

Department of Computer Science and Engineering

Bangabandhu Sheikh Mujibur Rahman Science and Technology University

EDGE-BSMRSTUCSE Digital Skills Training Final Exam



Final exam: 25 marks

Batch 27

Course title: Data analysis with python

Dataset link:

<https://archive.ics.uci.edu/dataset/53/iris>

Tasks:

1. Data Preprocessing:

- Load the dataset and convert it into a DataFrame with appropriate column names.
- Check for any missing values and handle them appropriately.
- Standardize the numerical features (sepal_length, sepal_width, petal_length, petal_width) using StandardScaler from sklearn.

2. Exploratory Data Analysis (EDA):

- Compute summary statistics (mean, median, variance) for each feature, grouped by species.
- Plot a box plot for each feature, separated by species.
- Create a violin plot to show the distribution of petal_length for each species.

3. Data Visualization:

- Create a scatter plot of sepal_length vs sepal_width, colored by species.
- Create a pair plot (using seaborn) to visualize the relationships between all numerical features, colored by species.

4. Class Distribution:

- Count the number of samples for each species and display the distribution in a bar plot.
- Calculate the percentage distribution of each species in the dataset.

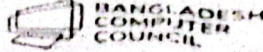
5. Correlation Analysis:

- Compute the correlation matrix for the numerical features and display it as a heatmap.
- Interpret which features have the strongest correlation and whether it varies significantly between species.

Department of Computer Science and Engineering

Bangabandhu Sheikh Mujibur Rahman Science and Technology University

EDGE-BSMRSTUCSE Digital Skills Training Final Exam



Final exam: 25 marks

Batch 27

Course title: Data analysis with python

6. Feature Engineering:

- Create new features `sepal_ratio` (`sepal_length/sepal_width`) and `petal_ratio` (`petal_length/petal_width`).
- Analyze which species has the largest average `sepal_ratio` and `petal_ratio`.
- Visualize the distribution of these new features using histograms, separated by species.

7. Dimensionality Reduction:

- Apply Principal Component Analysis (PCA) on the standardized data and reduce it to 2 dimensions.
- Plot a scatter plot of the two principal components, color-coded by species.
- Interpret how well PCA separates the species.

8. Clustering Analysis:

- Use the K-Means algorithm to cluster the data into 3 clusters (assuming you don't know the species labels).
- Compare the K-Means clusters with the actual species labels using a confusion matrix and calculate the clustering accuracy.
- Plot the clusters and centroids on a scatter plot of the first two principal components.

9. Classification Model:

- Train a Random Forest Classifier to predict the species of the flowers.
- Split the data into 80% training and 20% testing sets.
- Evaluate the model by calculating accuracy, precision, recall, and the F1-score.
- Plot the feature importance based on the Random Forest model.