C964: Computer Science Capstone

Enhancing Conservation Efforts: Building a Classification Model for Penguin Species Identification

Author: Jessica Short

Student ID: 010500487

Date: 12/20/2024

Part A: Letter of Transmittal

December 20, 2024

Carol Young
Executive Director
World Penguin Conservation Organization (WPCO)
50 Broad Street #1507
New York, NY 10004

Dear Mrs. Young,

My name is Jessica Short, and I am a cofounder and Senior Machine Learning Engineer at ThinkShort AI Solutions. I am writing to propose a machine-learning-based solution aimed at enhancing the efficiency, safety, and ethical practices of penguin species identification during field studies, directly supporting the World Penguin Conservation Organization's vital conservation efforts.

The World Penguin Conservation Organization (WPCO) has made significant contributions to the study and protection of penguins, often operating in Antarctica, where harsh weather conditions and time constraints are prevalent. A key aspect of these studies involves identifying penguin species—a process that can be time-consuming, disruptive to the animals, and potentially hazardous to field researchers.

To address these challenges, I propose the development of a non-invasive, accurate classification tool that utilizes machine learning to quickly identify penguin species based on simple physical measurements—bill length and bill depth. This tool will enable researchers to make accurate species identifications in seconds, improving efficiency while minimizing disruptions to the penguins and reducing the time researchers spend in extreme conditions, thereby enhancing their safety.

Proposed Solution

The solution is a machine-learning-based classification tool designed to predict penguin species based on physical measurements. It will be implemented in a Jupyter

Notebook hosted on Google Colab, ensuring ease of access and compatibility with remote field environments.

Key features of the tool include:

- Interactive Interface: Powered by Jupyter Notebook widgets, the interface will allow users to input bill length and bill depth measurements through intuitive fields. After clicking a button, the tool will return the predicted penguin species almost instantly.
- Performance Metrics: To ensure transparency and reliability, the tool will display an accuracy score and a confusion matrix, enabling users to assess the model's effectiveness in distinguishing between species.
- Ease of Use and Accessibility: Designed for simplicity and portability, the tool supports researchers in fieldwork with minimal disruption to their workflow or the animals being studied.

The model will be trained using the Palmer Penguins Dataset, available from Kaggle.com. This dataset, collected between 2007 and 2009 as part of the Palmer Station Long-Term Ecological Research program, provides high-quality data for this application.

Timeline

I anticipate that this project will take approximately 3 months to complete, with the following estimated timeline:

1. Business Understanding (2 weeks): Define objectives, understand constraints, and align project goals.
2. Data Understanding (1 week): Analyze the dataset to identify key patterns and ensure suitability for the model.
3. Data Preparation (2 weeks): Clean and preprocess the data to optimize performance.

4. Modeling (4 weeks): Develop, train, and refine the machine learning model for species classification.
5. Evaluation (1.5 weeks): Test the model for accuracy and ensure it meets performance requirements.
6. Deployment (1.5 weeks): Finalize the tool and deploy it in a Jupyter Notebook hosted on Google Colab.

Cost Estimate

The estimated cost for this solution is $110,116, covering all aspects required for its successful development and deployment. This includes:

- Salaries for machine learning engineers, a software developer and a project manager
- Hosting and platform fees for ensuring tool accessibility on Google Colab
- Integration and use of necessary machine learning frameworks and libraries
- Development and delivery of user training sessions and materials to ensure ease of use in the field
- Procurement of measuring tools (e.g., digital calipers) for fieldwork and data collection

Expected Benefits

By implementing this solution, the World Penguin Conservation Organization will experience the following advantages:

1. Increased Efficiency: Researchers will save valuable time in the field, allowing for more comprehensive data collection and observation. Additionally, the streamlined process will reduce research costs for the organization by minimizing the need for additional resources or extended fieldwork.
2. Non-Invasive Methods: The tool minimizes disturbance to penguins, supporting ethical research practices.

3. Safer Working Conditions: By reducing the time spent in extreme weather conditions, the tool helps mitigate the risk of injury to field researchers.
4. Cost-Effective Solution: Requiring only a digital caliper, the tool eliminates the need for expensive or complex laboratory analysis.
5. Enhanced Data Accuracy: The model reduces the potential for human error, ensuring more reliable species identification.

Qualifications

If you choose to proceed with this proposal, I will oversee all phases of the project to ensure its successful completion. With over 15 years of experience in machine learning and artificial intelligence, I bring the expertise needed to deliver impactful results.

My qualifications are further supported by a similar project I led in 2022 for the World Wildlife Fund (WWF). During that project, my team developed a U-Net model to map orangutan habitats in Borneo and Sumatra. The model achieved 95% pixel accuracy in distinguishing habitat types and successfully mapped 92% of orangutan habitats using high-resolution satellite imagery.

Next Steps

I kindly request your feedback and a formal decision on this proposal by January 31, 2025. Should you have any questions or require additional details, please feel free to contact me via phone or email. I am excited about the opportunity to collaborate with your organization and contribute to your critical conservation efforts.

Sincerely,

Jessica Short

Jessica Short

Senior Machine Learning Engineer

ThinkShort AI Solutions
61 Broadway #401
New York, NY 10006
Phone: (555) 555-8822
Email: jessicashort@thinkshortai.com

Part B: Project Proposal Plan

PROJECT SUMMARY

i. Problem Description

The World Penguin Conservation Organization (WPCO) plays a critical role in studying and conserving penguin species, particularly in the Antarctic region. One major challenge is the accurate, timely, and ethical identification of penguin species. Traditionally, researchers rely on physical measurements and observations, such as body mass, flipper length, behavioral cues, and plumage characteristics. However, this method is both time-consuming and prone to human error. Additionally, other techniques like blood/tissue sampling, tagging, and trapping are invasive and can harm penguins, causing stress, exhaustion, and even nest abandonment (Carroll, Turner, Dann, & Harcourt, 2016). Extreme weather conditions further complicate the data collection process and pose significant risks to the safety of field researchers, increasing the chances of injuries. Moreover, traditional methods are susceptible to errors, diminishing the accuracy and reliability of the data.

To overcome these challenges, there is a pressing need for a more efficient, accurate, and non-invasive solution for species identification. A machine-learning-based tool can provide rapid and precise species predictions with minimal disruption to the penguins, enhancing both researcher safety and field research efficiency.

ii. Client Needs

The World Penguin Conservation Organization (WPCO) spends approximately $2.5 million each year on fieldwork, monitoring, and research. By adopting our proposed solution, WPCO could reduce these costs by at least 10%. These savings stem from reduced labor hours spent on species classification in the field, as the new tool streamlines the process and allows researchers to work more efficiently. The savings can be reinvested into other initiatives, such as further research, expanding conservation programs, developing additional technological innovations, or increasing outreach and educational efforts to engage the public in penguin conservation.

Field researchers observing penguins in Antarctic regions face extreme weather conditions, including high winds, blowing snow, freezing temperatures, icy terrain, and low visibility. These hazards can lead to injuries such as frostbite, hypothermia, and falls. In 2023, WPCO reported 3 researcher injuries related to extreme weather, costing the organization $45,600. By adopting our proposed solution, WPCO could reduce researcher injuries by at least 30%. This reduction is primarily due to the tool’s ability to significantly shorten the time researchers spend exposed to the elements while measuring penguin physical features for classification.

The new solution will also improve penguin well-being by enabling researchers to identify species quickly and non-invasively. By using simple physical measurements, the model allows species classification with minimal physical interaction, reducing the risk of stress or harm to the penguins. Traditional methods of species classification can involve more invasive techniques that pose risks to penguin well-being. Such methods include blood/tissue sampling, tagging, and trapping/releasing. These invasive techniques can result in adverse effects such as stress, panic, exhaustion, weakened immune systems, and nest abandonment (Carroll, Turner, Dann, & Harcourt, 2016). Our proposed penguin classification tool aligns with ethical research practices, ensuring that penguins are studied with minimal disruption to their natural behaviors and environment, ultimately promoting their well-being.

iii. Deliverables

A description of all deliverables for this project are provided below:

1. **Penguin Classification Tool –** This tool will be an interactive, machine learning-based penguin species classification tool, implemented as a Jupyter Notebook application hosted on Google Colab. The tool will use an algorithm to predict penguin species based on physical measurements—bill length and bill depth. Featuring an intuitive user interface, researchers will be able to input measurements and receive species predictions in seconds. This tool will enable field researchers to quickly and accurately classify penguin species, significantly reducing the need for more invasive and time-consuming methods. The tool's interactive design and real-time predictions make it an invaluable resource for enhancing research efficiency while minimizing disruption to penguin well-being.

2. **Descriptive Methods and Visualizations Tool -** This tool will help users gain a deeper understanding of the Palmer Penguins Dataset. Hosted on Google Colab as a Jupyter Notebook application, the tool will allow users to explore key penguin features, such as bill length, bill depth, flipper length, and body mass, through various visualizations, including histograms, scatterplot matrices, and scatter plots. These visualizations will provide valuable insights into the data's distribution and potential relationships between features, such as how bill length and bill depth vary across different penguin species. By using this tool, users will be able to identify trends, correlations, and potential outliers, supporting informed decision-making when preparing for machine learning analysis. The tool's exploratory nature enhances understanding of the data's structure and supports feature selection, ultimately improving the effectiveness of subsequent predictive modeling.

3. **User Guide for the Penguin Species Classifier Tool –** This guide will provide step-by-step instructions for field researchers on how to effectively use the machine learning-based tool for penguin species identification.

4. **User Guide for the Descriptive Methods and Visualizations Tool –** This guide will provide step-by-step instructions for users on how to effectively use the Descriptive Methods and Visualizations Tool.

5. **Post Implementation Report –** This report will provide a comprehensive overview of the project, focusing on the technical aspects of the solution. It will explain the data source, its collection process, and how the data was processed throughout the project's lifecycle. The report will cover the machine learning method used, detailing what the method does, how it was implemented, and why it was selected for this specific use case. It will also include a thorough explanation of the validation process, with performance metrics to assess the model's effectiveness. Additionally, the report will highlight visualizations used to explore the data and demonstrate relationships between features.

6. **Training Sessions and Training Guide –** This deliverable will consist of training sessions for WPCO field researchers and employees, scheduled for April 2025. These sessions will provide an overview of the penguin species classification tool, including its key features, functionality, and benefits. Participants will experience live demonstrations and engage in hands-on practice, learning how to input measurements, generate predictions, and interpret results. A comprehensive

Training Guide will be provided, offering detailed instructions, best practices, and troubleshooting tips to support ongoing use of the tool in the field.

Project Benefits

The implementation of the machine-learning-based penguin species classification tool will provide several benefits for the World Penguin Conservation Organization (WPCO). These benefits are outlined below:

1. **Increased Efficiency**: The tool will streamline the species identification process by offering quick, accurate predictions based on simple physical measurements. This will significantly reduce the time researchers spend on data collection and analysis. As a result, researchers will be able to process more data in less time, enhancing the overall efficiency of their fieldwork. It is estimated that fieldwork, monitoring, and research costs will decrease by at least 10%, which translates to savings of approximately $250,000 for WPCO.

2. **Ethical and Non-Invasive**: Unlike traditional methods such as blood or tissue sampling, tagging, trapping, and weighing, the proposed classification tool offers a less invasive approach to species identification. This solution will help minimize disruption to penguin well-being, ensuring that research is conducted ethically and in alignment with conservation goals. By reducing the need for invasive procedures, the tool supports WPCO's commitment to minimizing stress and harm to penguin populations during research activities.

3. **Improved Data Accuracy**: By utilizing a machine learning model, the risk of human error in manually classifying penguin species is significantly reduced, leading to more consistent and dependable results. This approach ensures precise identification of species based on field measurements, providing a solid foundation of high-quality data that enhances the integrity and credibility of WPCO’s conservation efforts.

4. **Safer Fieldwork**: WPCO’s field research primarily takes place in Antarctica, where extreme weather conditions pose significant risks to researchers. Prolonged exposure to such environments can lead to exhaustion or injury. By streamlining the species identification process, the tool reduces the time researchers need to spend in the field, significantly lowering their exposure to

these hazards. This advancement not only enhances researcher safety and well-being but also allows for quicker data collection. WPCO estimates that the use of this tool will reduce researcher injuries by at least 30%.

5. **Cost-Effective**: The new model requires only a digital caliper for physical feature measurements, making it a cost-effective solution for species identification. In contrast to laboratory-based methods that demand costly equipment and materials, the tool's simplicity offers a low-cost alternative, allowing WPCO to conduct extensive research without the burden of additional financial resources.

DATA SUMMARY

i. Data Source

The model will be trained using the Palmer Penguins Dataset, available from Kaggle.com. This data was collected between 2007 and 2009 as part of the Palmer Station Long-Term Ecological Research program. The Palmer Penguins Dataset can be found at https://www.kaggle.com/datasets/satyajeetrai/palmer-penguins-dataset-for-eda/data.

ii. Data Source Advantages

The Palmer Penguins Dataset is ideal for this project due to its relevance, high quality, diversity, accessibility, and sufficient size. The dataset is of high quality, having been collected as part of the Palmer Station Long-Term Ecological Research program, ensuring its scientific validity and reliability. It offers a diverse range of data, including multiple penguin species such as Adelie, Chinstrap, and Gentoo, allowing the model to learn distinct physical traits and improving its ability to generalize. The dataset is easily accessible on Kaggle, making it simple for the team to download and begin data exploration and model development quickly. Its open availability eliminates legal or licensing barriers, enabling smooth integration into the project workflow. Furthermore, the dataset is sufficiently large, containing hundreds of samples that will help the model identify patterns, reduce overfitting, and improve prediction accuracy.

iii. Data Processing and Management

Data will be processed and managed during each phase of the project as follows:

Phase 1: Business Understanding – During this phase, the data processing requirements will be defined based on the project's objectives. This includes identifying the types of data needed, determining data sources, and setting data quality expectations. A plan will also be developed for how the data will be stored, secured, and processed.

Phase 2: Data Understanding – Raw data will be collected from various sources and analyzed for quality, structure, and relevance. Exploratory data analysis (EDA) will be performed to identify patterns, inconsistencies, and missing values.

Phase 3: Data Preparation – In this phase, the data will be preprocessed and prepared for modeling. Tasks will include handling missing values, addressing outliers, and splitting the dataset into training and testing subsets. The prepared data will be securely stored to ensure easy access during the modeling phase.

Phase 4: Modeling – The processed data will be used to train machine learning models. This includes selecting appropriate algorithms, tuning hyperparameters, and validating model performance.

Phase 5: Evaluation – The model's predictions will be evaluated using test data to ensure they are accurate and reliable. Any discrepancies or issues with the data will be addressed, and evaluation results will guide adjustments to the preprocessing or modeling processes. The final processed data and evaluation results will be documented for deployment.

Phase 6: Deployment – During the deployment phase, the trained and validated model will be integrated into the application. Users will preprocess incoming data, with the application prompting them to round measurements and ensure they fall within a specified range. The species identification results will be securely stored for further analysis and reporting. If any discrepancies or issues in the data pipeline arise, ThinkShort AI Solutions can be consulted for assistance. While periodic updates are not part of this project's scope, they remain a potential option for the future.

iv. Handling Data Anomalies

The Palmer Penguins Dataset may contain data anomalies, such as outliers and missing values, which could affect the quality and reliability of the model. These anomalies will be systematically addressed during the data preparation phase to ensure the dataset is ready for training.

1. **Outliers**: Extreme values that differ significantly from the overall data distribution will be identified through exploratory data analysis (EDA) and visual tools. Depending on their influence on the model's performance, outliers may either be removed or adjusted to reduce their impact on the model's accuracy. If outliers reflect legitimate, infrequent occurrences, they may be kept to ensure the model is capable of handling such variations.
2. **Missing Data**: Missing values will be detected and handled appropriately. If the amount of missing data is minimal, imputation methods, such as filling missing values with the mean, median, or mode, will be applied.

By addressing these anomalies, the dataset will be cleaned and prepared for model training, ensuring high-quality, consistent data that enhances the model's performance and accuracy.

v. Data Source Ethical and Legal Concerns

The Palmer Penguins Dataset, available on Kaggle and part of the Palmer Station Long-Term Ecological Research (LTER) program, is legally suitable for academic and research use. Collected through scientific research focused on monitoring penguin populations, the data was likely gathered ethically, with minimal impact on the animals. Since the dataset does not contain personal or sensitive information, privacy concerns are not relevant.

IMPLEMENTATION

For this solution, the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology will be implemented. The work that will be completed in each phase is summarized below. (Hotz, 2024).

Phase 1: Business Understanding – During this phase, the team will work closely with stakeholders to define project objectives, milestones, deliverables, and success metrics. A risk assessment will be carried out to identify potential challenges. Additionally, the team will develop a comprehensive project plan that details the remaining phases, including key milestones, timelines, and resource allocation.

Phase 2: Data Understanding – In this phase, the team will retrieve the Palmer Penguins Dataset from Kaggle.com. Exploratory Data Analysis (EDA) will be conducted to examine the dataset's structure, assess its quality, and identify trends and relationships among variables. The team will use visualizations such as histograms and scatter plots to explore the distribution of features and their relationships with the target variable. Insights gained from EDA will guide subsequent data preprocessing steps to ensure the data is clean and ready for modeling.

Phase 3: Data Preparation – In this phase, the team will clean the dataset by addressing missing values, handling outliers, and standardizing measurements if necessary. Key features, such as bill length and bill depth, will be selected and preprocessed for use in the model. The data will be split into training and testing subsets.

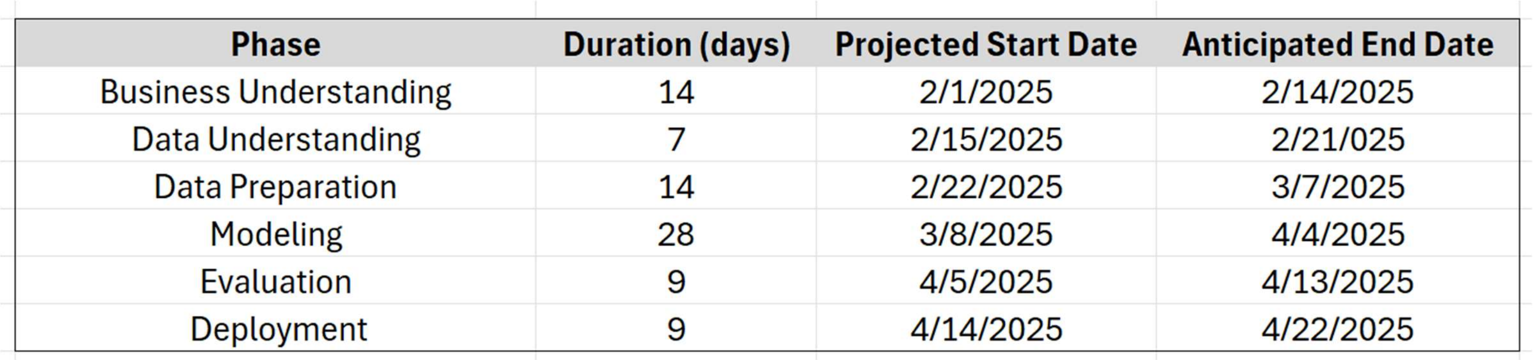
Phase 4: Modeling – In this phase, the team will select the most suitable algorithms for classification. The initial model will be trained on the prepared dataset. If needed, the model will be tuned to optimize performance.

Phase 5: Evaluation – In this phase, the final model will be evaluated on the test dataset. Performance metrics will be generated. The model will be reviewed against the success criteria established during the Business Understanding phase, and its readiness for deployment will be confirmed to ensure it meets all requirements.

Phase 6: Deployment – In this phase, the final model will be incorporated into a Jupyter Notebook hosted on Google Colab. A user-friendly interface will be created to enable input of bill length and depth for species predictions. The tool will undergo testing in simulated field conditions to verify its reliability and ease of use. Comprehensive documentation and training will be provided for WPCO staff and researchers, and the finalized tool, along with all relevant documentation, will be delivered for final deployment.

TIMELINE

The table below outlines the projected timeline for the project, with six key milestones corresponding to the phases of the CRISP-DM development cycle. For each milestone, the table provides the estimated duration, as well as the start and end dates.



Phase	Duration (days)	Projected Start Date	Anticipated End Date
Business Understanding	14	2/1/2025	2/14/2025
Data Understanding	7	2/15/2025	2/21/025
Data Preparation	14	2/22/2025	3/7/2025
Modeling	28	3/8/2025	4/4/2025
Evaluation	9	4/5/2025	4/13/2025
Deployment	9	4/14/2025	4/22/2025

Evaluation Plan

This Evaluation Plan outlines the verification and validation methods that will be employed at various stages of the project to ensure its success. These methods will evaluate both the technical functionality of the tools and the overall impact on WPCO's objectives.

1. Verification Methods during Development Phases:

Phase 1: Business Understanding

- Verification Method: A detailed review of project objectives, milestones, and success metrics will be conducted with stakeholders to confirm alignment with the needs of WPCO. This will ensure the project is on track to meet client expectations.

- Verification Activity: Risk assessment meetings, stakeholder interviews, and project planning documentation reviews.

Phase 2: Data Understanding

- Verification Method: Data quality checks will be performed during the exploratory data analysis (EDA) to ensure that the data is complete, accurate, and relevant for model development. Outliers and missing values will be identified, and necessary steps will be taken to handle them.
- Verification Activity: Data visualization (scatter plots, histograms) and summary statistics will be reviewed by the team to confirm the correctness of data features and distributions.

Phase 3: Data Preparation

- Verification Method: During the data preprocessing phase, checks will be made to ensure that all features are correctly processed, and the dataset is ready for training.
- Verification Activity: Spot checks and scripts for missing data imputation, outlier detection, and feature standardization.

Phase 4: Modeling

- Verification Method: The initial machine learning models will be evaluated to check for overfitting and underfitting. This step will ensure the selected model performs well on unseen data.
- Verification Activity: After training, the model's accuracy will be checked on both training and test datasets. A large gap in performance suggests overfitting (high accuracy on training, low accuracy on test), while poor performance on both indicates underfitting (IBM, n.d.).

Phase 5: Evaluation

- Verification Method: The final model will be tested on the testing subset to ensure that it generalizes well to new data. The tool's user interface will also be tested to ensure it is functional and easy to use.
- Verification Activity: Performance metrics, such as accuracy, confusion matrix, and user interface feedback, will be reviewed. User feedback will also be collected from internal testing sessions to identify areas for improvement.

Phase 6: Deployment

- Verification Method: In this phase, the model will be deployed into the production environment on Google Colab. It will be verified through field simulations and user testing to ensure it meets real-world conditions.

- Verification Activity: User acceptance testing (UAT) and end-to-end testing with real data from the field will be conducted to confirm that the system works as intended.

2. Validation Method upon Completion of the Project

Upon completion of the project, the validation process will assess whether the solution meets WPCO's objectives of improving research efficiency, reducing costs, ensuring ethical practices, and enhancing researcher safety. Some of these evaluations may extend beyond the project's completion date (April 22, 2025), as additional time will be required to gather financial, safety, and efficiency data. The validation process is expected to be completed by 2026.

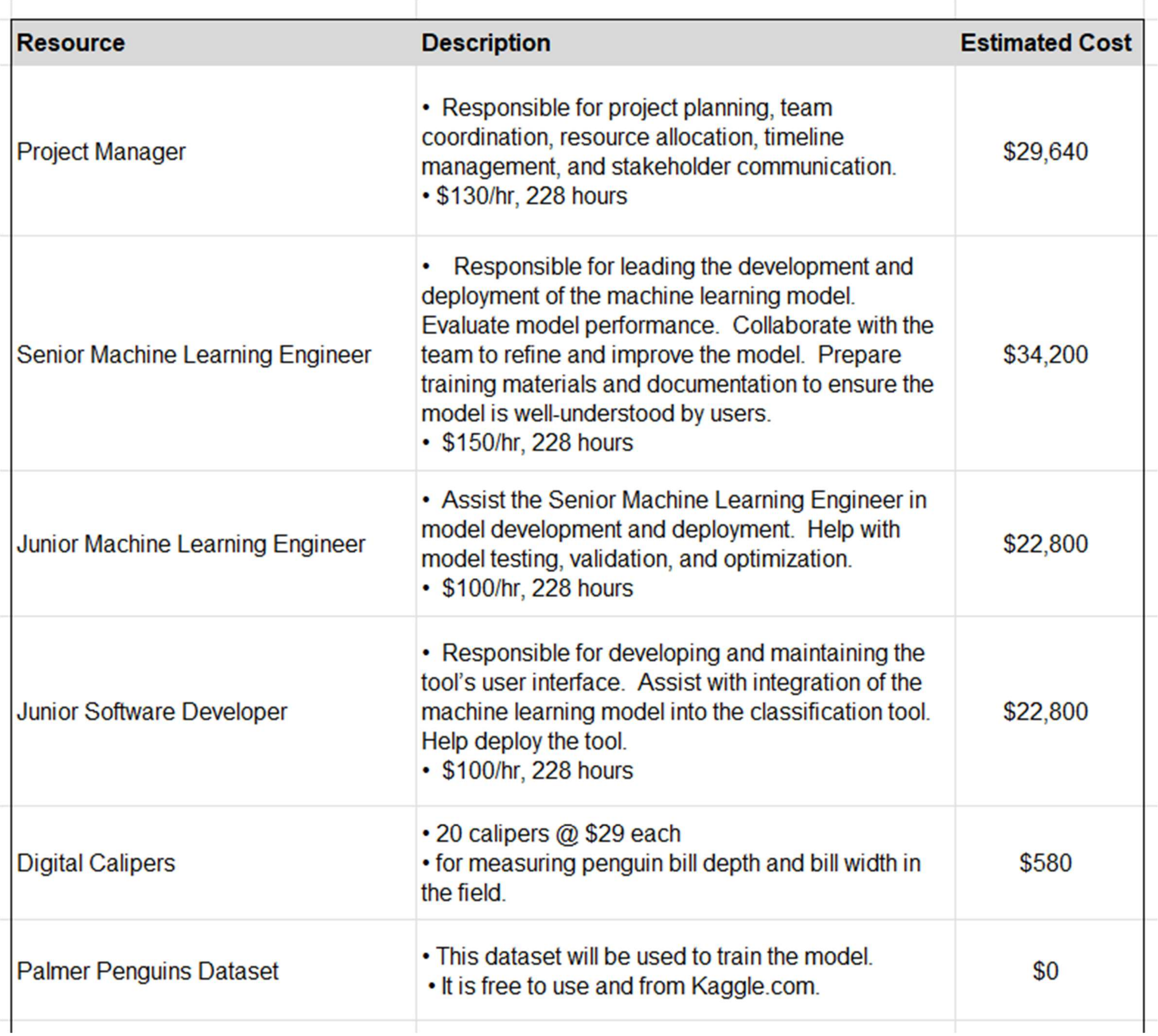
Validation Methods:

- **Accuracy and Reliability of the Model:** The final model will be validated against real-world test data. The Central Park Zoo has agreed to allow the team to test the model on 35 Gentoo and Chinstrap penguins, as the zoo does not currently house Adélie penguins (Penguins, n.d.). The model's accuracy will be assessed using the accuracy score, with the expectation of achieving at least 90% accuracy in species identification with this test group. Additionally, the model will be tested in fieldwork conditions in Antarctica in 2025, where it is also expected to meet or exceed the 90% accuracy benchmark.
- **Impact on Research Efficiency:** Field researchers will provide feedback on the time saved using the tool compared to traditional methods. This will be validated by comparing the time spent on species identification before and after tool implementation.
- **Ethical Considerations:** A review of the tool's non-invasive nature will be conducted to ensure it does not cause harm or stress to penguins. WPCO staff will provide feedback regarding the tool's impact on penguin well-being.
- **Field Researcher Safety:** The tool's ability to reduce time spent in harsh field conditions will be assessed through a post-deployment survey. The goal is to validate that the tool reduces the risk of injuries and improves safety. It is estimated that employee injuries will decrease by at least 30% annually compared to 2023.
- **Cost Savings:** The tool's ability to reduce costs associated with fieldwork, monitoring, and research will be assessed. This will be validated by comparing the cost of these activities before and after tool implementation, with an expected annual savings of at least $250,000.

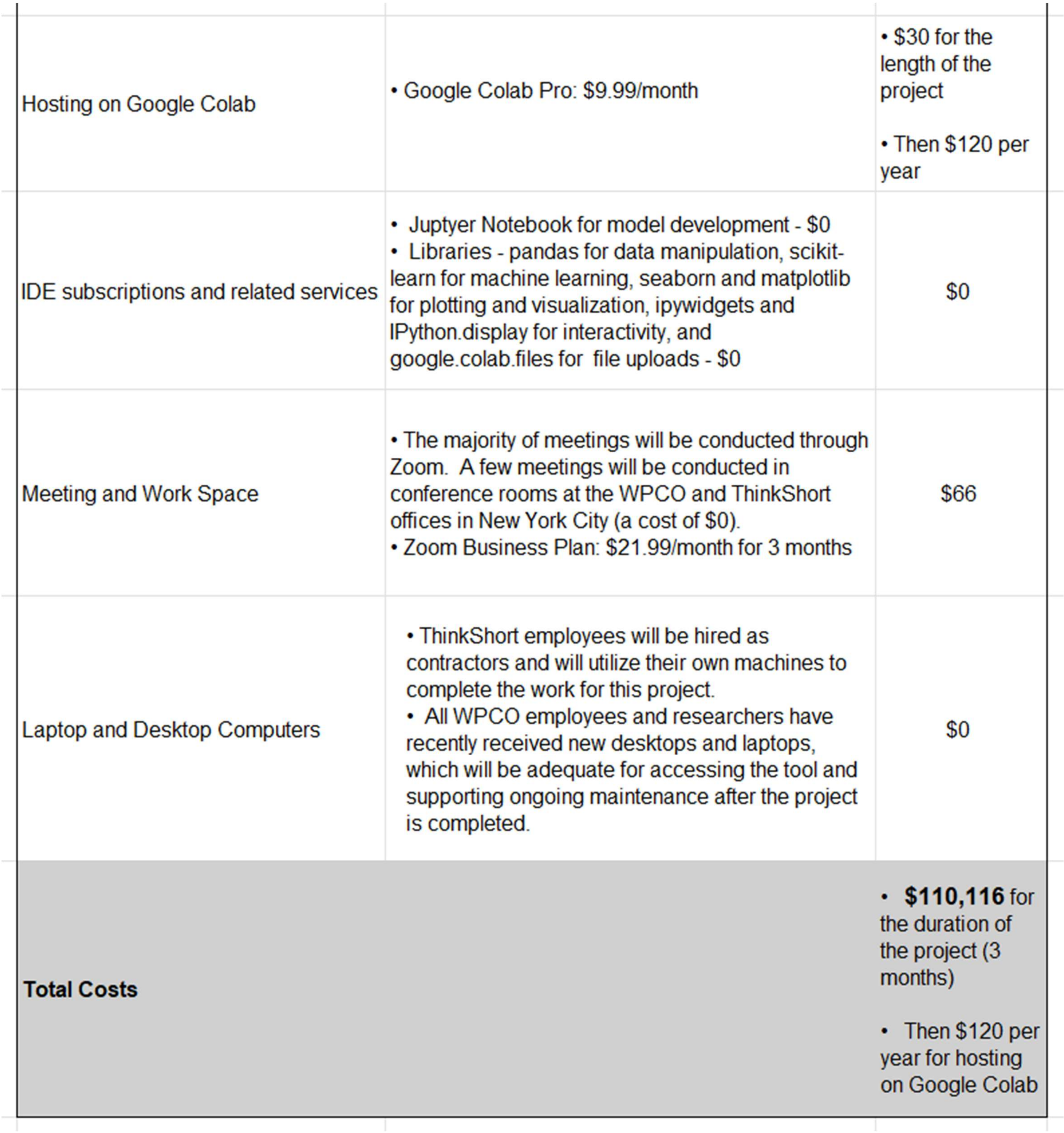
By combining verification and validation at every stage of development, the project will ensure that the proposed solution is reliable, effective, and aligned with the needs of WPCO.

RESOURCES AND COSTS

Below is a list of the resources required for this project, along with their estimated costs. The assumptions used in cost calculations are outlined in the "Description" column.



Resource	Description	Estimated Cost
Project Manager	• Responsible for project planning, team coordination, resource allocation, timeline management, and stakeholder communication. • $130/hr, 228 hours	$29,640
Senior Machine Learning Engineer	• Responsible for leading the development and deployment of the machine learning model. Evaluate model performance. Collaborate with the team to refine and improve the model. Prepare training materials and documentation to ensure the model is well-understood by users. • $150/hr, 228 hours	$34,200
Junior Machine Learning Engineer	• Assist the Senior Machine Learning Engineer in model development and deployment. Help with model testing, validation, and optimization. • $100/hr, 228 hours	$22,800
Junior Software Developer	• Responsible for developing and maintaining the tool's user interface. Assist with integration of the machine learning model into the classification tool. Help deploy the tool. • $100/hr, 228 hours	$22,800
Digital Calipers	• 20 calipers @ $29 each • for measuring penguin bill depth and bill width in the field.	$580
Palmer Penguins Dataset	• This dataset will be used to train the model. • It is free to use and from Kaggle.com.	$0



Hosting on Google Colab	• Google Colab Pro: $9.99/month	• $30 for the length of the project • Then $120 per year
IDE subscriptions and related services	• Juptyer Notebook for model development - $0 • Libraries - pandas for data manipulation, scikit-learn for machine learning, seaborn and matplotlib for plotting and visualization, ipywidgets and IPython.display for interactivity, and google.colab.files for file uploads - $0	$0
Meeting and Work Space	• The majority of meetings will be conducted through Zoom. A few meetings will be conducted in conference rooms at the WPCO and ThinkShort offices in New York City (a cost of $0). • Zoom Business Plan: $21.99/month for 3 months	$66
Laptop and Desktop Computers	• ThinkShort employees will be hired as contractors and will utilize their own machines to complete the work for this project. • All WPCO employees and researchers have recently received new desktops and laptops, which will be adequate for accessing the tool and supporting ongoing maintenance after the project is completed.	$0
Total Costs		• **$110,116** for the duration of the project (3 months) • Then $120 per year for hosting on Google Colab

Part C: Application

My complete application has been submitted on the project portal page. Here is a list of all items submitted:

1. “Penguin Species Classifier.ipynb”

   This is Jupyter source file for the Penguin Species Classifier Tool.

2. “Descriptive Methods.ipynb”

   This is the Jupyter source file for the Descriptive Methods and Visualization Tool.

3. “Penguin_dataset.csv”

   The is the Palmer Penguins Dataset csv file, cleaned and preprocessed for the model.

4. This is the link to the Penguin Species Classifier Tool, a Jupyter Notebook application hosted on Google Colab:

   Penguin Species Classifier Tool (Ctrl + click the link to open)

5. This is the link to the Descriptive Methods and Visualization Tool, another Jupyter Notebook application hosted on Google Colab:

   Descriptive Methods and Visualization Tool (Ctrl + click the link to open)

6. This is the link to the Palmer Penguins Dataset CSV file, which has been cleaned and preprocessed for the model. It is stored in the same Google Colab folder as the other related applications for easy access:

   Palmer Penguins Dataset (Ctrl + click the link to open)

Part D: Post-implementation Report

Solution Summary

The World Penguin Conservation Organization (WPCO) currently identifies penguin species through methods such as measuring body weight, height, flipper length, and bill dimensions (length and depth); recording bill shape, color, plumage patterns, habitat details, vocalizations, and behaviors; and occasionally trapping and releasing penguins for DNA sampling and tagging. While effective, these methods are time-intensive, invasive, and can cause stress or injury to penguins. Additionally, field researchers must endure prolonged exposure to the harsh Antarctic weather, which increases the risk of injury and health issues.

The new machine-learning-based classification tool offers a transformative solution by significantly improving the efficiency of fieldwork, enhancing accuracy in species identification, promoting penguin welfare, and reducing risks to researchers. The Penguin Species Classification Tool generates species predictions within seconds, requiring only two simple inputs: bill length and bill depth measurements. These inputs can be easily collected in the field using inexpensive digital calipers.

Data Summary

i. Data Source

A highly relevant and high-quality dataset, the Palmer Penguins Dataset, was located on Kaggle. You can access it at https://www.kaggle.com/datasets/satyajeetrai/palmer-penguins-dataset-for-eda/data. This dataset, collected between 2007 and 2009, includes observations of adult Adélie, Chinstrap, and Gentoo penguins on islands in the Palmer Archipelago near Palmer Station, Antarctica. The data was gathered and shared by Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program.

ii. Data Processing and Management

This section provides an overview of the data processing and management activities carried out during each phase of the project:

Phase 1: Business Understanding – During Phase 1, the dataset source was selected based on the project objectives defined by the team and stakeholders during the project kickoff meeting. Additionally, a plan for data storage was established to ensure secure and efficient management throughout the project. The dataset and related files will be stored securely in a Google Colab folder, with Jessica Short designated as the owner of the shared file. Access to the folder will be strictly controlled by Jessica Short, and only stakeholders and team members will be granted access to the file to ensure confidentiality and data integrity.

Phase 2: Data Understanding – During Phase 2, the team conducted Exploratory Data Analysis (EDA) to better understand the dataset's structure, assess its quality, and identify potential relationships among variables. The first step involved generating histograms to examine the distribution of the features within the dataset. A scatterplot matrix was then created to explore possible trends and relationships between the variables. Through this analysis, bill length and bill depth emerged as strong candidates for independent variable selection. To further investigate these relationships, a scatterplot of bill length versus bill depth, segmented by species, was generated, as shown in Figure 1. The Palmer Penguins dataset contained measurements for various features, but bill length and bill depth were particularly notable due to their clear relationship with different penguin species.

For additional visualizations created during Phase 2, please refer to the 'Visualizations' section below or consult 'Part 2 – Descriptive Methods and Visualizations' in the Quick Start User Guide

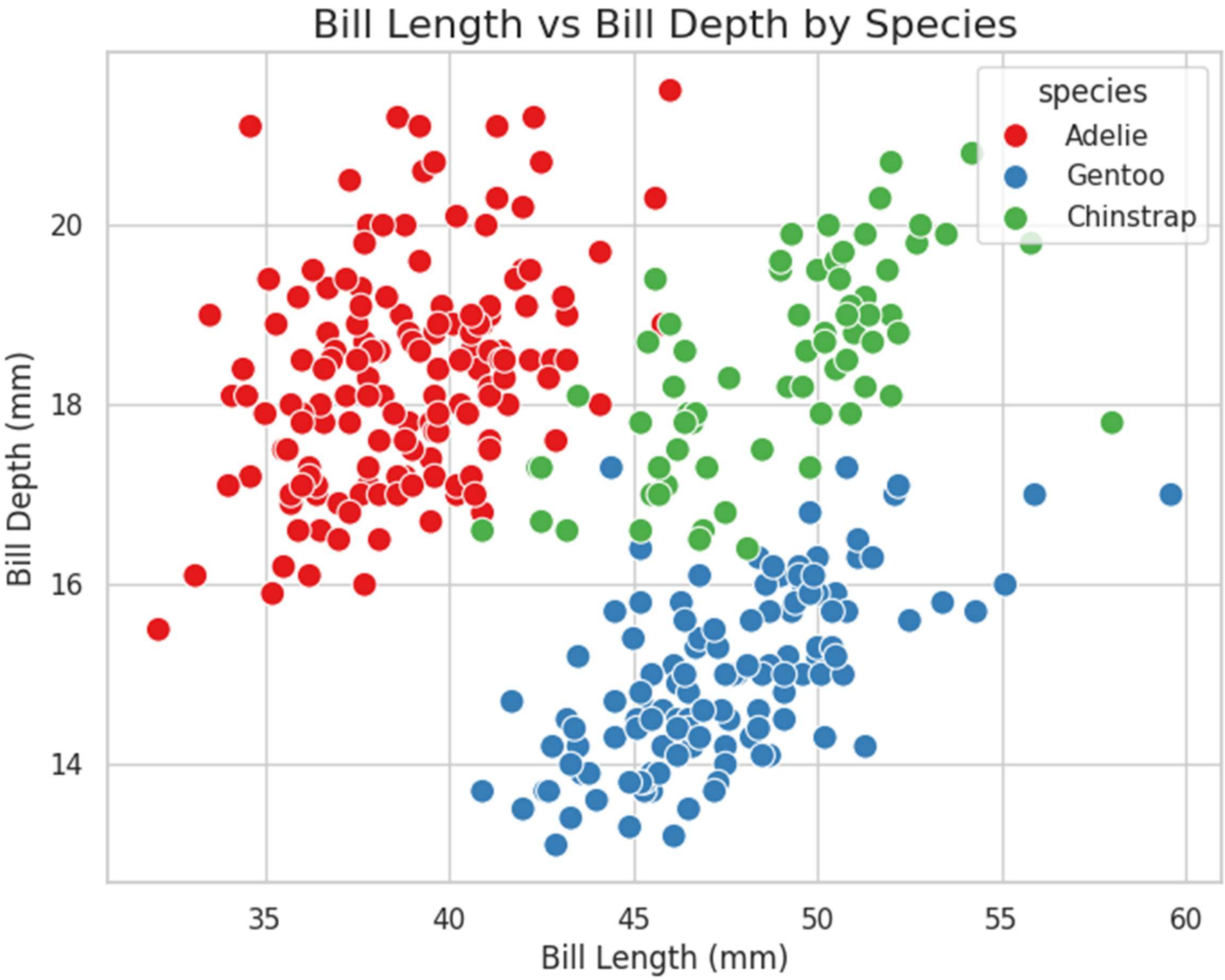


Figure 1. Bill Length vs. Bill Depth by Species

Phase 3: Data Preparation – During this phase, the data was cleaned and preprocessed for modeling. Rows corresponding to IDs 3 and 271 were removed due to multiple missing values in key attributes, including bill length, bill depth, flipper length, body mass, and gender. For rows with missing values only in the gender attribute (IDs 8–11, 47, 178, 218, 256, and 268), the mode value ('male') was imputed. Since gender was not selected as a feature for the model, these rows were retained to preserve the dataset's size.

An outlier detection assessment was conducted in Jupyter Notebook using boxplots, Z-scores, and the IQR method. The analysis revealed no significant outliers in the dataset. The results from this assessment are displayed in Figure 2 below.

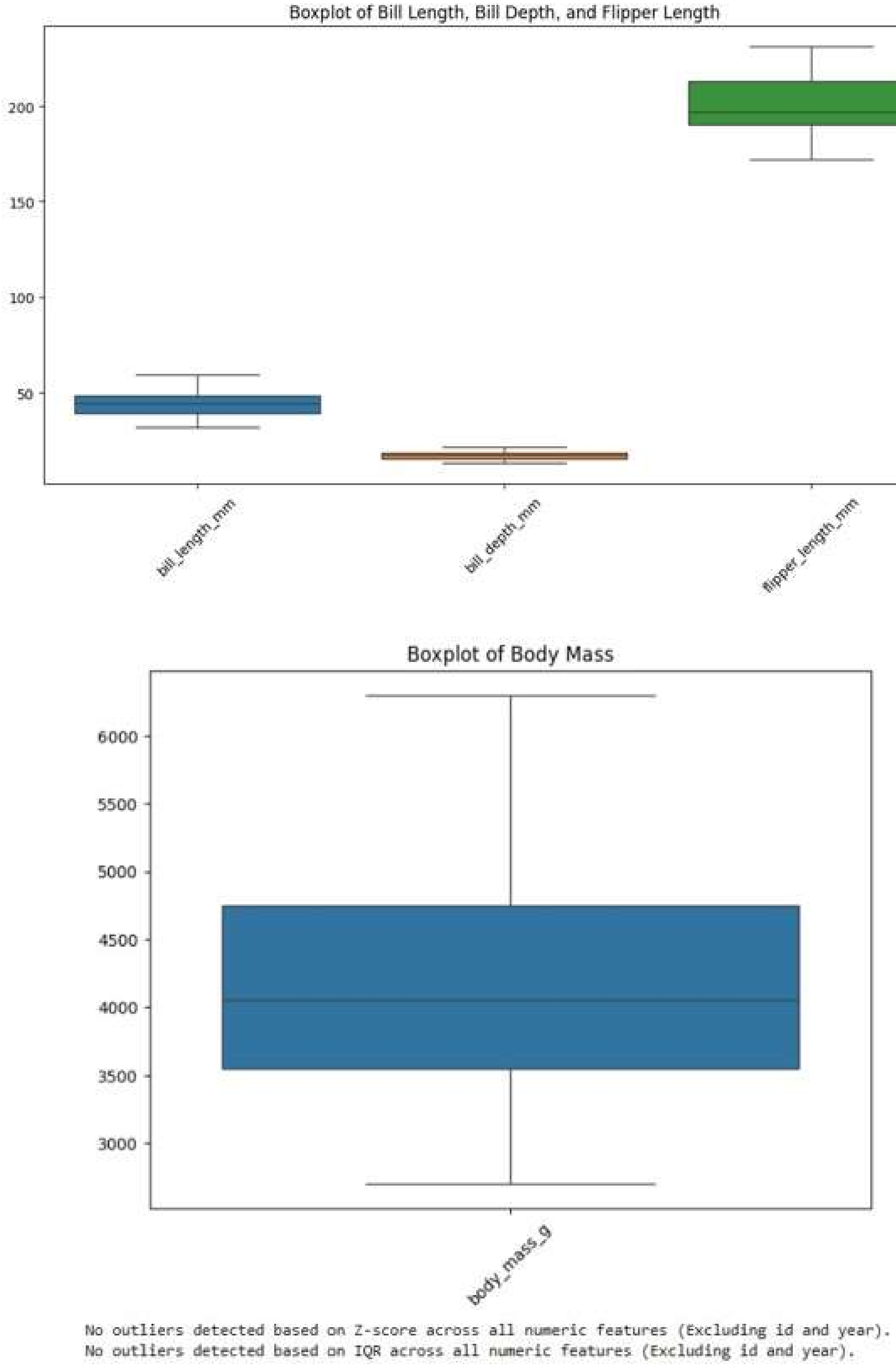
Next, key features for the model were selected, specifically bill length and bill depth measurements. The target variable for classification was the penguin species.

Finally, the data was split into training and testing subsets, with 70% allocated for training and the remaining 30% for testing. This split was achieved using the train_test_split function from Scikit-learn, with a random state of 42 to ensure reproducibility.

Phase 4: Modeling – During the modeling phase, a Logistic Regression model was selected for the classification task, as it is well-suited for binary and multiclass classification problems. The training subset was used to train the logistic regression model. The model was initialized using the LogisticRegression class from Scikit-learn, with the max_iter parameter set to 1000 to ensure the model converged. The model was then fitted to the training data using the fit method, which adjusted the model's internal parameters to minimize the error between predicted and actual species.

Phase 5: Evaluation – In this phase, the model's performance was evaluated to ensure it met the desired accuracy standards. After training the model using the training subset, it was tested on the testing subset, and the accuracy was calculated using the accuracy_score function from Scikit-learn. This accuracy score provided a general understanding of how well the model could predict penguin species based on the two selected features: bill length and bill depth.

Beyond just accuracy, a confusion matrix was generated to provide a more detailed view of the model's classification performance. The matrix displayed the number of correct and incorrect predictions for each species, helping to identify strengths and weaknesses in the model's predictions. The confusion matrix was visualized using a heatmap from Seaborn, offering a clear, easily interpretable representation of the results. This heatmap also highlighted any species that were misclassified more often, helping to pinpoint areas for potential model improvements. Please refer to the "Validation" section below to view the accuracy and confusion matrix results.



Boxplot of Bill Length, Bill Depth, and Flipper Length

Boxplot of Body Mass

```
No outliers detected based on Z-score across all numeric features (Excluding id and year).
No outliers detected based on IQR across all numeric features (Excluding id and year).
```

Figure 2. Results of the Outlier Detection Assessment

Phase 6: Deployment – In the deployment phase, the final model was integrated into the production environment, which is a Jupyter Notebook hosted on Google Colab. A user-friendly interface was developed to allow input of bill length and depth measurements for species predictions.

The model was tested under simulated field conditions to ensure its reliability and ease of use. In collaboration with the Central Park Zoo, WPCO researchers tested the model with 35 of the zoo's Gentoo and Chinstrap penguins. The zoo does not currently house Adélie penguins (Penguins, n.d.). The model achieved 98% accuracy in species identification with this test group.

Additionally, a long-term plan for data storage was established to securely store new field measurement data and model prediction results. These data will be stored in a Google Colab folder, with Jessica Short designated as the owner of the shared file. Access to the folder will be strictly controlled by Jessica Short, ensuring that only stakeholders and team members can access the file, thus maintaining confidentiality and data integrity.

Machine Learning

i. Method

In this project, a logistic regression model was used for the classification task. Logistic regression is a statistical method used for binary or multi-class classification problems, where the goal is to predict a categorical outcome based on one or more input features. It works by estimating the probability of each class and then making a prediction based on the highest probability. In this case, the model was trained to classify penguin species based on the features of bill length and bill depth.

ii. Method Development

In developing the penguin classification model, the first step was to preprocess the data and prepare it for machine learning. Using the Palmer Penguins dataset, relevant features were selected for the model. Specifically, the target variable was the penguin species (y), while the input features were bill length and bill depth (X). These features were chosen based on exploratory data analysis (EDA), where relationships between bill measurements and species were identified as significant. The dataset was split into a training set and a test set using a 70/30 split, ensuring that the model would have enough data to learn from while still being validated on unseen data. The train_test_split function from sklearn.model_selection was used to perform this split, with a random state of 42 to ensure reproducibility.

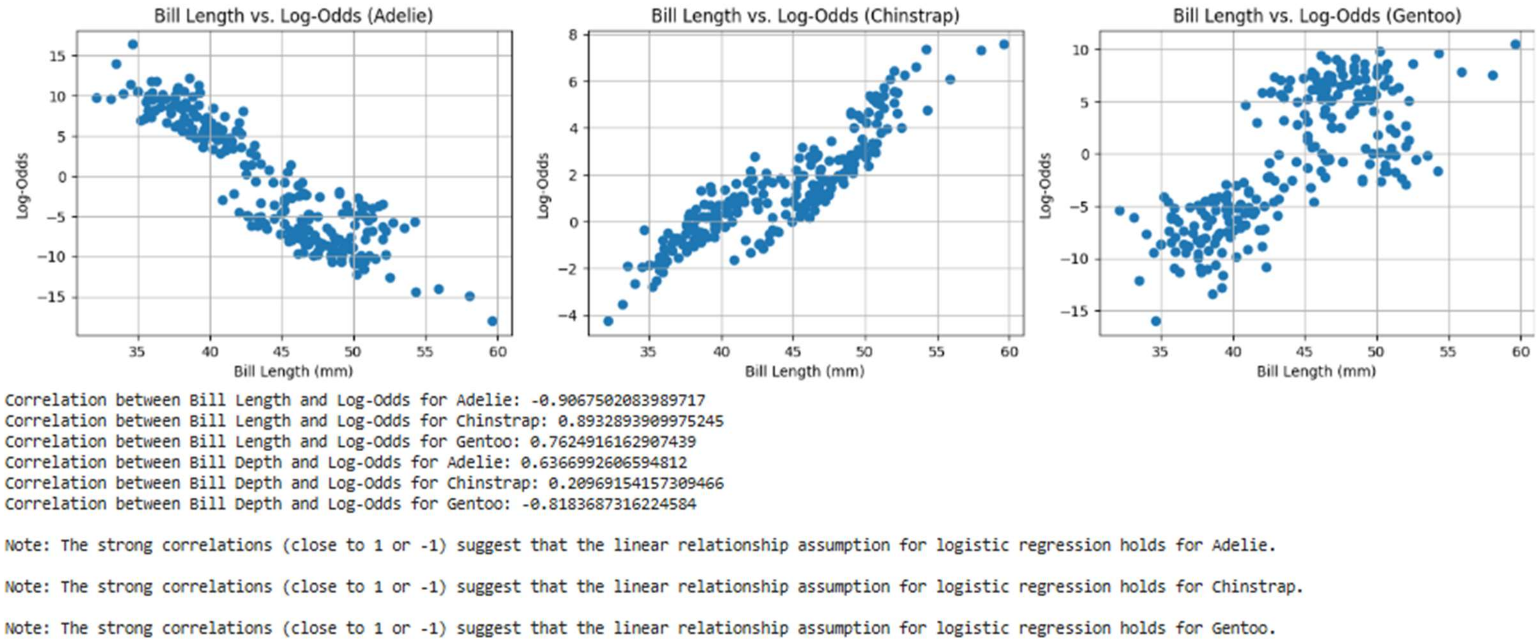
Once the data was prepared, a Logistic Regression model was chosen for classification. Logistic regression is a simple and interpretable machine learning model that works well for binary and multi-class classification tasks, making it suitable for this penguin species classification problem. The model was initialized and then trained using the training dataset, where it learned the relationship between the input features (bill length and bill depth) and the target variable (species). After the model was trained, its performance was evaluated on the test set. Predictions were made on the test data, and the model's accuracy was assessed using the accuracy_score function from the sklearn.metrics module. Additionally, a confusion matrix was generated to further assess the model's performance, visualizing how well the predictions matched the actual species labels. This validation step ensured that the model could generalize well to new, unseen data and provided insights into its classification performance.

iii. Method Justification

A Logistic Regression model was chosen for classification because it is well-suited for predicting categorical outcomes, particularly when the dataset is not too large. Additionally, logistic regression is easy to implement, computationally inexpensive, and relatively straightforward to understand (Shaw, 2020). These factors make it an ideal choice for this project.

During model development, the team recognized that logistic regression is effective when there is a linear relationship between the independent variables and the log-odds of the outcome. However, if the data exhibits non-linear patterns, the model may not perform well (Shaw, 2020). To ensure the assumptions were met, the team conducted a study to examine whether a linear relationship exists between the independent variables and the log-odds of the outcome.

In this study, log-odds were calculated for each class using a Jupyter notebook. Scatter plots were then created for each class to visualize the relationship between the features (bill_length_mm and bill_depth_mm) and the log-odds for that class. The correlation between the features and log-odds was calculated using np.corrcoef to assess the linearity of the relationships. The results of this analysis are shown below in Figure 3:



```
Correlation between Bill Length and Log-Odds for Adelie: -0.9067502083989717
Correlation between Bill Length and Log-Odds for Chinstrap: 0.8932893909975245
Correlation between Bill Length and Log-Odds for Gentoo: 0.7624916162907439
Correlation between Bill Depth and Log-Odds for Adelie: 0.6366992606594812
Correlation between Bill Depth and Log-Odds for Chinstrap: 0.20969154157309466
Correlation between Bill Depth and Log-Odds for Gentoo: -0.8183687316224584

Note: The strong correlations (close to 1 or -1) suggest that the linear relationship assumption for logistic regression holds for Adelie.

Note: The strong correlations (close to 1 or -1) suggest that the linear relationship assumption for logistic regression holds for Chinstrap.

Note: The strong correlations (close to 1 or -1) suggest that the linear relationship assumption for logistic regression holds for Gentoo.
```

Figure 3. Relationship Between Bill Dimensions and Log-Odds for Adelie, Chinstrap, and Gentoo Penguins

These results demonstrated that:

- Adelie and Chinstrap species show strong correlations (close to -1 or 1) with bill length, indicating that the linear relationship assumption holds well for these species. Bill length is therefore a reliable predictor for classifying these species.
- Gentoo shows a moderate correlation with bill length and a strong negative correlation with bill depth. This suggests that while the linearity assumption holds, bill depth has a stronger influence on classification than bill length for this species.
- Chinstrap shows a weaker correlation with bill depth, suggesting that the linear relationship assumption is less reliable for this feature.

Overall, the correlation analysis confirms that the linear relationship assumption holds for Adelie, Chinstrap, and Gentoo. Therefore, logistic regression remains a suitable choice for this classification model.

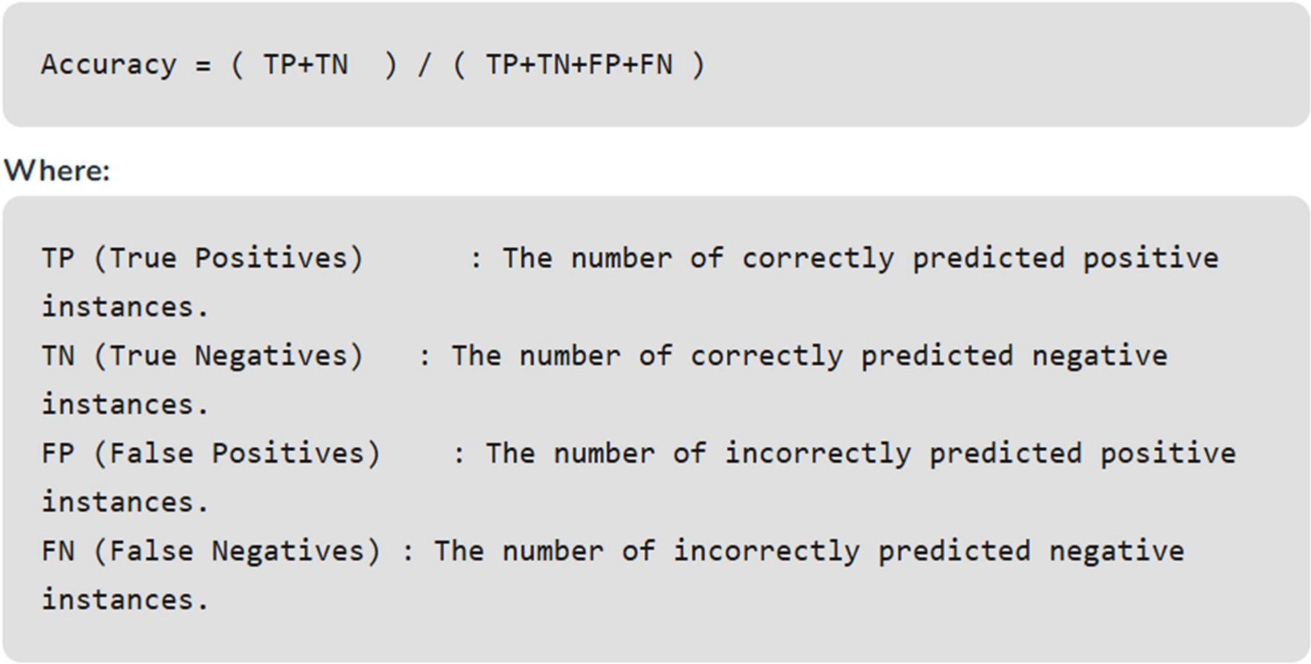
Validation

i. Validation Method

The train-test split method was used to evaluate the model's performance. The dataset was divided into two subsets: one for training the model and one for testing its performance. Specifically, 70% of the data was allocated to the training set, and 30% was reserved for testing. This was achieved using the train_test_split function from Scikit-learn, with a random seed of 42 to ensure reproducibility.

Once the dataset was split, a Logistic Regression model was trained using the training subset. The model was then used to make predictions on the testing subset, and the results were evaluated using two key metrics:

1. **Accuracy Score**: This metric the fraction of correct predictions (where the predicted label matches the true label) out of all predictions, providing an overall performance indicator for the model. The accuracy of the model is calculated using the accuracy_score function from Scikit-learn. The formula for accuracy is:



```
Accuracy = ( TP+TN  ) / ( TP+TN+FP+FN )
```

Where:

```
TP (True Positives)      : The number of correctly predicted positive
instances.
TN (True Negatives)   : The number of correctly predicted negative
instances.
FP (False Positives)    : The number of incorrectly predicted positive
instances.
FN (False Negatives) : The number of incorrectly predicted negative
instances.
```

(GeeksforGeeks, n.d.)

2. **Confusion Matrix**: This matrix visualizes the performance of the model, showing the number of correct and incorrect predictions across the different species categories. It was plotted using Seaborn's heatmap to facilitate a clearer interpretation.

ii. Validation Results

The model's performance was evaluated after making predictions on the test subset:

- **Accuracy**: The model achieved an impressive 97% accuracy in classifying penguin species based on bill length and depth measurements.
- **Confusion Matrix**: The confusion matrix (shown as a heatmap) displayed how well the model predicted the penguin species across all categories. The matrix helped identify any species that were misclassified more often, offering insights into areas for model improvement. The confusion matrix is shown in Figure 3 below:

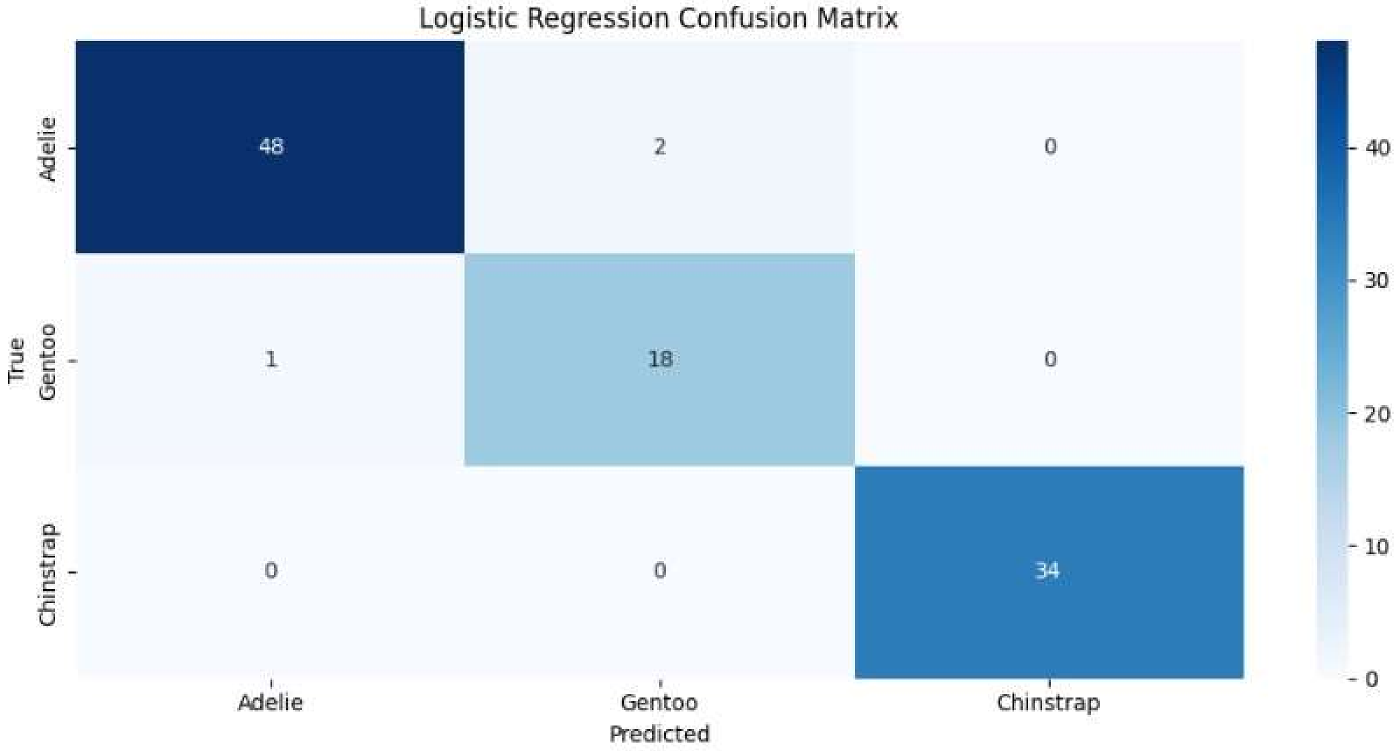


Figure 4. Confusion Matrix for the Penguin Species Classifier Model

The confusion matrix showed that the model performed best in identifying Chinstrap species, with 0 misclassification. Among the Adélie penguins, only 2 were misclassified and among the Gentoo penguins, only 1 was misclassified.

Overall, the validation method showed that the logistic regression model performed well in predicting penguin species with high accuracy, making it a reliable tool for species classification in this context.

Visualizations

During the Data Understanding phase of this project (phase 2), the Palmer Penguins dataset was explored through visualizations using the Descriptive Methods and Visualization Tool. This tool is provided to WPCO as a deliverable for this project. For detailed instructions on how to use this tool, please refer to the Quick Start User Guide (see "Part 2 – Descriptive Methods and Visualizations").

Here is the link to the tool:

Descriptive Methods and Visualization Tool

The visualizations generated by Descriptive Methods and Visualization Tool are shown below in Figures 5-7.

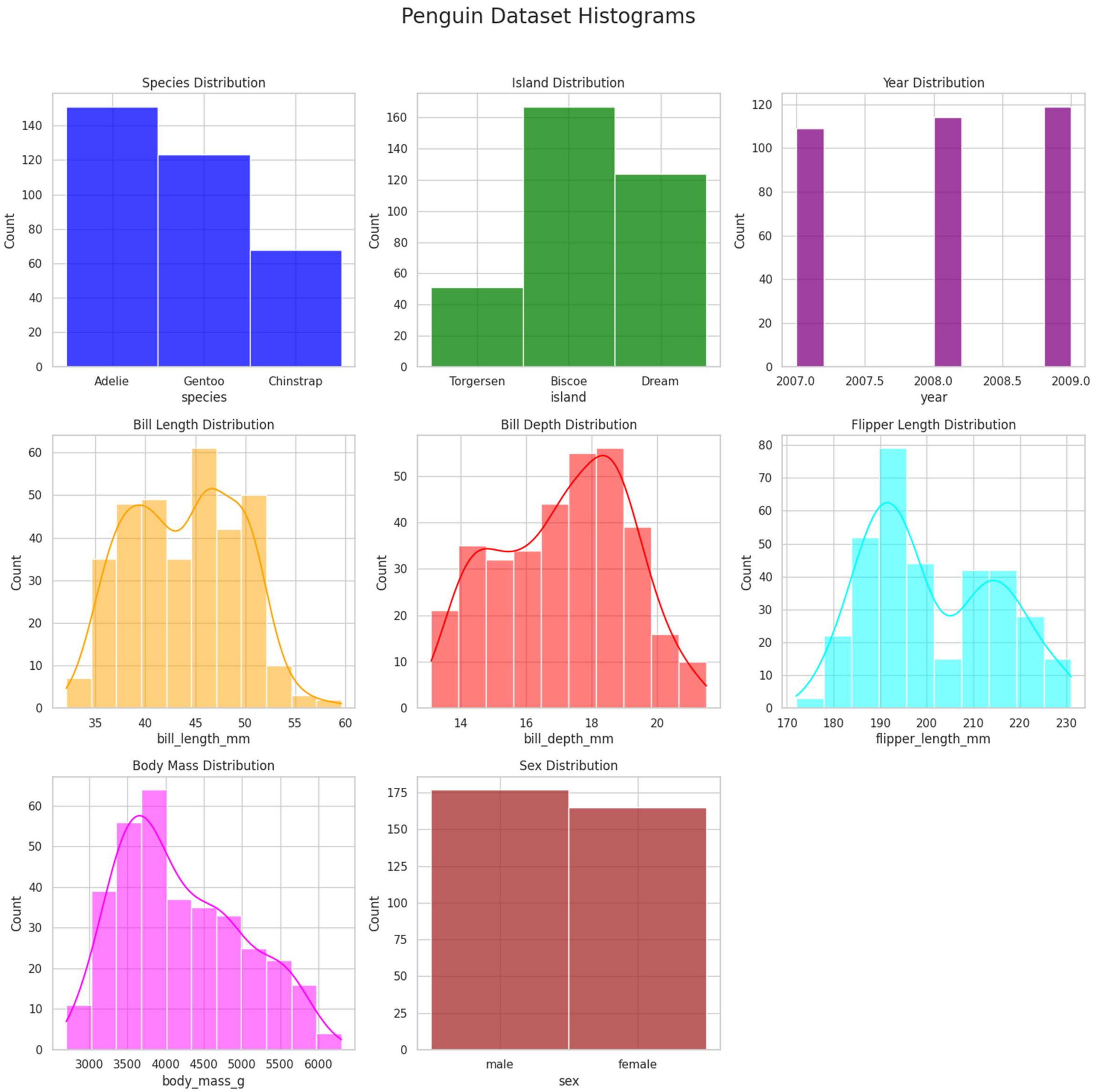


Figure 5. Histograms

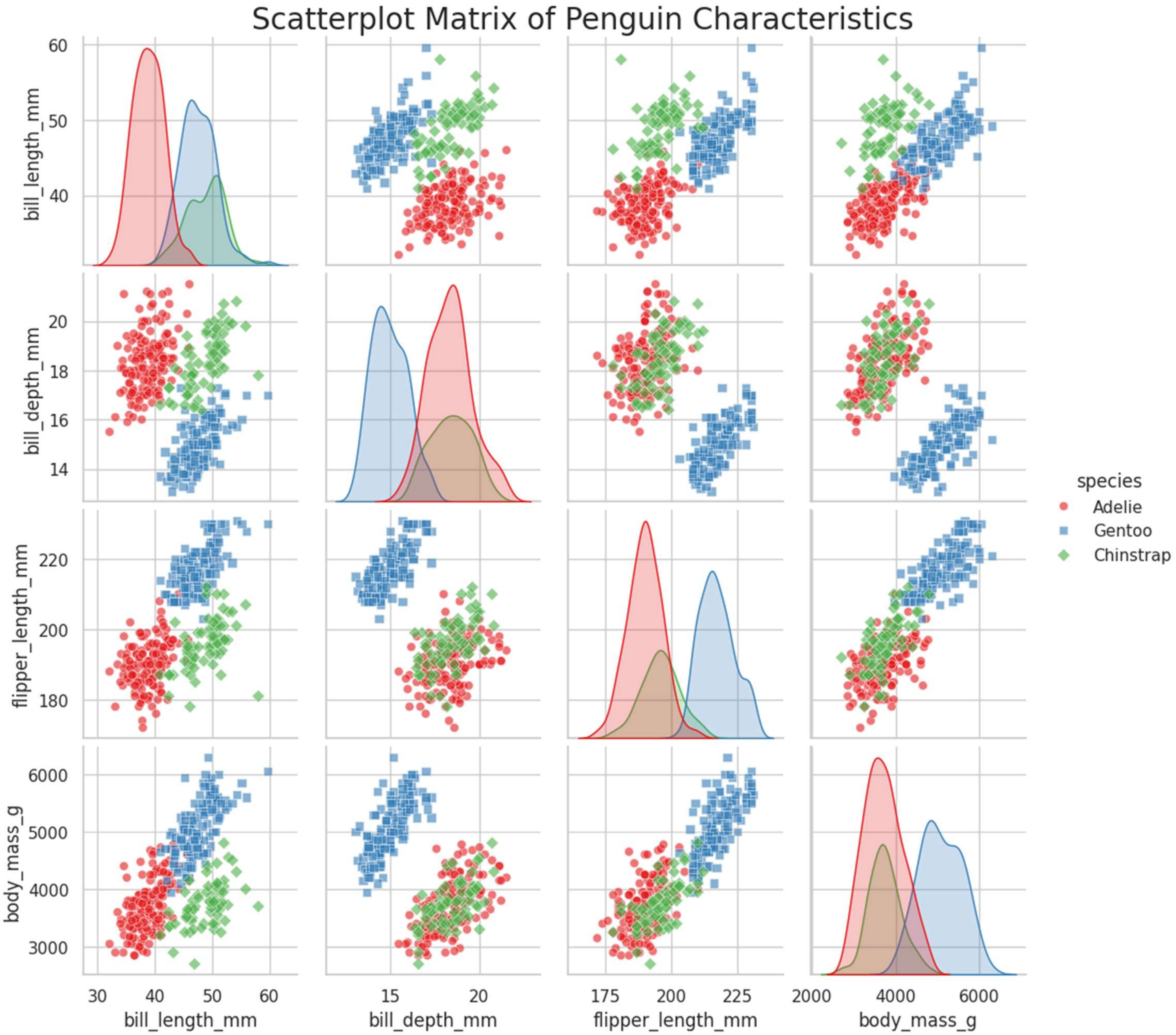


Figure 6. Scatterplot Matrix

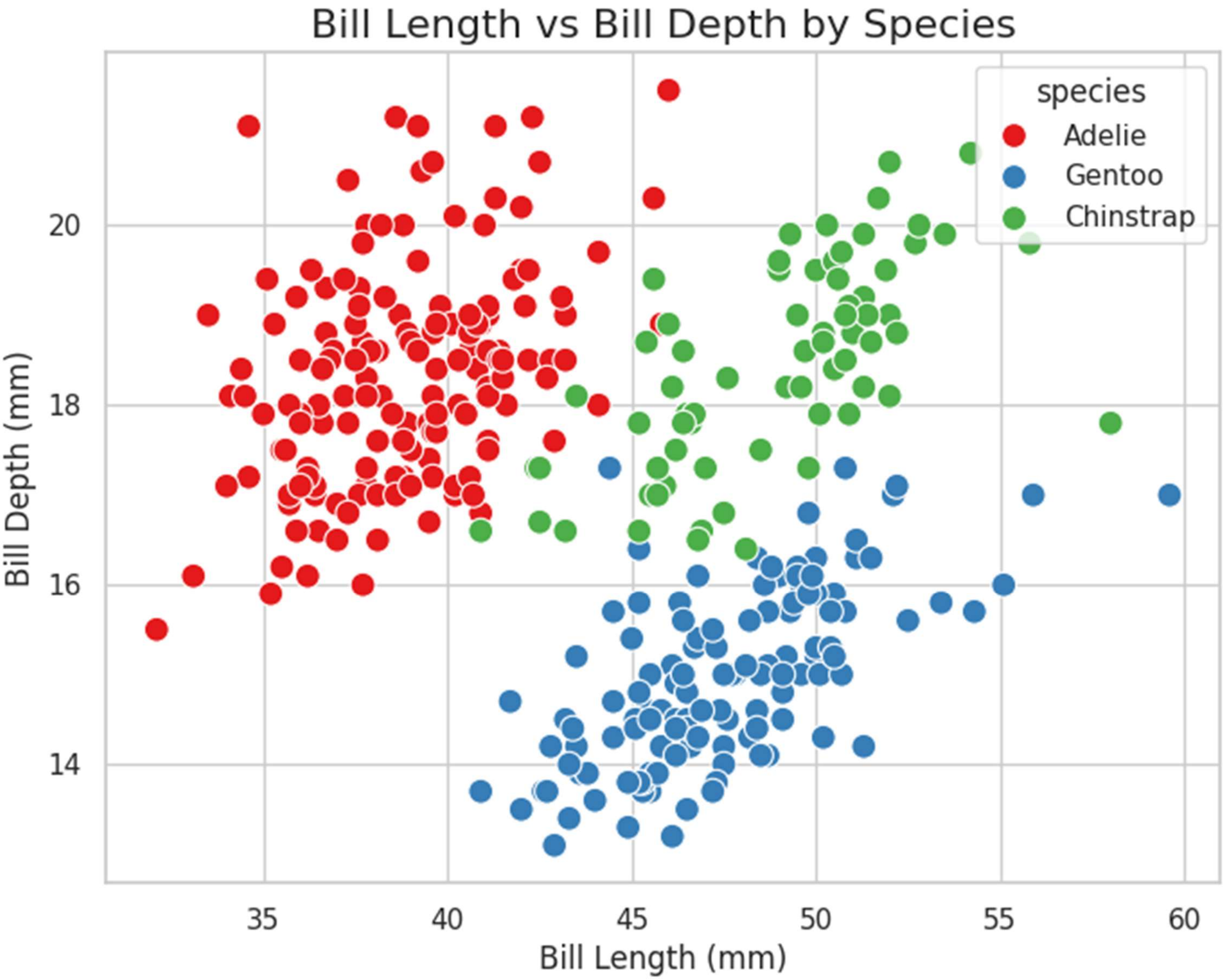


Figure 7. Scatterplot of Bill Length vs Bill Depth by Species

Quick Start User Guide

To open and use the Penguin Species Classification Tool and the Descriptive Methods and Visualizations Tool, you will need the following:

1. Windows 10 or higher.
2. A Google account - Since the tool is hosted on Google Drive, a Google account is required. If you don't have one, you can create a free Google account here:

   Create a Google Account

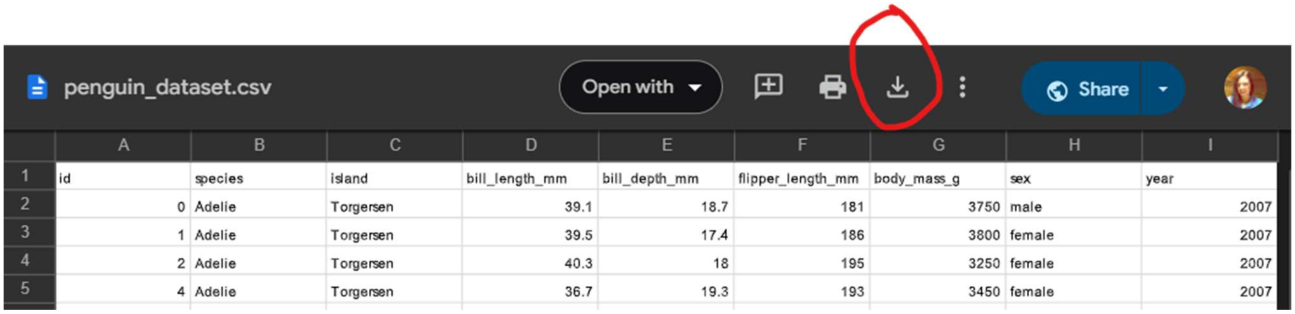
3. A stable internet connection.
4. A modern web browser - Google Chrome or Safari is recommended.

Part 1 – Penguin Species Classification Tool

Follow these instructions to use the Penguin Species Classification Tool:

1. **Download the file "penguin_dataset.csv"** to your local machine. Ctrl + click the link below and then click ⤓ to save the file to your downloads folder. Make sure to move this file to a location where you can easily find it later. **And make sure that you name it "penguin_dataset.csv" (<<Use this name exactly!)**

   penguin_dataset.csv

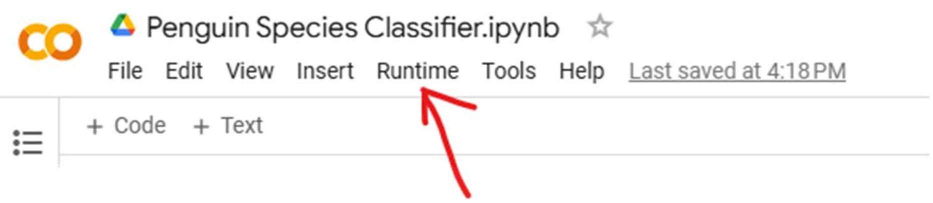


	A	B	C	D	E	F	G	H	I
1	id	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
2	0	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
3	1	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
4	2	Adelie	Torgersen	40.3	18	195	3250	female	2007
5	4	Adelie	Torgersen	36.7	19.3	193	3450	female	2007

2. **Open the Jupyter Notebook in Google Colab** by Ctrl + clicking the link below:

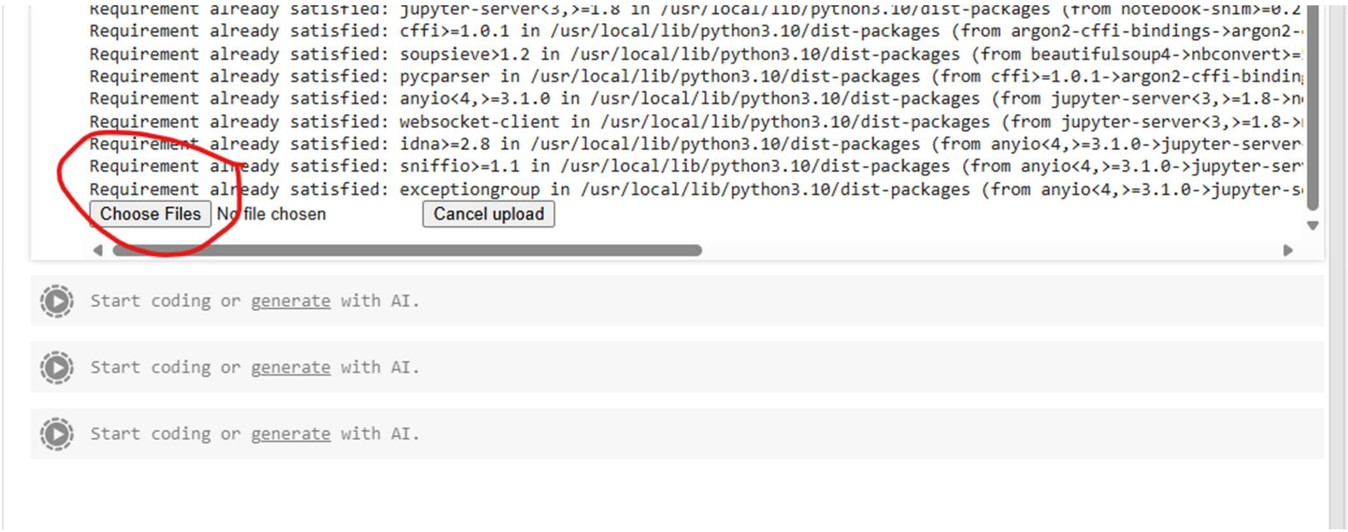
   [Penguin Species Classifier Tool]

3. From the Menu Bar at the top of the page, click on **“Runtime”** > **“Run All”**.



4. After a few seconds, scroll to the bottom of the output cell. A button labeled **“Choose Files”** will appear. Click this button, select the “penguin_dataset.csv” file from the location where you saved it, and click **“Open”**. This will upload the file to Colab’s temporary storage.

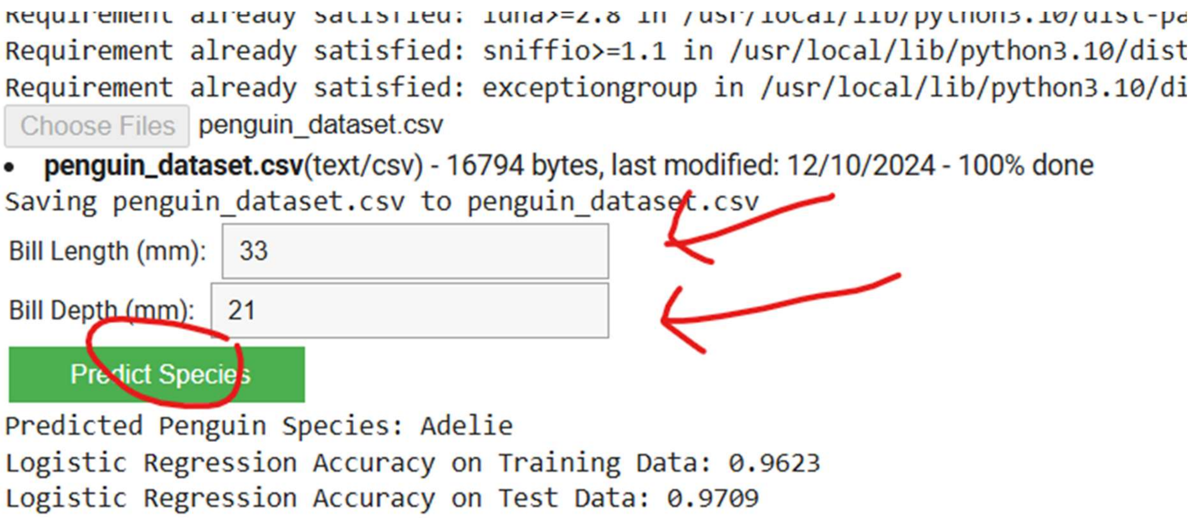
Note: Make sure that you named the csv file “penguin_dataset.csv” in Step 1 or it won’t work correctly.



```
Requirement already satisfied: jupyter-server<3,>=1.8 in /usr/local/lib/python3.10/dist-packages (from notebook-shim>=0.2
Requirement already satisfied: cffi>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from argon2-cffi-bindings->argon2-
Requirement already satisfied: soupsieve>1.2 in /usr/local/lib/python3.10/dist-packages (from beautifulsoup4->nbconvert>=
Requirement already satisfied: pycparser in /usr/local/lib/python3.10/dist-packages (from cffi>=1.0.1->argon2-cffi-bindin
Requirement already satisfied: anyio<4,>=3.1.0 in /usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->n
Requirement already satisfied: websocket-client in /usr/local/lib/python3.10/dist-packages (from jupyter-server<3,>=1.8->
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.10/dist-packages (from anyio<4,>=3.1.0->jupyter-server
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.10/dist-packages (from anyio<4,>=3.1.0->jupyter-ser
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from anyio<4,>=3.1.0->jupyter-s
```

Choose Files
No file chosen
Cancel upload
Start coding or generate with AI.
Start coding or generate with AI.
Start coding or generate with AI.

5. Once the code finishes running, scroll down to the bottom of the output cell. You will see a **green button labeled "Predict Species"**. Above this button, there are text boxes where you can input values for bill length and bill depth to get a prediction for the penguin species.



```
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.10/dist-pa
Requirement already satisfied: sniffio>=1.1 in /usr/local/lib/python3.10/dist
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/di
```

Choose Files penguin_dataset.csv

- **penguin_dataset.csv**(text/csv) - 16794 bytes, last modified: 12/10/2024 - 100% done

```
Saving penguin_dataset.csv to penguin_dataset.csv
```

Bill Length (mm): 33

Bill Depth (mm): 21

```
Predicted Penguin Species: Adelie
Logistic Regression Accuracy on Training Data: 0.9623
Logistic Regression Accuracy on Test Data: 0.9709
```

6. **Input the bill length and bill depth** (in millimeters) into the respective boxes. **Enter a bill length between 32.0 mm and 60.0 mm, and a bill depth between 13.0 mm and 22.0 mm. Round your inputs to the nearest 0.1 mm.** Then, Click the "Predict Species" button.

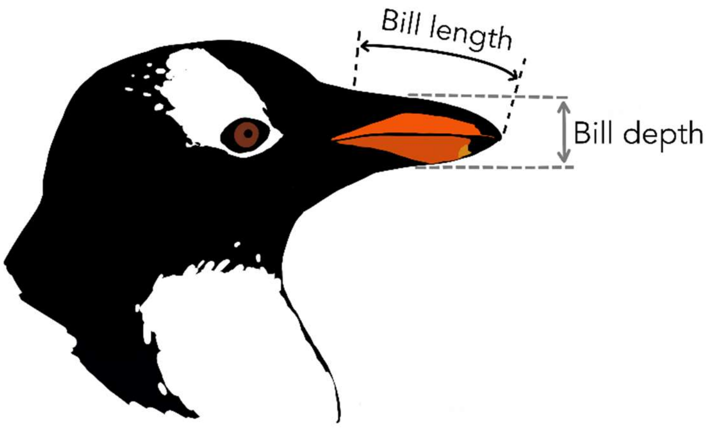
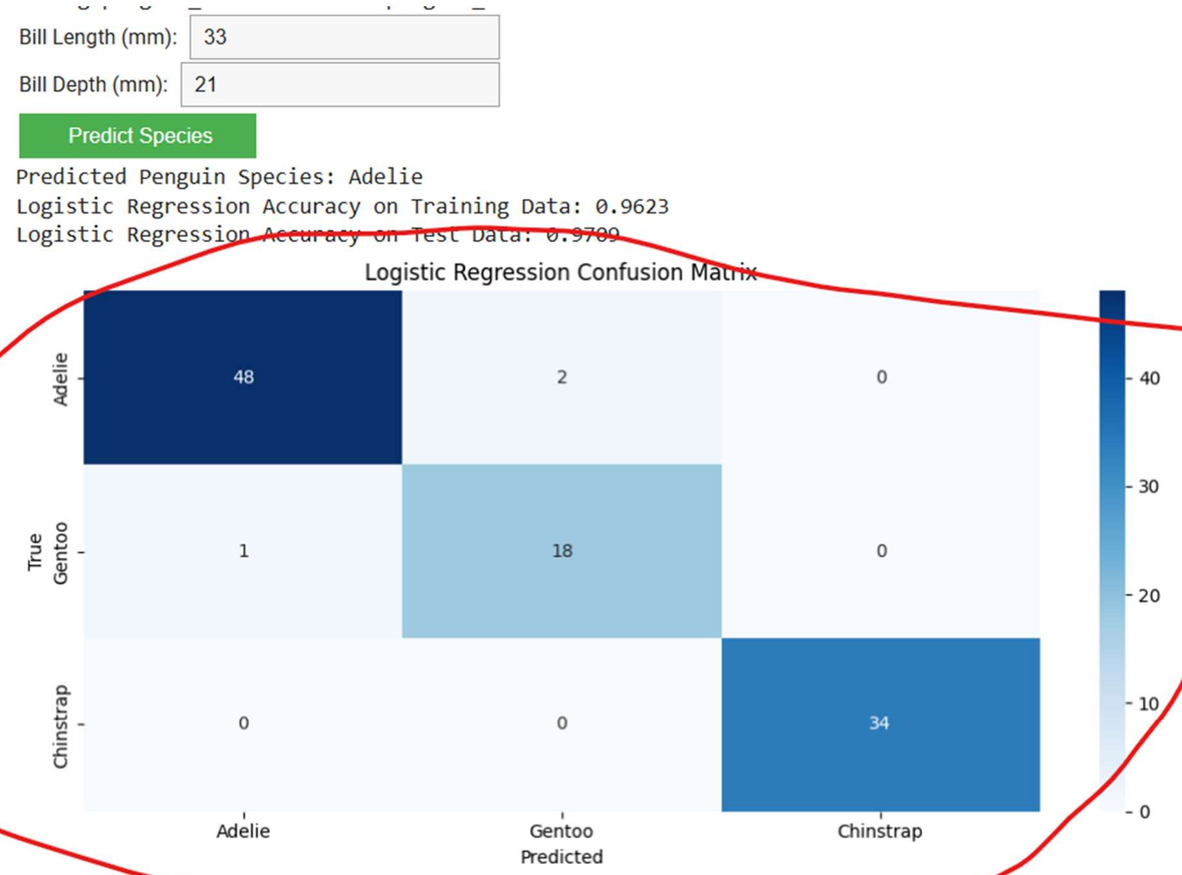


Figure 1: Measuring bill length and bill depth. (Horst, n.d.)

7. Your **prediction result** will appear under the “Predict Species” button, labeled as “Predicted Penguin Species:”.

8. Below the species prediction tool, you can view the **accuracy analysis** for the model, including the accuracy percentage and the confusion matrix.



Bill Length (mm): 33

Bill Depth (mm): 21

```
Predicted Penguin Species: Adelie
Logistic Regression Accuracy on Training Data: 0.9623
Logistic Regression Accuracy on Test Data: 0.9709
```

Logistic Regression Confusion Matrix

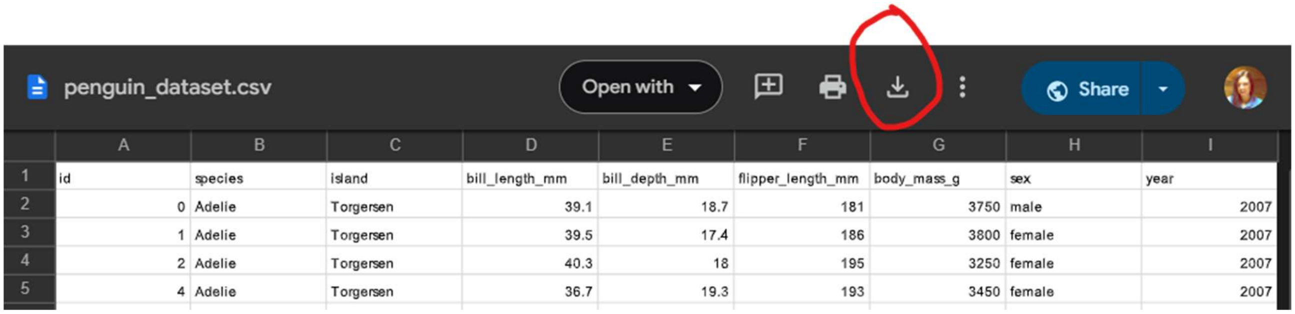
True \ Predicted	Adelie	Gentoo	Chinstrap
Adelie	48	2	0
Gentoo	1	18	0
Chinstrap	0	0	34

Part 2 – Descriptive Methods and Visualizations

Follow these instructions to explore the dataset via descriptive methods and visualizations:

1. Download the file “penguin_dataset.csv” to your local machine. Ctrl + click the link below and then click ⬇ to save the file to your downloads folder. Make sure to move this file to a location where you can easily find it later. **And make sure that you name it “penguin_dataset.csv” (<<Use this name exactly!)**

[penguin_dataset.csv]

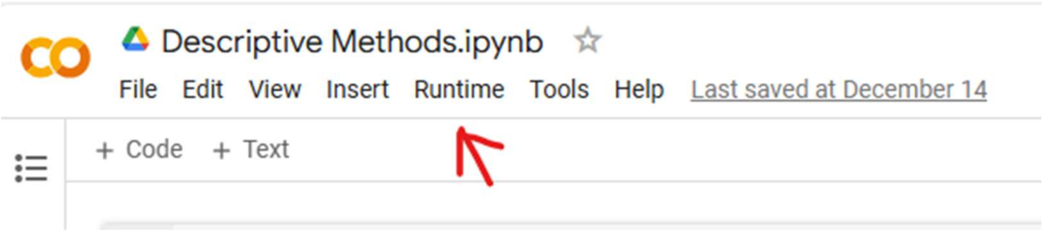


	A	B	C	D	E	F	G	H	I
1	id	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
2	0	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
3	1	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
4	2	Adelie	Torgersen	40.3	18	195	3250	female	2007
5	4	Adelie	Torgersen	36.7	19.3	193	3450	female	2007

2. Ctrl + click the link below to open the Jupyter Notebook in Google Colab:

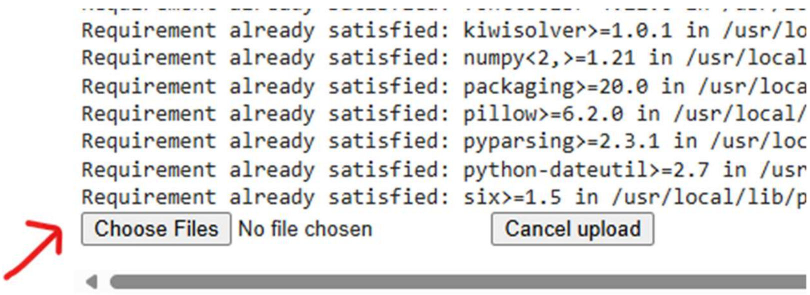
[Descriptive Methods and Visualization Tool]

3. From the Menu Bar at the top of the page, click on “Runtime” > “Run All”.



4. Scroll down to the bottom of the output cell. After a few seconds, a button will appear that says “Choose Files”. Click this button and select “penguin_dataset.csv” in the location that you saved it to your local machine. Then select “Open”. This will upload the file to Colab's temporary storage.

Note: Make sure that you named the csv file “penguin_dataset.csv” in Step 1 or it won’t work correctly.



```
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/lo
Requirement already satisfied: numpy<2,>=1.21 in /usr/local
Requirement already satisfied: packaging>=20.0 in /usr/loca
Requirement already satisfied: pillow>=6.2.0 in /usr/local/
Requirement already satisfied: pyparsing>=2.3.1 in /usr/loc
Requirement already satisfied: python-dateutil>=2.7 in /usr
Requirement already satisfied: six>=1.5 in /usr/local/lib/p
```

5. Once the code is finished running, scroll down to the bottom of the output cell. Here, you can view the generated histograms and scatterplots.

Part E. References

1. Carroll, G., Turner, E., Dann, P., & Harcourt, R. (2016). Prior exposure to capture heightens the corticosterone and behavioural responses of little penguins (Eudyptula minor) to acute stress. *Conservation Physiology, 4*(1), cov061. https://doi.org/10.1093/conphys/cov061

2. Central Park Zoo. (n.d.). *Penguins*. https://www.centralpark.com/things-to-do/central-park-zoo/penguins/

3. GeeksforGeeks. (n.d.). *Classification metrics using sklearn*. GeeksforGeeks. https://www.geeksforgeeks.org/sklearn-classification-metrics/

4. Hotz, N. (2024, April 28). What is CRISP DM? *Data Science PM*. https://www.datascience-pm.com/crisp-dm-2/

5. Horst, A. (n.d.). *Penguin bills* [Photograph]. *Palmer Penguins*. https://allisonhorst.github.io/palmerpenguins/articles/art.html

6. IBM. (n.d.). *Overfitting vs. underfitting*. IBM. Retrieved December 20, 2024, from https://www.ibm.com/think/topics/overfitting-vs-underfitting

7. Shaw, A. (2020, July 27). *Logistic regression and its role in classification problems*. Medium. https://medium.com/@abhishekshaw020/logistic-regression-and-its-role-in-classification-problems-504ff348bb41