CS 5710 Project

Project Report on Machine Learning

Name	Student ID
Venkata Sri Vamsi Ponnada	700762561
Sai Akshith Akshintala	700757137
Mothukur Venkata Deepthi	700761693

Dataset description:

The Palmer Penguins dataset is a popular dataset in the field of data science and machine learning. It contains data on various measurements of penguins collected from the Palmer Archipelago in Antarctica. This dataset was created by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER.

The dataset consists of several features, including measurements such as culmen_length_mm, culmen_depth_mm, flipper length, and body mass. It also includes categorical variables such as species (e.g., Adelie, Chinstrap, Gentoo), island (Torgersen, Biscoe, Dream), and sex (male, female).

The primary purpose of the Palmer Penguins dataset is to study the characteristics and behavior of penguin species inhabiting the Palmer Archipelago. Researchers and data scientists often use this dataset for tasks such as exploratory data analysis, classification, clustering, and predictive modeling.

Due to its well-structured nature and real-world relevance, the Palmer Penguins dataset serves as an excellent resource for learning and practicing various data science techniques and algorithms.

We have performed a sequence of steps on the dataset before proceeding to model training. They are:

1) Data Preprocessing

- **Data Cleaning:**
    - Handled missing values: Checked for missing values in the dataset and imputed them using appropriate methods.
    - Removed duplicates: Identified and removed duplicate records from the dataset.

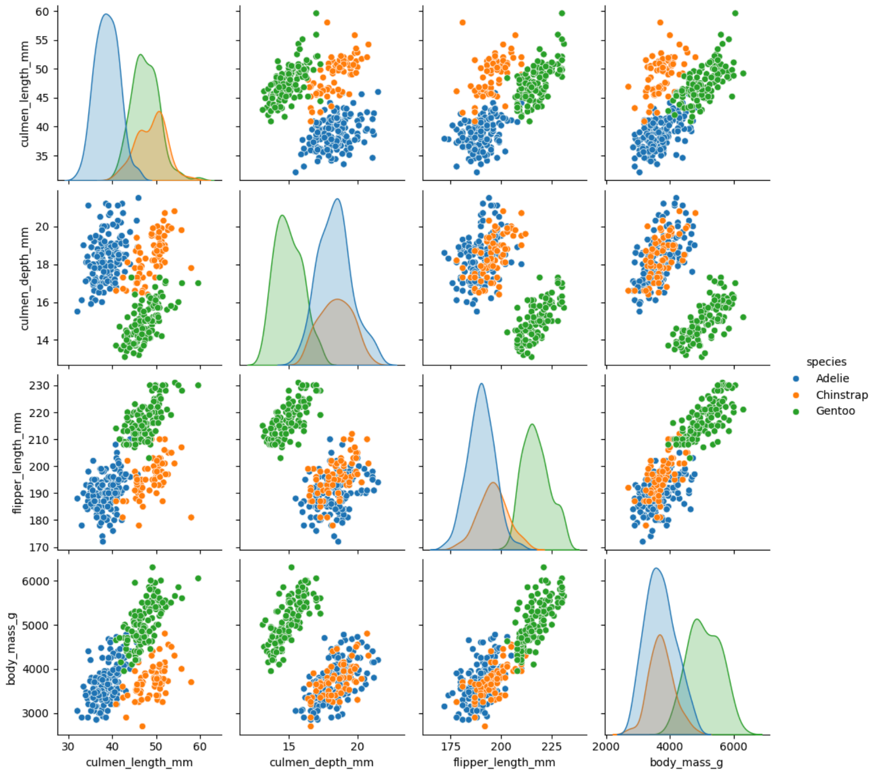
- **Data Transformation:**
    - Encoded categorical variables: Converted categorical variables into numerical format using techniques like one-hot encoding.
    - Scaled numerical features: Scaled numerical features to ensure they are on a similar scale.
- **Feature Engineering:**
    - Created new features: Extracted additional features from existing ones to enhance the dataset.

2) **Exploratory Data Analysis (EDA)**

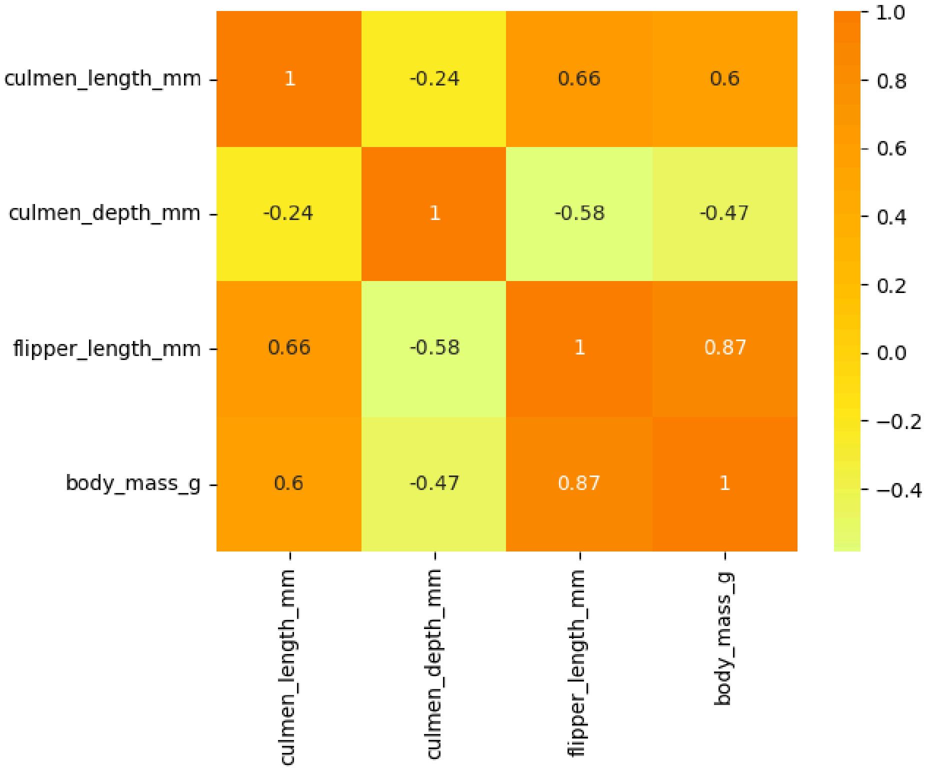
- **Descriptive Statistics:**
    - Calculated summary statistics such as mean, median, minimum, maximum for numerical features.
    - Examined frequency distribution for categorical features.
- **Univariate Analysis:**
    - Visualized histograms for numerical features to understand their distributions.
    - Plotted bar plots for categorical features to observe their frequencies.
- **Bivariate Analysis:**
    - Investigated relationships between numerical features using scatter plots.
    - Analyzed the relationship between categorical and numerical features

3) Data Visualization

- **Distribution Visualization:**
    - Displayed distributions of various features using histograms and KDE plots.
- **Relationship Visualization:**
    - Utilized pair plots to visualize pairwise relationships between features.

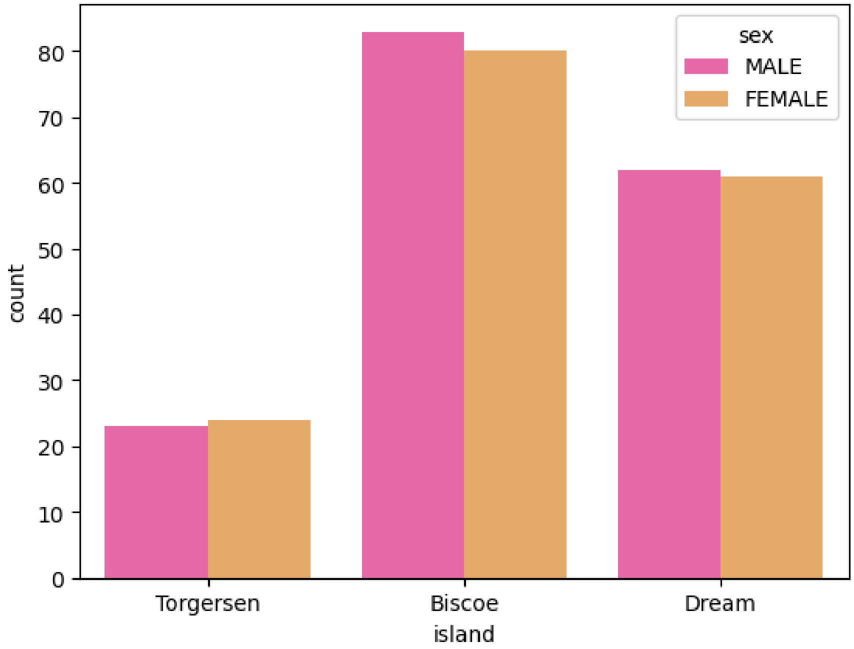


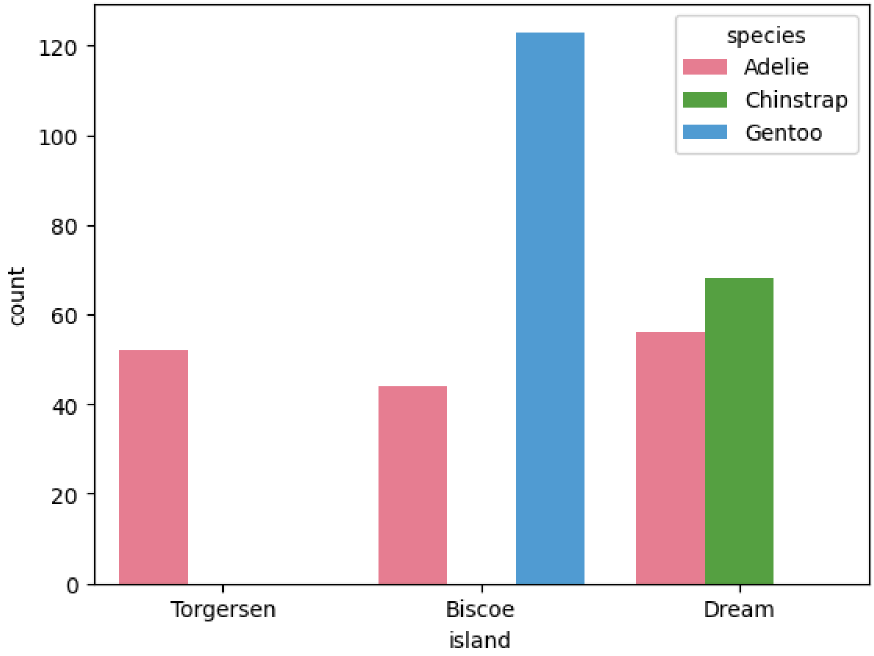
- ○ Created heatmaps to visualize correlations between features.

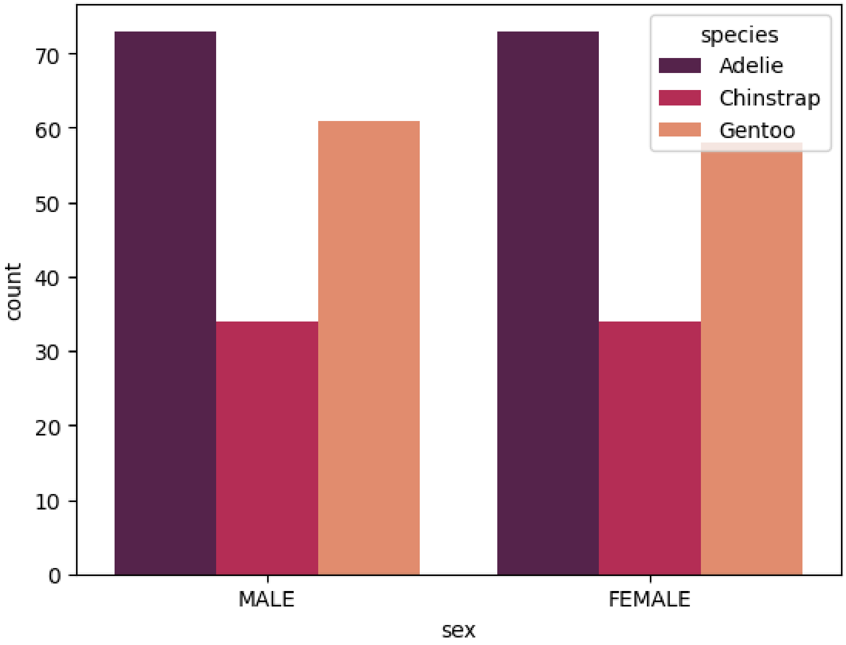


	culmen_length_mm	culmen_depth_mm	flipper_length_mm	body_mass_g
culmen_length_mm	1	-0.24	0.66	0.6
culmen_depth_mm	-0.24	1	-0.58	-0.47
flipper_length_mm	0.66	-0.58	1	0.87
body_mass_g	0.6	-0.47	0.87	1

- **Categorical Visualization:**
  - Illustrated distributions of categorical variables using bar plots and pie charts.







We have evaluated the performance of four classifiers trained on the Palmer Penguins dataset. The dataset has been split into a 70% training set and a 30% testing set. The goal is to classify penguin species based on various features such as culmen_length_mm, culmen_depth_mm, flipper length, and body mass.

Classifiers

We have selected the following classifiers for evaluation:

- Support Vector Classifier (SVC)
- Random Forest Classifier
- Decision Tree Classifier
- K-Nearest Neighbors Classifier

Training and Evaluation

Data Split:

- Training Set: 70%
- Testing Set: 30%

Metrics:

- We have evaluated the classifiers based on the following metrics:
    - Accuracy: Overall classification accuracy.

- Precision: Ability of the classifier not to label as positive a sample that is negative.
- F1-Score: Harmonic mean of precision and recall.

Results:

1. Support Vector Classifier (SVC):
   - Accuracy: Accuracy score of svc is 0.78.
   - Precision: Precision of SVC is 0.8158672248803828
   - F1-Score: f1_score of SVC is 0.7686794717887157
2. Random Forest Classifier:
   - Accuracy: Accuracy score of Random Forest Classifier is 0.99
   - Precision: Precision of Random Forest Classifier is 0.9904166666666667
   - F1-Score: f1_score of Random Forest Classifier is 0.9900446545836615
3. Decision Tree Classifier:
   - Accuracy: Accuracy score of Decision Tree Classifier is 0.78
   - Precision: Precision of Decision Tree Classifier is 0.8158672248803828
   - F1-Score: f1_score of Decision Tree Classifier is 0.7686794717887157
4. K-Nearest Neighbors Classifier:
   - Accuracy: Accuracy score of K-Nearest Neighbors Classifier is 0.78
   - Precision: Precision of K-Nearest Neighbors Classifier is 0.8158672248803828
   - F1-Score: f1_score of K-Nearest Neighbors Classifier is 0.7686794717887157

Based on the results obtained from training and testing the classifiers on the Palmer Penguins dataset, we can draw the following conclusions:

1. **Accuracy:**
   - The Random Forest Classifier achieved the highest accuracy score of 0.99, indicating that it made the most accurate predictions on the testing set.
   - Support Vector Classifier (SVC), Decision Tree Classifier, and K-Nearest Neighbors Classifier all achieved an accuracy score of 0.78, indicating that they performed equally well in terms of overall classification accuracy.
2. **Precision:**
   - The Random Forest Classifier also achieved the highest precision score of 0.9904, indicating that it had the highest proportion of true positive predictions among all positive predictions made.
   - Support Vector Classifier (SVC), Decision Tree Classifier, and K-Nearest Neighbors Classifier all achieved a precision score of approximately 0.8159, suggesting that they had similar performance in terms of precision.

3. **F1-Score:**
    - The F1-score, which combines precision and recall into a single metric, was consistent across all classifiers, with a value of approximately 0.7687.
    - This indicates that all classifiers achieved similar balance between precision and recall, resulting in comparable overall performance.

Conclusion:

- The Random Forest Classifier outperformed the other classifiers in terms of accuracy and precision, making it the most suitable choice for this classification task.
- However, it's essential to consider other factors such as computational complexity, interpretability, and scalability when selecting the final model.
- Further analysis, including hyperparameter tuning and feature engineering, could potentially improve the performance of all classifiers.