Data Mining and Analysis

Team Members

- Sneh Patel
- Malav Shah
- Dhruvrajsinh Chavda
- Nandiniben Patel

Abstract-*The iris.csv dataset is a well-known dataset that is often used in data science and machine learning projects for classification tasks. It contains data on various attributes of different species of iris flowers, including sepal length, sepal width, petal length, and petal width.*

❖ INTRODUCTION

- The dataset consists of 150 rows and 5 columns, where each row represents a single flower sample, and each column represents a different attribute of that sample.
- The first four columns are numeric variables representing the measurements of the flowers' sepal length, sepal width, petal length, and petal width, respectively.
- The fifth column is a categorical variable representing the species of the flower, with three possible values: setosa, versicolor, and virginica.
- The dataset is often used to explore various data analysis techniques, including data visualization, exploratory data analysis, and classification modeling.
- Its popularity is due in part to the fact that it is a well-understood dataset with a clear structure, making it easy to work with and analyze.

Overall, the iris.csv dataset is a valuable resource for anyone looking to explore and develop their skills in data science and machine learning.

❖ Perform Statistical Analysis

- The iris dataset is a well-known dataset that is often used for statistical analysis and machine learning tasks. In this report, we will provide an overview of the statistical analysis of the iris dataset.
- We can see that the sepal length ranges from 4.3 to 7.9 centimeters, with a mean of 5.84 and a standard deviation of 0.83.
- Similarly, the sepal width ranges from 2.0 to 4.4 centimeters, with a mean of 3.05 and a standard deviation of 0.43.
- The petal length ranges from 1.0 to 6.9 centimeters, with a mean of 3.76 and a standard deviation of 1.76.
- Finally, the petal width ranges from 0.1 to 2.5 centimeters, with a mean of 1.20 and a standard deviation of 0.76.

- We can generate some visualizations of the dataset to gain a better understanding of the relationships between the different attributes. One useful visualization for exploring these relationships is a pair plot, which shows scatter plots and histograms for each pairwise combination of the numeric attributes in the dataset.
- In the context of the iris dataset, sns.pairplot(iris) creates a grid of scatterplots and histograms for the four numerical variables (sepal length, sepal width, petal length, and petal width) and the categorical variable (species). The resulting plot shows how the variables are related to each other and how they differ between the three species of iris flowers.

❖ Box Plot

- A box plot is a graphical representation of the distribution of data based on the five-number summary, which includes the minimum value, first quartile (Q1), median (Q2), third quartile (Q3), and maximum value. It can help to identify outliers and give an overview of the spread and skewness of the data.

- In the context of the iris dataset, a box plot can help us visualize the distribution of the four numerical variables (sepal length, sepal width, petal length, and petal width) and detect any potential outliers. It can also provide insights into the differences between the three species of iris flowers.

❖ Regression Line

- To demonstrate the connection between sepal width and sepal length, petal width and sepal length, and petal length and sepal length, respectively, the algorithm generates three scatterplots with linear regression lines and 95% confidence intervals using the iris dataset. The aspect number determines the plot's aspect ratio, and scatter_kws regulates the scatterplot's point transparency.

- We can determine how closely the factors are correlated and whether the connection is positive or negative by looking at the regression lines and the scatter of the data points around them.
- As the regression line slopes lower in the scatterplot of sepal width versus sepal length, for instance, we can see that there is a faint negative association between the two variables.
- As the regression line slopes upward in the scatterplot of petal length versus sepal length, however, we can see that there is a significant positive association between the two factors.

❖ Clustering

- The sklearn.neighbors package is used to load the KNeighborsClassifier class. The k-nearest neighbours algorithm, a kind of supervised learning method used for categorization tasks, is implemented by this class. In this method, a new data point's class is forecast based on the class of its k nearby training data neighbours.

- In many machine learning and statistical models, choosing X and Y and standardizing the features are crucial preparatory stages. We can create models to forecast the target variable based on the characteristics by dividing the data into features and target. By guaranteeing that each characteristic is on the same scale and is given the same significance, standardization can enhance the efficacy of these models.

- Sklearn is a Python module that offers a variety of tools and methods for applying machine learning techniques. The metrics and classification report functions are helpful for assessing the performance of any machine learning model, and the KNeighborsClassifier class is just one illustration of the many algorithms accessible in the module.

❖ Fitting The Model

- The iris dataset is being used to build and train a K-nearest neighbour (K-NN) model.
- The data is first preprocessed by changing it with fit_transform and setting a standard scaler on the independent variable X_pca.()
- The training data is then fitted to a K-NN classification object using fit. (). Using predict, the model is used to make projections about the test results.().

❖ Classification Report

- Finally, we use classification report () to produce a classification report that displays the target variable y's target variable's precision, recall, f1-score, and support for each class.
- The study demonstrates that for all three classes—Iris-setosa, Iris-versicolor, and Iris-virginica—the model was able to attain perfect accuracy, recall, and f1-score.
- The success of the model is further supported by the fact that the weighted average of all the measures is flawless at 1.0

❖ Conclusion

- Correlation shows whether a linear relationship between variables exists and shows if one variable tends to occur with large or small values of another variable.
- The scatter plot shows quite a strong positive relationship overall between the petal length and petal width measurements.
- The relationship between petal length and petal width is not as strong for the Iris Setosa as with the other two species.
- The correlation matrix showed a strong relationship between sepal length and sepal width for the Iris Setosa only and a strong relationship between the petal length and petal widths for the Iris Versicolor only.
- However, the relationship between the petal lengths and sepal lengths is very strong for the Iris Virginica with Iris Versicolor being only a little bit weaker.