Weekend Masters in Applied Statistics and Data Science

Department of Statistics
Jahangirnagar University
Savar, Dhaka-1342

A report on

Predicting species of Iris flowers by using machine learning

Course Title: Introduction to Data Science with Python
Course Code:WM-ASDS04

Submitted to

Farhana Afrin Duty
Assistant Professor
Department of Statistics
Jahangirnagar University

Submitted by

Sl	Name	Roll	Section	Batch
1	Mahbubur Rahaman	20229024	A	9th
2	Abdus Salam Sarkar	20229032	A	9th

Introduction:

Machine learning is a field of computer science and artificial intelligence that focuses on developing algorithms and models that enable computers to learn from and make predictions or decisions based on data. It involves using statistical and mathematical techniques to automatically detect patterns in data and use those patterns to make predictions or decisions without being explicitly programmed to do so. It can be classified as Supervised learning, unsupervised learning and reinforcement learning. If the output value of a supervised learning is categorical then that is called classification and if output value is numeric value then that will be called regression.

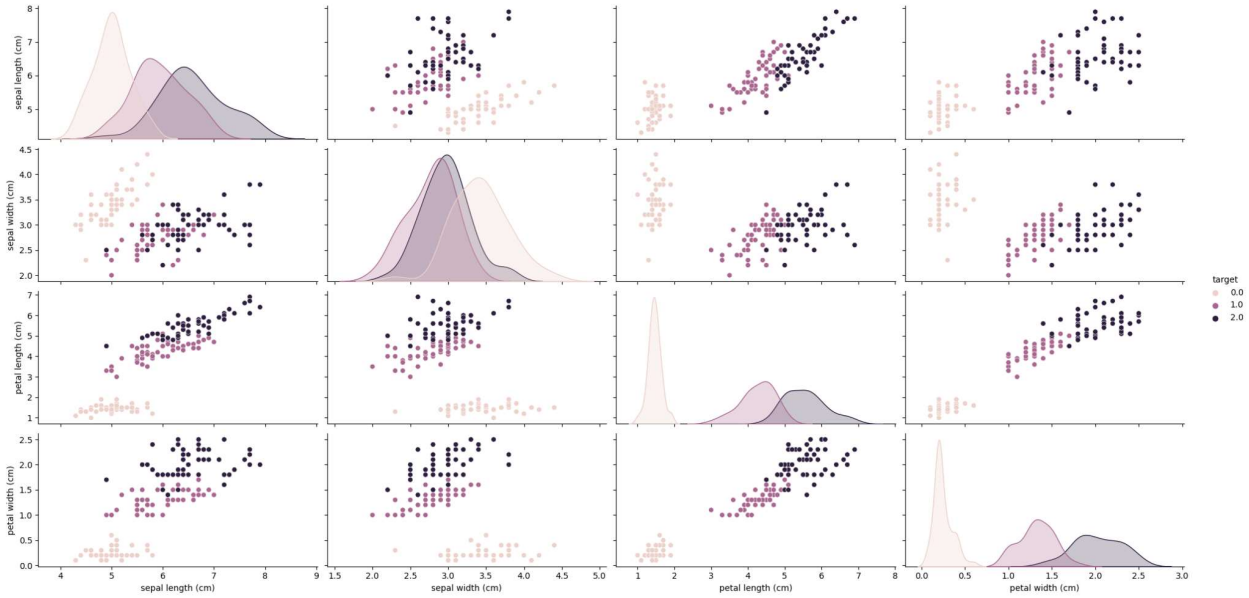
This project was done based on Iris flower data which have three species 'setosa', 'versicolor', 'virginica' and these species need to be identified based on provided data. So the problem that needs to be solved is a classification problem.

Objective:

The data provided have information about species, sepal length, sepal width, petal length & petal width of iris flower. Objective of this project is to develop a prediction model based on provided data to identify species of Iris flower by giving sepal length, sepal width, petal length & petal width as input data.

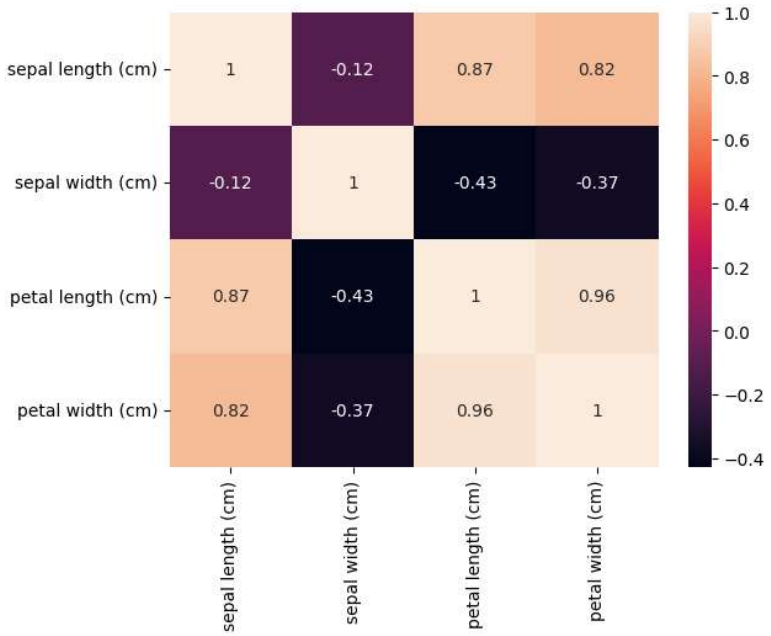
Findings from exploratory data analysis:

1. Feature names provided in the data set are 'sepal length (cm)', 'sepal width (cm)', 'petal length (cm)', 'petal width (cm)'
2. Target names provided in the data set are 'setosa', 'versicolor', 'virginica'
3. Total rows in provided data is 150 and each target have 50 data
4. Pairplot of data is given below:



From the above figure we can see that 'setosa' species can be clearly distinguished by comparing petal length, petal width and pair of any features. But sepal length or sepal width alone cannot distinguish that species. For other species petal length and petal width data is very important to predict the species.

5. Correlation data is shown below:



Above correlation chart shows that sepal length has weak negative correlation, sepal width with petal length and petal width have moderate negative correlation, petal length with petal width and sepal length have strong positive correlation.

data preprocessing:

1. There are no null values in the data set
2. 70% data are used as training data and 30% are used as test data
3. Features data are scaled and target data are encoded.

Model selection:

Different models are tried to select a suitable model before applying the model to the whole data set. Below are the list of models and their accuracy level:

1. Logistic Regression model predicting with 97.78% accuracy
2. K nearest neighbor classifier model predicting with 97.78% accuracy
3. Decision tree classifier model predicting with 97.78% accuracy
4. Random forest classifier model predicting with 97.78% accuracy
5. Support vector machine model predicting with 97.78% accuracy
6. Naive Bayes model predicting with 100% accuracy

So, Naive Bayes model was selected as a prediction model and the model was implemented on the whole data set.

Prediction result:
Some results with new observations are listed below:

Sl	Input value	Predicted species
1	[6.1, 3.4, 4.1, 1.8]	versicolor
2	[4.5, 5, 5, 2.7]	virginica
3	[5.0, 3.3, 1.0, 0.2]	setosa

Conclusion:
The model chosen for predicting species of iris flower provided a very decent accuracy level. As the problem selected is of classification type, RMSE, MSE, MAE or R_squared error calculation isn't applicable. Accuracy, precision, recall, F1 score and confusion matrix are the measures to calculate performance of prediction models for classification problems. Accuracy is applied as the performance measure for the problem and found the naive bayes model accuracy value is 1.0 that means prediction the selected model was doing is 100% accurate.