BIRZEIT UNIVERSITY

Department of Electrical & Computer Engineering
ENCS5141 - INTELLIGENT SYSTEMS LAB

Assignment#1 - Case Study

Prepared by: Ahmad Abbas
Instructor: Aziz Qaroush
Assistant: Eng. Mazen Amria
Date: November 20, 2023

Abstract

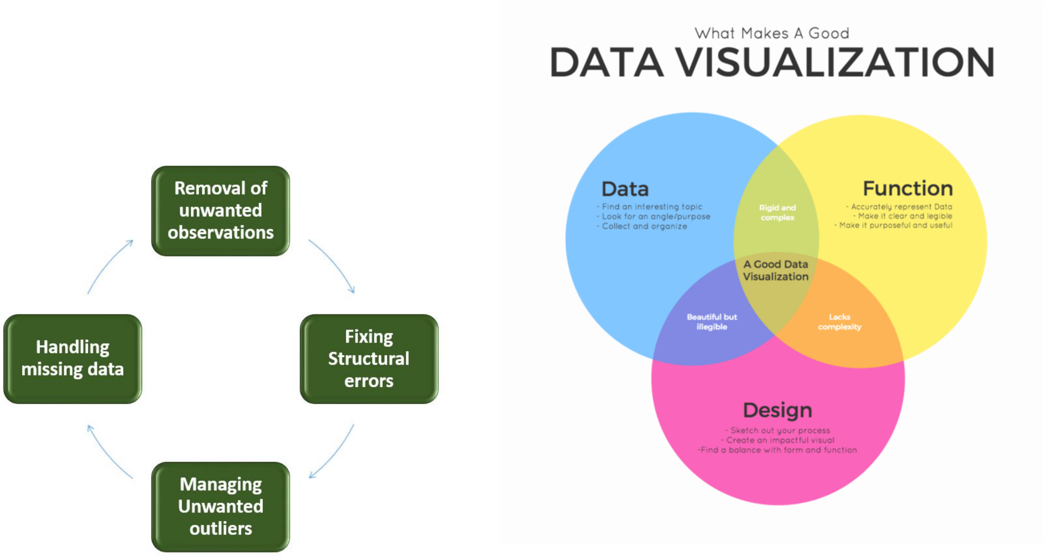
In this case study, we apply data science concepts, particularly focusing on data preprocessing techniques such as data cleaning, feature engineering, normalization, and dimensionality reduction. We employ these methods on the Seaborn "penguins" dataset, which includes both numerical and categorical features of penguins, such as bill length and species, to study the enhancement of machine learning models' performance. Techniques like Variance Threshold, SelectKBest, and PCA are used in our case study. Our approach involves using Random Forest model to evaluate the performance of preprocessing, especially studying how specific techniques effect model accuracy. The study shows the improvement on performance from processing the data, which we figured a difference on it.

Contents

1 Introduction

The fields of data science and machine learning are growing rapidly, and at the same time, businesses have an increasing need for professionals who can effectively analyze large amounts of data to help on efficient decision-making. Data science involves dealing with large datasets, including tasks such as data cleansing, preprocessing, and analysis. Data scientists integrate data from various sources and use machine learning, predictive analytics, and sentiment analysis to extract valuable insights from datasets [1].

Data Visualization and Data Cleaning are essential pillars in data analysis with significant roles in the field of Data Science and Machine Learning. Data Visualization involves representing data graphically to extract patterns, insights, and valuable information that may not be clear in the raw data [2]. On the other hand, data cleaning is the process of ensuring that the data doesn't contain incorrect, noisy, duplicate, or incomplete raw's within your dataset. Dealing with multiple sources increases the probability of having issues on the data. All these issues can lead to unreliable and inaccurate models. The process of cleaning data depends on the dataset and the objectives, but it is important to have a plan in place to ensure that you perform it correctly [3].


Removal of unwanted observations
Fixing Structural errors
Managing Unwanted outliers
Handling missing data
What Makes A Good
DATA VISUALIZATION
Data
- Find an interesting topic
- Look for an angle/purpose
- Collect and organize
Rigid and complex
Function
- Accurately represent Data
- Make it clear and legible
- Make it purposeful and useful
A Good Data Visualization
Beautiful but illegible
Lacks complexity
Design
- Sketch out your process
- Create an impactful visual
-Find a balance with form and function

(a) Data cleaning cycle[4]

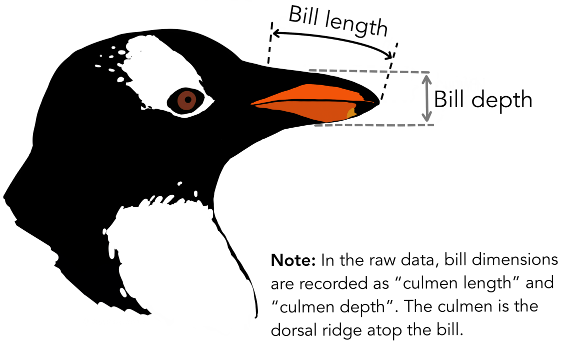
(b) good data visualization mix of good data, design, and functionality [5].

Figure 1.1: Good Data Visualization and Data Cleaning Cycle

When discussing Data Science, we must talk about Feature Engineering. Feature engineering is an important technique that use data to create a new feature set not present in the training set, with the goal of simplifying data transformations and enhancing model accuracy. In machine learning, feature engineering typically involves four key steps: Feature Creation, Transformations, Feature Extraction, and Feature Selection. However, in this experiment, our focus is on feature selection and reduction [6, 7].

1.1 Penguins Dataset

The popular Seaborn dataset "penguins" is commonly used for tasks involving exploratory data analysis and data visualization. This dataset includes physical attributes such as body mass, sex, bill length, and flipper length for a variety of penguin species. It also mentions the islands where the penguins were seen to be. This dataset is frequently used by researchers to investigate the relationships between different penguin species, their physical characteristics, and their environments [8, 9]. We will be using the Pandas DataFrame, a useful tool for data manipulation and analysis, to work with the "penguins" dataset. This will enable us to effectively organize, filter, and analyze the data in order obtain further knowledge and support our study [10].



Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

Figure 1.2: How bill length and depth are measured [11].

1.2 Main Objectives

Our main objectives include initial data exploration, data cleaning, feature engineering, encoding of categorical variables, data normalization, dimensionality reduction, and validation of our preprocessing pipeline by measuring the accuracy before and after it for different preprocessing pipelines.

Our first first step is doing an exploratory data analysis for the dataset in order to understand the structure of this dataset, detecting any missing values, and get patterns from visualization. We also summarize the dataset's statistics to gain deeper insights. This process helps in understanding the data and plays a crucial part in the future steps.

After that, we focus on cleaning the data by handling missing values and outliers. For missing values, we will use two approaches: mean replacements for missing numerical data depending on categorical , where missing values are replaced with the mean of the respective feature, and linear regression techniques based on correlations among features. For categorical data, we will use KNN (K-nearest neighbor) imputation to replace the missing values. For outliers, a z-score analysis will be used to detect outliers and deleting them. Moreover, We will use chopping method, considering data points between the 5^{th} and 90^{th} percentiles, to reduce the effect of outliers on our analysis.

The third objective of our case is to feature selection in our pipeline by selecting the most relevant features from the dataset to improve the performance and reduce the computation cost. We will use many techniques, which we have used in the experiment, like Variance Threshold for removing features with low variance, SelectKBest with mutual information, which measure of the mutual dependence between the two variables [12], and F-class scores, which uses the ANOVA f-test for feature selection and considers only linear dependency, as opposed to mutual information-based feature selection, which may account for any type of statistical dependency [13], for selecting the best features based on their statistical significance. Furthermore, Principal Component Analysis (PCA) is a dimensionality reduction strategy that is frequently used to reduce the dimensionality of big data sets by reducing a large collection of variables into a smaller one that retains the majority of the information from the large set [14].

A significant part of the "penguins" dataset preparation is the encoding of categorical variables into numerical representations, which is required by most ML methods. This encoding is using methods like label encoding, which are selected based on the unique properties of the categorical data.

The final stages of preprocessing involve splitting the dataset into training and testing subsets, which is important for evaluating the performance of machine learning models and avoid overfitting. After that, we train a model, Random Forest [15], to validate the performance of our preprocessing steps. By comparing the model's performance on the preprocessed data vs the raw data, we will see the effect of our preprocessing, including feature filtering, transformation, and reduction.

2 Procedure and Discussion

2.1 Data Exploration

The ”penguins” dataset have 344 data rows, combining both categorical and numerical features across seven attributes. The categorical data, which are species, island, and sex, have few limited values. In the other hand, the numerical data, representing physical measurements like bill length and body mass, we can see different ranges across the numerical data.

With more exploring, the dataset contains missing values, with two rows missing all numerical values, and eleven rows with missing ’sex’ category. In the other hand, ’species’ and ’island’ are not missing in any row. This requires data processing for better analysis.

2.2 Data Visualization

We will represent a series of visualizations from the dataset, each of them shows a specific point of the data and offering insights into the penguin population under study.

The below pie chart shown in Figure 2.1 provides a clear visual representation of penguin species within the dataset. It shows that Adelie penguins are the most frequent, accounting for 44.19% of the entries, followed by Gentoo penguins at 36.05%, and Chinstrap penguins with 19.77%. This visualization shows the proportional differences between the species, highlighting the dominance of the Adelie species in the dataset.

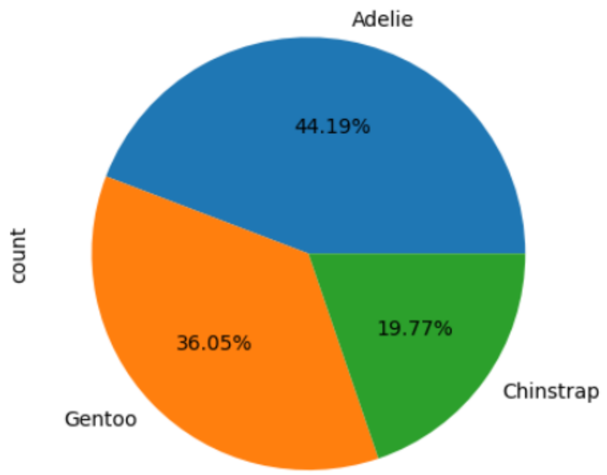


Figure 2.1: Distribution of Penguin Species in the Dataset.

The pie chart shown below, Figure 2.2, represents the distribution of of penguins across the three islands represented in the dataset. Biscoe Island have the largest number of penguins, with nearly half of the observations (48.84%). Dream Island follows with 36.05%, and Torgersen Island has the smallest number at 15.12%. This distribution can be a reflection of the ecological factors affecting penguin populations in these areas, which is mentioned on [16].

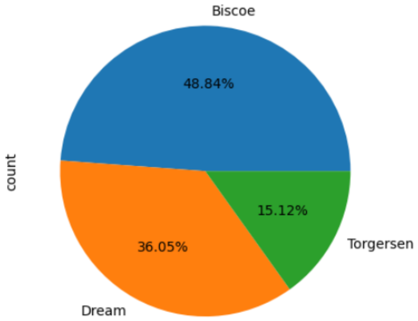


Figure 2.2: Distribution of Penguin Islands in the Dataset.

The pairwise plot in Figure 2.3 provides a complete overview of the relations between the all numerical values on the dataset, grouped by by species. Each plot combines scatter plots for attributes relation and density plots for distribution of single attributes. Clear clusters are seen among different species, depending on that, these values can be strongly used for species classification task. As shown, Gentoo penguins tend to have longer flipper lengths and higher body masses, while Adelie and Chinstrap species show more overlap between their values but can still be distinguished based on bill dimensions.

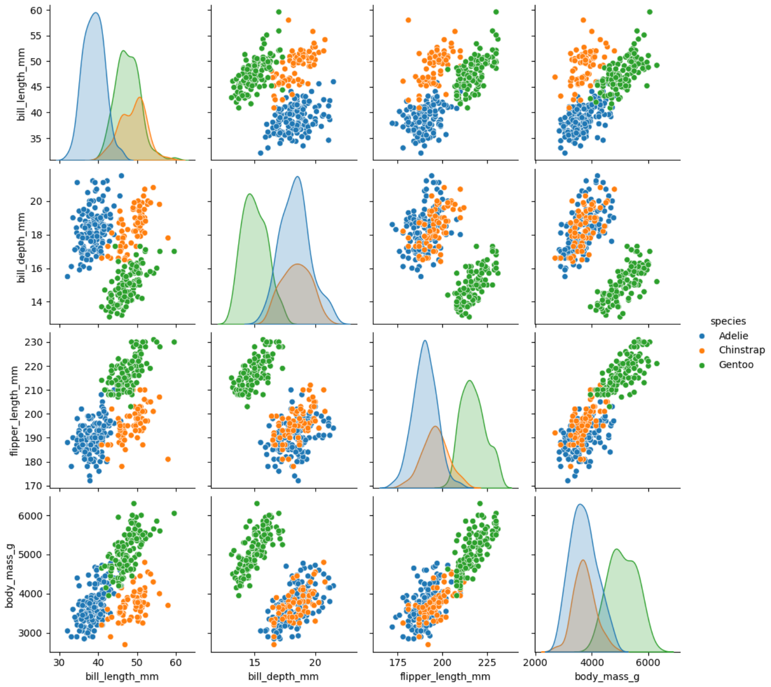


Figure 2.3: Pairwise Relationships and Distributions of Numerical values Among Penguin Species.

The jointplot we have made shows how the lengths of flippers and bills are linked across different penguin species. It seems like the bigger the flipper, the bigger the bill. Gentoo penguins, shown in green, usually have bigger flippers and bills, while Adelie and Chinstrap types mix together a bit but still show their own unique sizes.

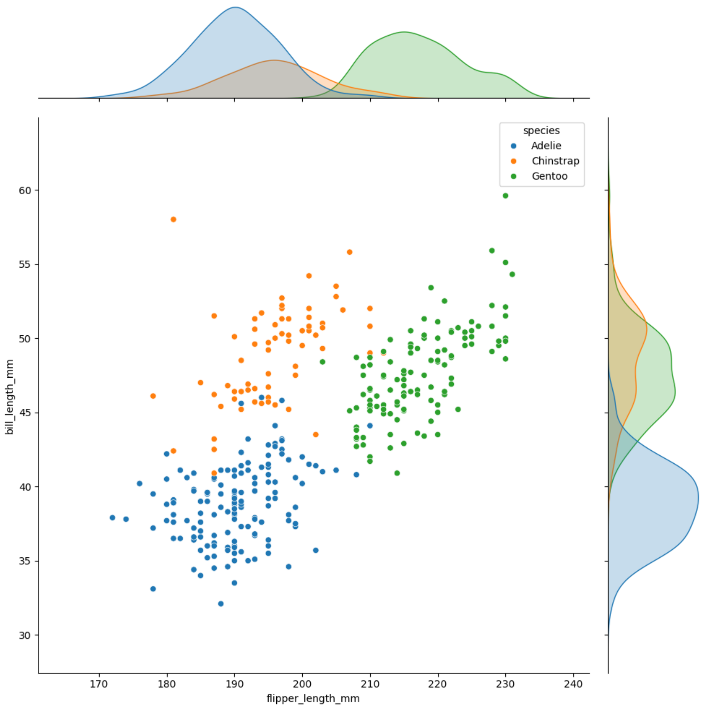


Figure 2.4: Scatter plot and Density Plots of Flipper Length vs. Bill Length by Penguin Species.

2.3 Statistical Analysis

Pandas dataframe has a powerful method ‘df.describe()‘ which provides statistical summary of the dataset’s numerical features. Which shown in the table 1 below. As shown, mean values are with their standard deviations indicates the small variability within each measurement. On the other hand, the standard deviation of body mass reflects a big spread in the weight of the penguins.

Table 1: The result of pandas describe function on penguins dataset

	bill_length_mm	**bill_depth_mm**	**flipper_length_mm**	**body_mass_g**
count	342.000000	342.000000	342.000000	342.000000
mean	43.921930	17.151170	200.915205	4201.754386
std	5.459584	1.974793	14.061714	801.954536
min	32.100000	13.100000	172.000000	2700.000000
25%	39.225000	15.600000	190.000000	3550.000000
50%	44.450000	17.300000	197.000000	4050.000000
75%	48.500000	18.700000	213.000000	4750.000000
max	59.600000	21.500000	231.000000	6300.00000

The histograms with kernel density estimations (kde) below, shows a visualization of the frequency distribution for each of the numerical values in the dataset: bill length, bill depth, flipper length, and body mass. The distribution suggest multiple degrees of skewnes . For instance, bill length and bill depth shown to have a more symmetric distribution, whereas flipper length and body mass show a slight skew.

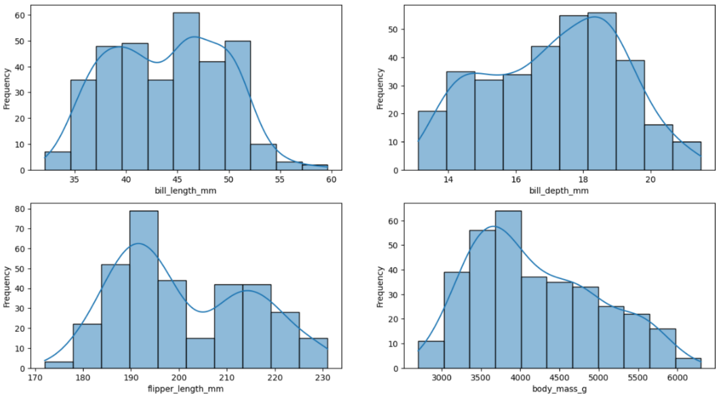


Figure 2.5: Frequency Distribution of Penguins’ numerical values with KDE.

With deeper analysis, we have got the statistics grouped by species. We observe unique patterns for each species. Adelie penguins, comparing means, have shorter bill lengths and also have the smallest body mass and flipper length. Chinstrap penguins have longer bills than Adelie and Gentoo, and their body mass is a much smaller than Gentoo’s.

Gentoo penguins have the longest flippers and heaviest body mass, with a clear shorter bill depth compared to the other species.

Grouping the data by islands shows another view. Penguins from Biscoe are, on average, larger in all measured values, except bill depth, than those from Dream and Torgersen. Penguins from Dream have comparable bill lengths to those from Biscoe but are lighter and have shorter flippers. Torgersen's penguins have the shortest bill lengths and flipper lengths, so a smaller size overall.

The tables for the both species and islands analysis will be provided on the notebook of the case study. After these investigations, we can start out work on processing the data and evaluate the performance before and after processing. For evaluation, we will consider islands as our label on the classification.

Moreover, we have found the correlation between all features which is shown in the Table 2 below. This helps us in understanding the relations between different measurements alongside the encoded categorical features, which encoded using label encoder library. As we can see, flipper length and body mass have a high positive correlation (0.871202), so the penguin with longer flippers tend to be heavier. On the other hand, bill depth and flipper length show a negative correlation (-0.583851). Moreover, the correlation between species and other numerical values represents a string relationship between the species classification and their physical structures.

Table 2: Correlation Table Between Features

	bill_length_mm	**bill_depth_mm**	**flipper_length_mm**	**body_mass_g**	**species_encoded**	**island_encoded**	**sex_encoded**
bill_length_mm	1.000000	-0.235053	0.656181	0.595110	0.731369	-0.353647	0.271440
bill_depth_mm	-0.235053	1.000000	-0.583851	-0.471916	-0.744076	0.571035	0.311460
flipper_length_mm	0.656181	-0.583851	1.000000	0.871202	0.854307	-0.565825	0.215992
body_mass_g	0.595110	-0.471916	0.871202	1.000000	0.750491	-0.561515	0.361224
species_encoded	0.731369	-0.744076	0.854307	0.750491	1.000000	-0.635659	0.008559
island_encoded	-0.353647	0.571035	-0.565825	-0.561515	-0.635659	1.000000	0.029246
sex_encoded	0.271440	0.311460	0.215992	0.361224	0.008559	0.029246	1.000000

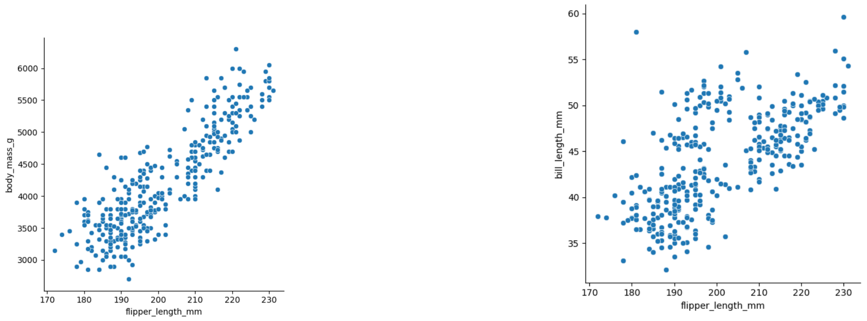
2.4 Data Cleaning

This section will discuss two aspects, handling missing data and outlier with representing our methods for each of them.

2.4.1 Handling Missing Data

After observing the correlation table, we have seen the strong correlation between the flipper length and species and island combined. So we have started handling the missing data for flipper length feature by replacing them with the mean of flipper length for each combination of island and species. Having two rows with missing values, we have the replaced the row of the Adelie penguin in Torgersen island with the value (191.196078) and the other row, which is a Gentoo penguin in Biscoe island, with the value (217.186992).

After that, we have started handling the body mass feature by plotting a scatter plot between flipper length and body mass as shown in the figure 2.6a below. According to the high and clear linear relation between the two features, we have used linear regression model to predict the missing values using flipper length. We have used a training set that exclude any outliers in the flipper length feature based on the lower quartile (25^{th} percentile). We have used the same approach with bill length.



(a) Flipper length VS body mass

(b) Flipper length VS bill length

Figure 2.6: Handling missing data with linear regression

After getting a linear relationship between flipper length and body mass. We have used a scatter plot, which is shown below in figure 2.7, to get the relationship between flipper length and bill depth and observed two different area. By using a K-Means clustering algorithm, we have divided the penguins into two clusters. Then applied a linear regression within each cluster to predict missing values. After getting a threshold from the cluster centers to split the dataset, we have filled in the missing values for each area. This approach allowed us to replace the missing data with more accurate way to avoid noise.

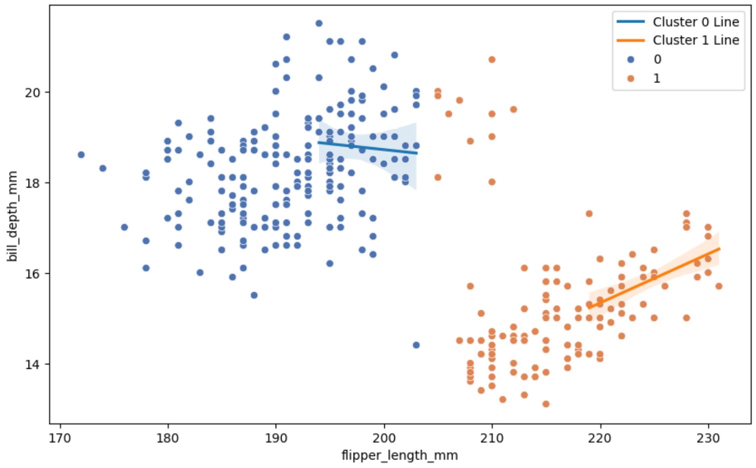


Figure 2.7: Flipper length VS bill depth clusters

For the missing 'sex' values, we have used machine learning KNN model with choosing the closest neighbor. After setting aside the rows with known 'sex' values to train our model and scaling them. This approach is grounded in the assumption that penguins with similar species and island attributes have a higher likelihood of sharing the same sex.

When dealing with the outliers, we have used two different approaches for different sets of features, based on their distribution that shown in Figure 2.5. For bill length and depth, we calculated the Z-scores, setting our bounds at 2.5 standard deviations from the mean. This method ensures that we get data that are within a reasonable range of the expected values.

For flipper length and body mass, we used percentile capping. We set the lower bound at the 10^{th} percentile and the upper bound at the 95^{th} percentile, cutting the values outside this range. This method is particularly effective for features that may have extreme values which could skew the analysis. By capping these features at the specified percentiles, we decreases the effect of outliers, ensuring a more normalized distribution.

2.5 Feature Engineering

We have started by encoding our categorical variables 'species', 'sex', and 'island' using label encoding. Next, we scaled our features to normalize the distribution of data points using StandardScaler. This is important because it ensures that each feature contributes equally to the distance calculations in our models.

For feature selection, we have used multiple methods:

1. Variance Threshold: This technique filters out features with a variance below a certain threshold, under the assumption that features with low variance do not contribute significantly to the model's performance.

2. SelectKBest: We then applied SelectKBest with mutual information and f-classifier as the scoring functions to select the top k features.

Then we incorporated a feature reduction step using Principal Component Analysis (PCA), By applying PCA with n_components=4 as a default value. Moreover, We have used cross validation scores in order to enhance the performance of our models.

3 Results

After all our processing pipeline, we have evaluated our processing using Random Forrest model. To get clear results, we have tested all possible combinations for the label 'island' and all results is shown on the Table 3 below.

Table 3: Best Accuracy Scores Across Different Configurations

Configuration	**Non-Cleaned DF**	**Non-Cleaned DF with CV**	**Cleaned DF**	**Cleaned DF with CV**
Without VT & PCA	0.6418	0.7182	0.7101	0.7251
With VT, Without PCA	0.6567	0.7251	0.6667	0.7251
Without VT, With PCA	0.6418	0.6955	0.7101	0.6775
With VT & PCA	0.6567	0.7251	0.6667	0.7251
SelectKBest (MI), No PCA	0.671642	0.7482	0.652174	0.7397
SelectKBest (MI), PCA	0.671642	0.7482	0.666667	0.7397
SelectKBest (F-Clf), No PCA	0.671642	0.7482	0.652174	0.7251
SelectKBest (F-Clf), PCA	0.656716	0.7297	0.666667	0.7251

3.1 Performance with/without Cleaning

The cleaned dataset generally showed improved or comparable performance to the non-cleaned dataset. As a result, the data cleaning process was effective in increasing model accuracy, emphasising on the importance of preprocessing steps in machine learning workflows.

3.2 Impact of Variance Threshold Selection and PCA

The usage of Variance Threshold (VT) and PCA often represented nearly the same performance without any enhancement on the cleaned data, small enhancement on non-cleaned data. However, the slight enhancement in the non-cleaned data indicates that VT and PCA were more beneficial in managing hidden noise and irrelevant features present in the raw data.

3.3 SelectKBest: Mutual Information (MI) vs F-Classifier

Models using Mutual Information (MI) in the SelectKBest method generally outperformed those using the F-Classifier by the ability of MI to capture non-linear relationships in the dataset more effectively than the F-Classifier [17]. But in our dataset, the values performance was nearly the same. This indicates that the relations in our dataset seems to be more linear.

3.4 Importance of Cross-Validation

Cross-validation consistently enhanced model accuracies across various configurations. This highlights the importance of cross-validation which ensures that the model is tested on multiple subsets of the dataset.

3.5 Best Performing Configurations

The highest accuracies were observed with SelectKBest using Mutual Information, both with and without PCA, in both cleaned and non-cleaned datasets when cross-validation was applied. This result underscores the efficacy of this feature selection technique for the specific characteristics of the dataset in question.

4 Conclusion

In conclusion, the preprocessing pipelines plays an important part on the performance of the machine learning models. Data cleaning enhanced model accuracy in general, highlighting its importance in the machine learning workflow. Variance Threshold (VT) and PCA improved model accuracy slightly for non-cleaned data but had little effect on cleaned data. This implies that VT and PCA are more powerful in datasets containing noise and non related features. Models in the SelectKBest approach that included Mutual Information (MI) and f-cassifier did well in general, confirms the efficiency of them in capturing connections on the data. Their close performance indicates that the rrelations in our dataset seems to be more linear. Cross-validation enhanced model accuracies regularly, highlighting its importance in providing well model testing.

The study concludes that the preprocessing steps significantly impacted the performance of the machine learning models. Data cleaning generally improved or maintained model accuracy, highlighting its importance in the machine learning workflow. Variance Threshold (VT) and PCA showed a slight enhancement in model accuracy for the non-cleaned data but had minimal impact on the cleaned data. This suggests that VT and PCA are more beneficial in datasets with inherent noise and irrelevant features. Models using Mutual Information (MI) in the SelectKBest method generally performed well, indicating the effectiveness of MI in capturing non-linear relationships. Cross-validation consistently improved model accuracies, underlining its significance in ensuring robust model testing. The highest accuracies were observed with SelectKBest using MI, both with and without PCA, across cleaned and non-cleaned datasets when cross-validation was applied. These results underscore the importance of thoughtful preprocessing in machine learning, particularly in feature selection and dimensionality reduction, to enhance model accuracy and reliability.

References

[1] Nov 2023. [Online]. Available: https://www.simplilearn.com/data-science-vs-data-analytics-vs-machine-learning-article

[2] "What is data visualization?" Nov 2023. [Online]. Available: https://www.ibm.com/topics/data-visualization

[3] "Data cleaning: Definition, benefits, and how-to — tableau," Nov 2023. [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-cleaning

[4] "Difference between data cleaning and data processing," Nov 2023. [Online]. Available: https://www.geeksforgeeks.org/difference-between-data-cleaning-and-data-processing/

[5] "Data visualization infographic," Nov 2023. [Online]. Available: https://venngage.com/templates/diagrams/data-viz-edf4207a-7fb4-47ce-ae1e-df730cc93d78

[6] "What is feature engineering? definition and faqs — heavy.ai," Nov 2023. [Online]. Available: https://www.heavy.ai/technical-glossary/feature-engineering

[7] "What is feature engineering — importance, tools and techniques for machine learning," Nov 2023. [Online]. Available: https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

[8] "The penguins datasets — scikit-learn course," Nov 2023. [Online]. Available: https://inria.github.io/scikit-learn-mooc/python_scripts/trees_dataset.html

[9] "palmerpenguins," Nov 2023. [Online]. Available: https://github.com/allisonhorst/palmerpenguins

[10] "pandas: Powerful data structures for data analysis, time series, and statistics," Nov 2023. [Online]. Available: https://pandas.pydata.org

[11] [Online]. Available: https://allisonhorst.github.io/palmerpenguins/

[12] W. contributors, "Mutual information," Nov. 2023. [Online]. Available: https://en.wikipedia.org/wiki/Mutual_information

[13] H. Ertan, "Feature selection methods in scikit learn — medium," *Medium*, Nov. 2023. [Online]. Available: https://medium.com/@hertan06/which-features-to-use-in-your-model-350630a1e31c

[14] Z. Jaadi, "A step-by-step explanation of principal component analysis (pca)," Apr. 2021. [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[15] [Online]. Available: https://www.ibm.com/topics/random-forest

[16] W. Z. Trivelpiece, S. G. Trivelpiece, and N. J. Volkman, “Ecological segregation of adelie, gentoo, and chinstrap penguins at king george island, antarctica,” *Ecology*, vol. 68, no. 2, p. 351–361, Apr. 1987. [Online]. Available: https://doi.org/10.2307/1939266

[17] [Online]. Available: https://scikit-learn.org/stable/auto_examples/feature_selection/plot_f_test_vs_mi.html

List of Figures

List of Tables