Report - 1

Objective: Compute the accuracy of the model with all three datasets IRIS, TITANIC and mobile_price.

Solution:

Note:

- Step-1: Open the given Colab file by clicking the link
- Step-2: Upload the mobile_price.csv dataset and remaining dataset access by the seaborn module
- Step-3: Run it
- Step-4: Get the result

Machine Learning:

- Machine learning is a field of artificial intelligence that focuses on creating algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed.
- It involves training a computer system using data, allowing it to automatically learn patterns, relationships, and insights to make accurate predictions or perform specific tasks.

Dataset:

- A dataset is a structured collection of data that represents a particular set of observations, measurements, or information.
- It can consist of various types of data, such as numbers, text, images, or other formats, organized into rows (instances) and columns (variables) for analysis or machine learning tasks.

Multiscaling value in dataset:

It typically refers to a situation where the dataset contains features or variables that have different scales or units of measurement. In other words, the values in the dataset are not uniformly distributed or measured in the same units.

Here's a small sample of the dataset:

Area (sq. ft.)	Price ($)	Bedrooms

1500	200000	3
2000	300000	4
1200	180000	2
2500	350000	4
1800	250000	3

In this example, we can see that the "Area" feature has values ranging from 1200 to 2500, the "Price" feature has values ranging from 180,000 to 350,000, and the "Number of Bedrooms" feature has discrete values of 2, 3, and 4.

The multiscaling values in this dataset arise due to the different units and ranges of the variables. "Area" is measured in square feet, "Price" is measured in dollars, and "Number of Bedrooms" is a discrete variable with no specific unit.

To address this issue, we can apply feature scaling techniques to bring all the variables to a similar scale. Let's use min-max scaling to transform the "Area" and "Price" features to a range between 0 and 1:

Area (scaled)	Price (scaled)	Bedrooms
0.4167	0.2857	3
0.6667	0.4286	4

0.1667	0.1429	2
1.0000	0.5714	4
0.5000	0.3571	3

Now, the "Area" and "Price" features are scaled between 0 and 1, allowing for a more standardized representation of the data. The "Number of Bedrooms" feature remains unchanged since it is a discrete variable.

Multi Data Types :

A dataset with multiple data types refers to a collection of data that includes variables of different types, such as numerical, categorical, text, or date/time. Each data type represents a distinct kind of information and may require specific handling and analysis techniques. Understanding and appropriately managing these data types is crucial for accurate analysis and modeling of the data.

Here’s a small sample of the dataset:

Age	Gender	Education Level	Income	Review	Date of Purchase
30	Male	Bachelor's	50000	"The product is excellent. Highly recommended!"	2022-05-10
45	Female	Master's	80000	"Not satisfied with the quality. Disappointed."	2022-06-20

28	Female	High School	25000	"Great value for money. Will buy again!"	2022-03-15
35	Male	Bachelor's	60000	"Average product. Could be better."	2022-07-05
50	Male	Master's	90000	"The best product I've ever used!"	2022-02-01

In this example, we have a dataset with multiple data types:

- "Age" and "Income" are numerical variables representing age and income, respectively.
- "Education Level" is an ordinal variable representing different levels of education.
- "Review" is a text variable containing customer reviews.
- "Date of Purchase" is a date/time variable representing the date of purchase.

When the dataset contains multiscaling and multi data types values:

The steps for data analysis are,

- Missing value
- Balancing
- Normalization
- Splitting
- Classifier or model (using SVM model)

Some Terms Definition:

- SVM Model: Support Vector Machine (SVM) is a machine learning model that aims to find the best decision boundary or hyperplane to separate data points of different classes, maximizing the margin between the classes.

- Training data: The portion of the dataset used to train or fit a machine learning model. It consists of input features and corresponding target labels. The model learns from this data to understand the underlying patterns and relationships in the data.

- Testing data: The portion of the dataset used to evaluate the performance of the trained machine learning model. It is separate from the training data and contains input features, but the corresponding target labels are typically withheld. The model makes predictions on this data, and the predictions are compared to the actual labels to assess the model's accuracy or performance.

We worked on three datasets:

1. IRIS

    - The Iris dataset is a famous dataset in machine learning and statistics.
    - It consists of measurements of sepal length, sepal width, petal length, and petal width of three different species of iris flowers.
    - The goal of using the Iris dataset is often to classify the iris flowers into their respective species based on these measurements.
    - It is commonly used for tasks such as classification, clustering, and data visualization.

    Case 1: training data is 50% and testing data is 50%
    The accuracy of the model is 98%.

    Case 2: training data is 60% and testing data is 40%
    The accuracy of the model is 100%.

    Case 3: training data is 70% and testing data is 30%
    The accuracy of the model is 95%.

    Case 4: training data is 80% and testing data is 20%
    The accuracy of the model is 93%.

2. TITANIC

    - The Titanic dataset is a dataset that contains information about the passengers onboard the Titanic ship.
    - It includes attributes like age, gender, passenger class, fare, and survival status (whether the passenger survived or not).
    - The dataset is often used for predictive modeling, aiming to predict whether a given passenger survived the Titanic disaster based on the available attributes.
    - It is a popular dataset for learning data analysis and machine learning techniques, particularly for classification or survival prediction tasks.

    Case 1: training data is 50% and testing data is 50%
    The accuracy of the model is 75%.

    Case 2: training data is 60% and testing data is 40%
    The accuracy of the model is 75.36%.

    Case 3: training data is 70% and testing data is 30%

The accuracy of the model is 75.98%.

Case 4: training data is 80% and testing data is 20%
The accuracy of the model is 78.67%.

3. PricePrediction :

- The PP dataset contains features of the cell phone.
- It contains features such as Battery, Screen size, Ram, Storage and so on.
- The goal of this dataset is to train the model and model predict the price of cellphone by taking features as input.

Case 1: training data is 50% and testing data is 50%
The accuracy of the model is 65%.

Case 2: training data is 60% and testing data is 40%
The accuracy of the model is 64.6%.

Case 3: training data is 70% and testing data is 30%
The accuracy of the model is 57%.

Case 4: training data is 80% and testing data is 20%
The accuracy of the model is 54%.

Conclusion:

- Nature of features present in the database.
- Classifier learning on those features of the dataset.
- Considering different cases, the model accuracy varies.