

sepal_length
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
sepal_width
4.5
4.0
3.5
3.0
2.5
2.0
petal_length
7
6
5
4
3
2
1
petal_width
2.5
2.0
1.5
1.0
0.5
0.0
4
6
8
sepal_length
2
3
4
5
sepal_width
2
4
6
8
petal_length
0
1
2
3
petal_width
species
Iris-setosa
Iris-versicolor
Iris-virginica

02.

UNSUPERVISED LEARNING

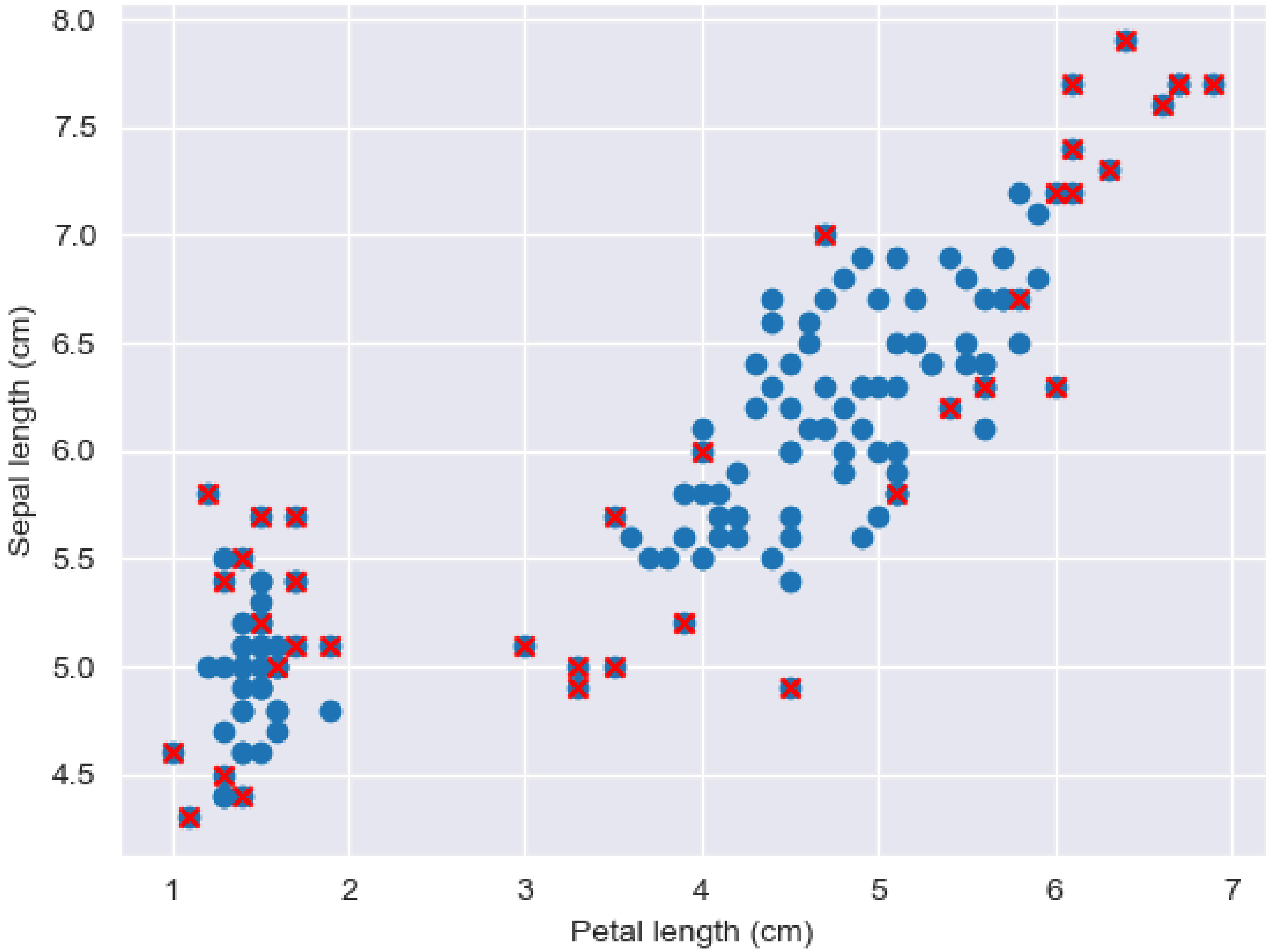
K-MEANS CLUSTERING

- We used K-means clustering to cluster the data.
- We can see how easily K-means clustered the data correctly assuming we know the correct k value, but we can use the elbow method to find the best k value.



ISOLATION FOREST

- We used Isolation Forest to find outliers in the data.
- The Isolation Forest found outliers but it failed to recognize that the third species is not an outlier.



03.

SUPERVISED LEARNING

SUPERVISED LEARNING

- we chose F1 score (micro) as an evaluation method for the classifier.
- split the data to training and testing sets.
- we chose Logistic Regression as a baseline classifier and it managed to reach an F1 score of 96.6%.

04.

MODEL COMPARISON

MODEL COMPARISON

- We chose the following models: SVC, RandomForestClassifier, GradientBoostingClassifier, AdaBoostClassifier, CategoricalNB.
- Out of the chosen models the best ones were LogisticRegression, SVC, RandomForestClassifier.

04.

MODEL TUNING

MODEL TUNING

- Preformed hyperparameter on logistic regression using grid search.
- We tuned the model on the best hyperparameters and it reached an F1 score of 97.5%.
- Implemented an ensemble (Voting ensemble) of a group of the top performing models and used bagging on each one and got 94.1% as an F1 score.

THANK YOU