Introduction

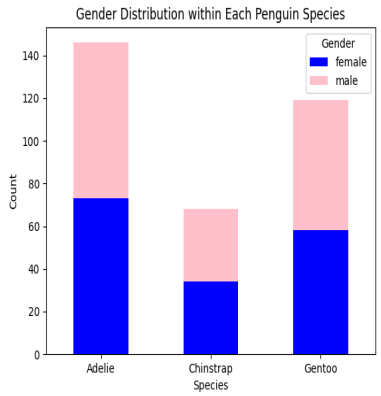
This report explores the "Palmer Penguins" dataset, which details three penguin species [1] in the Palmer Archipelago, Antarctica. It aims to identify penguin species using machine learning techniques, assessing data types and preprocessing to handle missing entries. The study tests various features through unsupervised and classification algorithms, comparing outcomes with a baseline to validate the methods. The goal is to derive insightful and constructive feedback from the analysis.

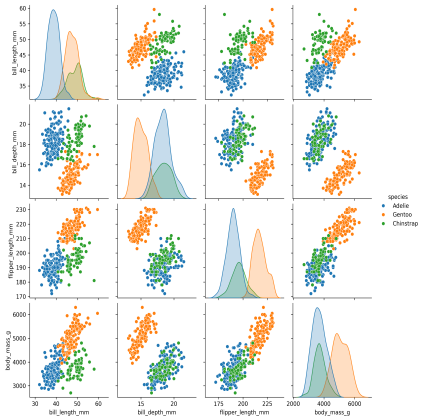
Exploratory data analysis

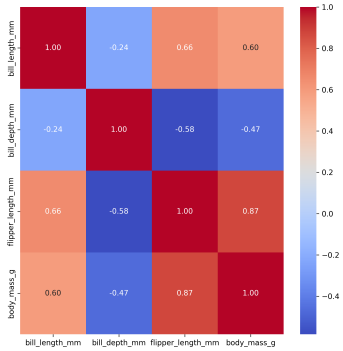
Exploratory Data Analysis (EDA) is a crucial step in understanding the characteristics of a dataset. This section will provide a detailed introduction to EDA process for the "Palmer Prnguins" datasets. This section reviews the penguin dataset, revealing it contains 344 records, each with 9 attributes. Using Python, we can extract details like species, gender, habitat, and various biometric data. Notably, there are missing entries for gender and some measurements.

preprocess data and Feature selection

The initial data inspection shows that most of the data are complete, but some missing values still exist. Specifically, there are some missing data points. Addressing these missing values is important for subsequent data analysis and prediction. These data can be handled by using median or mean values. Additionally, some columns, like "rowid" and "year of observation," can be removed initially, because they have no impact on future predictions and analysis. Since the gender ratio within each penguin species is nearly evenly distributed, with males and females almost equally represented, gender is not considered a classifying feature. The scatter plot matrix assesses relationships between penguin biometrics, differentiating species with colors: Adelie (blue), Gentoo (green), Chinstrap (orange). It highlights how "bill length vs. flipper length" provides clear species separation, while "body mass" and "bill depth" show significant overlap, making them less effective for classification. High correlation between certain metrics aids in prediction but can also complicate it due to overlap. Consequently, "flipper length" and "bill length" are chosen for further analysis and model training due to their distinct inter-species distribution.







Classification

k-nearest neighbors algorithm

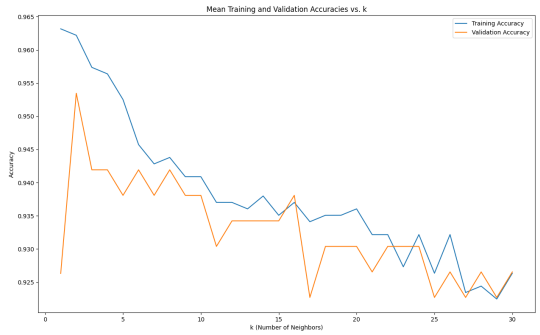
The KNN algorithm is a straightforward [2], non-parametric supervised learning method that uses proximity to classify or predict data points. It requires no assumptions about the data, making it suitable for biological datasets like the Palmer Penguins. KNN has few hyperparameters, mainly the number of neighbors (k) and the distance metric, which simplifies its use and implementation. It performs efficiently on small to medium-sized datasets without demanding extensive computational resources.
After initially dividing the dataset into training and testing sets, KNN uses the training set with `knn.fit(X_train, y_train)`. In this phase, KNN examines the relationships between features and labels to identify patterns, calculating distances using Euclidean metrics and selecting *k* nearest neighbors to build its prediction mechanism.

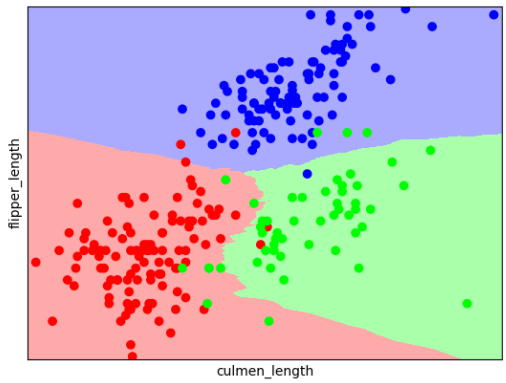
$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{1}$$

Finally, it uses majority voting to make classification decisions.

$$y = \text{mode}\{y_1, y_2, \ldots, y_k\} \tag{2}$$

K-fold cross-validation divides the dataset into k subsets [3], training on k-1 subsets while testing on the remaining subset, rotating until each subset serves as the test set. This method effectively estimates the model's performance by averaging the outcomes from all k iterations, ensuring comprehensive utilization of the dataset for training and testing. To optimize performance, this model requires selecting suitable hyperparameters, such as the number of neighbors (k), achieved through cross-validation. The ideal k value leads to the highest and closest training and validation accuracies, as demonstrated by k = 6 and k = 16 in the KNN k-fold graph.





Decision tree classifier

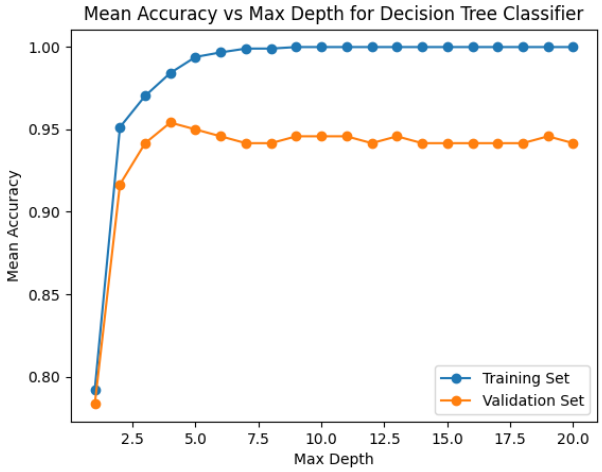
Decision trees [4] emulate human-decision-makning processes, This is valuable for explanation of model prediction. What is more, it requires low preprocessing for data and the capability to handle non-linear datasets. This method can identify penguin species based on penguin biometric characteristics. Although decision trees can cause over-fitting, but this issue can be solved by adjusting hyperparameters, such as the depth of the tree and the minimum number of samples per leaf.
In this report, the 'DecisionTreeClassifier' uses entropy as its criterion for splitting nodes, which is one rule for node division in decision tree. Entropy can capture the complicated relationships within the biometric

data of penguins and handle multiple classification like three species classfication in penguin.

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{3}$$

- $H(S)$ is the entropy of set SS,
- pi is the proportion of the class i elements within set S,
- n is the number of classes.

In this part, k-fold cross-validation is also used to select the best hyperparameter for the decision tree depth. Both training accuracy and validation accuracy are observed to determine at which depth the validation achieves the highest accuracy and has the smallest gap with training accuracy. A significant difference between them might indicate overfitting or underfitting, suggesting that the model is either too complex or too simplistic. In this case, depths of 3 and 4 would be decent values to achieve high accuracy.



Baseline Testing and Comparison

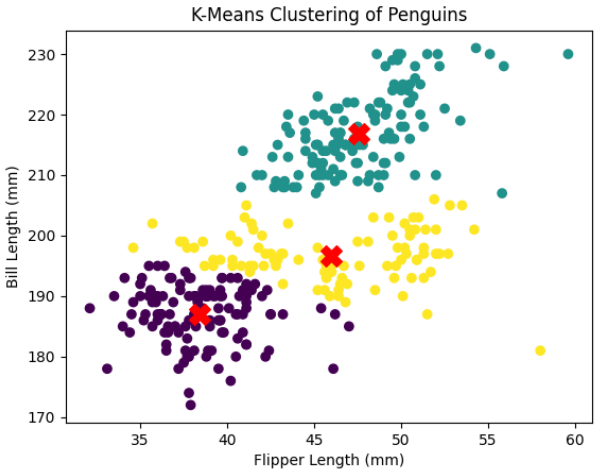
A dummy classifier, also known as a baseline classifier or a null model, is a simple machine learning model that focuses on data distribution or simple rules in a dataset. If the performance of the KNN or Decision Tree is inferior to that of the dummy classifier, it suggests that these models may not be suitable for penguin classification, or some technical issues exist in the training process. Moreover, by comparing with a dummy classifier, overfitting can also be checked. If KNN and Decision Trees show higher accuracy on the training set compared to the dummy classifier but fail to demonstrate similar improvements on unseen data, it could indicate overfitting.

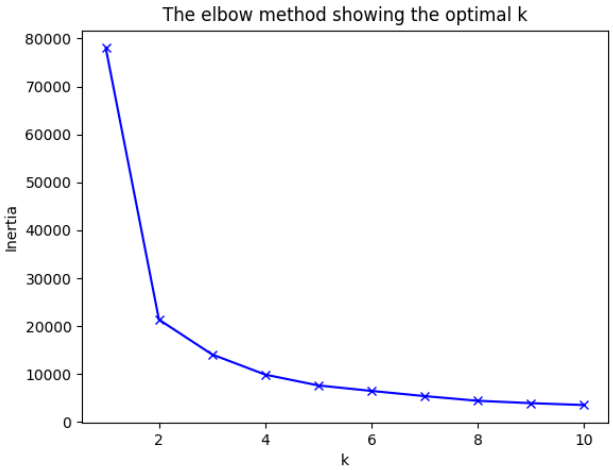
	Dummy Classifier	**KNN**	**Decision Tree**
Accuracy	46.51%	96.5%	94.23%
Precision score	15.5%	95.76%	93.33%
Recall score	33%	95.24%	92.87%
F1 score	21.2%	95.43%	93.05%

The dummy classifier as a baseline achieved 46.61% accuracy, 15.5% precision, 33% recall, and an F1 score of 21.2%. Both the KNN and decision tree models significantly outperformed the baseline. This indicates that they have better learning ability and generalization than simple or random chance. Every metric from the KNN model surpasses 95%, and the decision tree also performed well in this task, reaching over 90% in all aspects, although it was not as distinctively efficient as the KNN. This suggests that KNN might be a better model for this specific dataset, but the decision tree also shows that it can make great predictions.

K-means Cluster (Unsupervised Learning)

Penguin species can be classified by biometric characteristics [5], like flipper length and bill length, which vary between species. K-means can identify these variations as it calculates the inertia, or within-cluster sum of squares, which it seeks to minimize. Since penguins evolve separately, it is reasonable to assume that they would form clusters based on observed physical features.
Using the Elbow method, inertia represents how internally coherent the clusters are against the number of clusters (k). This indicates that coherence does not necessarily decrease with more clusters. The point of inflection on the plot suggests the best k (number of clusters). For example, according to the graph, k=3 would be ideal for a K-means model. After determining the best k, K-means initializes the cluster centroids randomly. Each centroid represents one cluster. The centroid is then recalculated as the mean data value inside the cluster, efficiently finding the center point. Subsequently, each data point is assigned to the cluster with the nearest (newly calculated) centroid. This iterative process continues until the centroids stabilize and the cluster assignments no longer change, indicating convergence.





Silhouette Score: 0.485 This score measures the degree of separation between clusters, ranging from -1 to 1. A high score indicates well-separated clusters. A score of 0.485 suggests moderate separation. While not close to 1, this score indicates that the clusters are more distinct than overlapping, but there's room for improvement.

Homogeneity Score: 0.635 This score examines if all clusters contain only data points that are members of a single class. A score of 0.635 indicates a moderate level of homogeneity. This means that the clusters largely contain data points from a single class, but some mixed points are included, reducing the score.

Challenge: Racial and Socioeconomic Disparities in Credit Scoring

According to research and report statements, compared to white or higher-income groups [6], people from minority and low-income backgrounds typically have lower credit scores. This can be attributed to several reasons: **Historic Inequality:** Historically, minorities and low-income individuals have had fewer opportunities to receive credit and financial services. This has limited their chances of establishing good credit. **Data Bias:** Since credit scores might inherit past biases, they rely heavily on historical data. This data may reflect more commonly experienced discriminatory practices or socio-economic disadvantages in some communities.

Ethical Issue: Transparency: Consumers often do not have sufficient information about how their credit scores are calculated, what factors affect them, or methods for improving their scores.**Influence:** A low credit score can limit opportunities to apply for loans and basic financial products like mortgages. It can also impact the accumulation of assets and affect employment opportunities, as credit checks are sometimes used during the hiring process.

Addressing the Issue: Algorithmic Fairness: Apply advanced machine learning techniques to identify and mitigate biased data in models. This includes using fairness-aware algorithms that can adjust for imbalances in data which might disproportionately affect minority groups. **Policy Change:** Encourage policies that limit the use of credit checks in job applications, aimed at reducing discrimination based on poor credit history, which may stem from factors beyond an individual's control. **Reinforce Supervision:** Implement stricter regulations that require regular audits of credit scoring algorithms to ensure they do not perpetuate historical biases. Ensure that the calculation of credit scores is transparent, helping consumers understand how the system works and informing them of ways to improve their credit score.

Challenge: AI in Mental Health Interventions

Description: French artificial intelligence startup company Nabla is testing a chatbot robot based on GPT-3 [7] in a healthcare setting. The system warns that the robot may ask sensitive questions about suicide. The chatbot's inappropriate response—"I think you should"—underscores the risks involved in deploying such technology in healthcare settings, which may affect human well-being and provide life advice.

Safety Threat: There is a direct risk to individuals' physical and mental health when AI systems provide harmful advice.**Reliability Threat:** AI chatbots, especially those based on large language models like GPT-3, can produce unpredictable statements due to their training sets that inevitably include unsafe, biased, or inappropriate content.**Ethical Threat:** The use of artificial intelligence may impact vulnerable individuals, which could raise ethical concerns regarding the development of AI and the responsibilities of those deploying it.

Addressing Ethical Problems and Enhancing Transparency

Reliability Testing: regularly test AI systems across a wide range of scenarios to evaluate their responses in sensitive situations. This would help identify potential flaws in the model’s understanding and interaction capabilities.**Ethical Guidelines and Training:** Design and adhere to ethical principles for healthcare applications and the use of artificial intelligence. These guidelines should prioritize patient safety and well-being. Train AI developers and stakeholders on ethical issues specific to AI use in sensitive areas. **Audit Trails:** Continuously update comprehensive logs of AI interactions to ensure traceability and accountability. This is not only beneficial for assessing the performance of AI but also supports investigations into cases where advice from AI might lead to adverse outcomes.

Bibliography

1. C. Mazzaroli, “Exploratory Data Analysis to Predictive analysis (Palmer Penguins),” *Deepnote*. https://deepnote.com/app/mazzaroli/Exploratory-Data-Analysis-to-Predictive-analysis-Palmer-Penguins-e6bd8932-9ca1-4363-b78a-79f4f4dc818f (accessed Apr. 29, 2024). ↵

2. IBM, “What is the k-nearest neighbors algorithm? | IBM,” *www.ibm.com*, 2023. https://www.ibm.com/topics/knn ↵

3. J. Brownlee, “A Gentle Introduction to k-fold Cross-Validation,” *Machine Learning Mastery*, Oct. 04, 2023. https://machinelearningmastery.com/k-fold-cross-validation/ ↵

4. IBM, “What is a Decision Tree | IBM,” *www.ibm.com*, 2023. https://www.ibm.com/topics/decision-trees ↵

5. Wikipedia Contributors, “k-means clustering,” *Wikipedia*, Feb. 22, 2019. https://en.wikipedia.org/wiki/K-means_clustering ↵

6. J. Ney, “Credit Scores and Inequality,” *Age of Awareness*, Oct. 11, 2021. https://medium.com/age-of-awareness/credit-scores-and-inequality-1df9d80074d2 ↵

7. T. Writer, Entrepreneur, M. over Marketing, F. Geek, and C. Thinker, “Famous AI Gone Wrong Examples In the Real World we Need to Know,” *Analytics Insight*, Mar. 09, 2021. https://www.analyticsinsight.net/famous-ai-gone-wrong-examples-in-the-real-world-we-need-to-know/ ↵