February 17th, 2024

Penguin Species Prediction System

Task 2 parts A, B, C and D

[Part A: Letter of Transmittal 2](#_Toc159107196)

[Part B: Project Proposal Plan 3](#_Toc159107197)

[Project Summary 3](#_Toc159107198)

[Data Summary 3](#_Toc159107199)

[Implementation 3](#_Toc159107200)

[Timeline 4](#_Toc159107201)

[Evaluation Plan 4](#_Toc159107202)

[Resources and Costs 5](#_Toc159107203)

[Part C: Application 7](#_Toc159107204)

[Part D: Post-implementation Report 8](#_Toc159107205)

[Solution Summary 8](#_Toc159107206)

[Data Summary 8](#_Toc159107207)

[Machine Learning 10](#_Toc159107208)

[Validation 11](#_Toc159107209)

[Visualizations 13](#_Toc159107210)

[User Guide 16](#_Toc159107211)

[Guide for Google Colab with a Google Account 16](#_Toc159107212)

[Guide for local python IDE 17](#_Toc159107213)

# Part A: Letter of Transmittal

February 17th, 2024

Anrosha Lewbarggs, CEO

Penguin Research Group

2509 Grasselli Street

Portsmouth, NH 03801

Dear Mr. Lewbarggs,

I am contacting you to present the proposal for implementing the Penguin Species Classification System at PRG. The Penguin Species Classification System is a data product developed by our team that addresses a crucial need within the Penguin Research Group by revolutionizing how penguin species are identified. As the Penguin Research Group strives to advance in research and conservation efforts, it can be a challenge to manually identify species of penguins. Manual identification can be time-consuming and prone to human error. The proposed Penguin Classification System solution can fully automate this process using machine learning algorithms. Our Penguin Species Classification System will leverage these machine learning algorithms to provide the Penguin Research Group with a more reliable and efficient method for identifying the species of penguins.

This proposed project consists of data acquisition, preprocessing, developing the model, integrating the system, creating desired documentation, and training researchers within the Penguin Research Group. I estimate the total funding requirement for this project to be $169,080. Within the attached project proposal plan below I further describe the resources and costs for hardware and software, labor, and environments. After approval, the project's estimated start date is March 18th, 2024, and the estimated end date is June 7th, 2024.

I currently have 4 years of professional experience as a machine learning engineer. In the 4 years of my professional career, I have assisted in the development of many systems using machine learning algorithms for animal researchers to use as tools. With my ongoing history of furthering the enhancement of systems and tools for researchers, I can confidently recommend our system. The implementation of this system will address the Penguin Research Group's needs for identifying the species of penguins.

Thank you for considering my proposal! If you would like to address this opportunity further or have any questions or concerns, you can contact me with the given information below.

Sincerely,

Tenny Akihary

Machine Learning Engineer

takihary@hotmail.com

# Part B: Project Proposal Plan

## Project Summary

The Penguin Research Group faces the challenge of identifying the species of penguins accurately and efficiently. Researchers manually identifying penguins is time-consuming and is prone to human error. Research requirements and the number of penguins or data can grow or fluctuate, which also increases the need for more time and accuracy from researchers. This can hinder the Penguin Research Group in advancing their research, understanding, and conservation efforts.

The proposed Penguin Species Classification System can address these issues. This is done by providing a system that accurately and automatically identifies penguins based on the morphological features researchers record. The Penguin Research Group may research three different types of penguins, but the system will be compatible with any number of penguin species. Support Vector Machine (SVM), the machine learning algorithm to be used will be trained and tested to maximize accuracy. This will automate the process of identifying the species from the morphological features of penguins.

Implementing this system will provide researchers with more time to further Penguin Research Group’s research, understanding, and conservation efforts. Seamlessly identifying penguins will provide researchers with more time to dedicate to furthering penguin research, understanding, and conservation efforts. The user-friendly system will combat errors that derive from manual human identification. This will provide more accurate and complete data for Penguin Research Group’s overall understanding and research on penguins.

## Data Summary

Penguin Research Group has been researching and recording morphological data for 8 years. We will use the combined data over the years to train and test the SVM machine learning system. The system will input the CSV file of the combined data and set accurate and aesthetically pleasing column titles. Provide the source of the raw data, how the data will be collected, or how it will be simulated. Irrelevant or incomplete data will be dropped from the data frame for consistency, accuracy, and model performance. The data measurements of the morphological features recorded are unique to each species of penguin. This ensures the correct data will be passed to the model for training to provide accurate species predictions from user data inputs. The system will be compatible with scaling when Penguin Research Group increases its research efforts. The data frame that will be used can be easily modified to handle any number of penguin species.

## Implementation

An agile methodology will be used in this project’s Penguin Species Classification System implementation. Below is an outline of the implementation that uses an Agile methodology:

* **Planning**: Information and relevant input from stakeholders will define the requirements and goals for the project. This will assist in creating a backlog for tasks based on priority. Determine the scope of all phases and iterations of the project.
* **Design**: The classification system can then be outlined with the design of the user interface and data processing. Closely involving and informing stakeholders to assist in the creation of the user interface. Design the model architecture for the SVM machine learning classification.
* **Development**: Use the backlog from the planning phase to iteratively implement features and functionalities. Receiving feedback after each increment and modifying features as needed. Facilitate continuous integration and maintain accurate and quality code through collaboration.
* **Testing**: Performing stress tests to gauge system stability. Ensure functionality in different environments with compatibility testing. Load testing to benchmark the system’s capacity. Involve stakeholders and conduct user acceptance testing for further validation.
* **Deployment**: Prepare for fully implementing the system. Ensure coordination and collaboration between all teams for a seamless deployment. Continue to run performance metrics and address any errors or faults that occur during implementation.
* **Review**: Conduct regular meetings to help promote returning feedback (both good and bad) and collaboration throughout the project. Review the performance of each team and recognize potential improvements. Gather potential ideas and spread awareness of any known issues.

## Timeline

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone or deliverable** | **Duration**  **(hours or days)** | **Projected start date** | **Anticipated end date** |
| Planning | 5 Days | March 18th, 2024 | March 22nd, 2024 |
| Design | 10 Days | March 25th, 2024 | April 5th, 2024 |
| Development | 25 Days | April 8th, 2024 | May 10th, 2024 |
| Testing | 5 Days | May 13th, 2024 | May 17th, 2024 |
| Deployment | 5 Days | May 20th, 2024 | May 24th, 2024 |
| Review | 10 Days | May 27th, 2024 | June 7th, 2024 |
| ***Total*** | ***60 Days*** | ***March 18th, 2024*** | ***June 7th, 2024*** |

## Evaluation Plan

Verification methods used for each stage of development are as follows:

* **Planning**: Review the overall requirements with stakeholders and define user acceptance for guidance throughout the project.
* **Design**: Gather team members and stakeholders for an overview of designs. Designs include user interface, data processing, and the machine learning model.
* **Development**: Double-check the priorities and values of the items in the backlog with stakeholders. Participate in various code review activities and test the integration of components.
* **Testing**: Verify the accuracy of the machine learning model through training and testing. Conduct functional and regression testing to make sure the system is always functional.
* **Deployment**: Check the status of the deployment in the environment. Verify basic functionality to ensure the system is working as intended.
* **Review**: In collaboration with stakeholders, identify topics that are reviewed multiple times throughout the development to ensure issues or feedback have been properly handled. Reflect on areas for improvement that have yet to be implemented.
* **Completion**: Validate the machine learning model accurately classifies penguin species given a variety of morphological inputs. Involve researchers and stakeholders in further user acceptance testing to ensure the system is at full functionality. Record the times at which a researcher uses the machine learning model to correctly predict a species. The time for researchers to identify a species using the Penguin Species Classification System will be quicker than the time it takes for researchers to identify the species manually.

## Resources and Costs

* Hardware and Software: Open-source libraries will be used, which assists in limiting the cost of certain software. Penguin Research Group has an abundance of computers that have been purchased themselves, donated, or funded through the government. Penguin Research Group does not have servers to assist in deploying, operating, and maintaining the product. Three servers must be purchased for current users, maintenance, and backup. Each server is $4,500, resulting in a total of $13,500 for servers.

|  |  |  |  |
| --- | --- | --- | --- |
| **Resource** | **Amount** | **Single Cost** | **Total Cost** |
| Libraries | 6 | $0 | $0 |
| Computers | 38 | $0 | $0 |
| Servers | 3 | $4,500 | $13,500 |
| **Total** |  |  | **$13,500** |

* Labor: Four software developers will be utilized throughout the entire project. Contracting these software developers will be at a rate of $2,115/week for each developer. Which results in a total sum of $101,520. Two software quality assurance positions will be filled at a rate of $1,346/week for each position. For 12 weeks that results in a sum of $32,304. One project manager will be recruited for the entirety of the project at a rate of $1,730/week, which results in $20,760. Researchers and stakeholders will be involved throughout the project. However, costs for those employed at Penguin Research Group will not be included in the costs for this project.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Resource** | **Amount** | **Single Cost** | **Weeks** | **Total Cost** |
| Software Developers | 4 | $2,115 | 12 | $101,520 |
| Software Quality Assurance | 2 | $1,346 | 12 | $32,304 |
| Project Manager | 1 | $1,730 | 12 | $20,760 |
| Researchers | 18 | $0 | 12 | $0 |
| Stakeholders | 6 | $0 | 12 | $0 |
| **Total** |  |  |  | **$154,584** |

* Environment: The language is Python and the IDE of choice is PyCharm which is $249 for each developer, resulting in a total of $996 for the IDE.

|  |  |  |  |
| --- | --- | --- | --- |
| **Resource** | **Amount** | **Single Cost** | **Total Cost** |
| PyCharm | 4 | $249 | $996 |
| **Total** |  |  | **$996** |

# Part C: Application

Refer to the uploaded files:

* prgnotebook.ipynb
* penguins\_size.csv

# Part D: Post-implementation Report

## Solution Summary

The Penguin Research Group needed a method for seamlessly predicting the species of penguins. Time is important when it comes to utilizing researchers. Researchers had previously manually identified the species of penguin, which leaves data susceptible to human error. Using the Penguin Species Classification System researchers can now more quickly, efficiently, and accurately identify penguin species. This will facilitate a rise in the speed and accuracy of research to promote Penguin Research Group’s research and conservation efforts.

## Data Summary

The dataset originated and was obtained from Kaggle at the following link: <https://www.kaggle.com/datasets/amulyas/penguin-size-dataset>

The raw data in the CSV file ‘penguins\_size.csv’ was read and stored into a variable using the pandas library. The class reads and divides the comma-separated values into a dataset. A screen shot of a computer

Description automatically generated

Although the raw data contained somewhat relevant column names, they were renamed to cleaner, more aesthetically pleasing titles.

A screen shot of a computer code

Description automatically generated

Afterward, the data frame was printed to view the data with the new column names to confirm the name change and ensure data was properly stored.

A screenshot of a computer

Description automatically generated

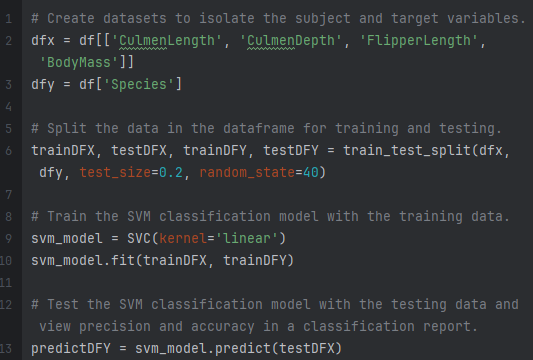
Although the column names and data were accurately stored, there were records (rows) with missing data points. The data also included columns that are irrelevant to our intended system. The irrelevant columns and records with missing data points were dropped. Then the data frame was viewed once more using a print statement to review and finalize the data preprocessing. A screenshot of a computer

Description automatically generated

## Machine Learning

For the penguin species prediction system, a Support Vector Machine (SVM) module using its Support Vector Classification (SVC) class from the Scikit learn library was the proposed solution. This supervised learning algorithm is typically used for tasks involving classification and regression. This system used the SVM for classification tasks by calculating the best hyperplane that separates data into different classes, increasing the margin between classes. The SVM library is proficient in high-dimensional feature spaces and has a variety of kernel functions and features.

Different kernels from the library were experimented with to determine the one that best suited this system resulting in using a linear kernel. After the data was preprocessed, it was split into training and test sets. Different splits of the testing and training data were tested for accuracy. The lower the percentage of training data, the less accurate the model’s outcome. However, all testing percentages below 20% were also more inaccurate. Given the results of the tests, the split was set to 20% testing data and 80% training data to give the model abundant information for more accurate training with enough data left over for testing the model. A random state was set to control the shuffling applied to the data and assist in reproducibility. Any other values or parameters for classes in the SVM remained default.



The SVM was selected for its ability to calculate a prediction for a categorical feature (species) from numerical features (culmen length, culmen depth, flipper length, body mass). The linear kernel parameter in the SVC class was set for its simplicity and interpretability. Improved accuracy in handling high-dimensional data and its overall flexible model in multi-class classification is why the SVM was selected for predicting the species of penguins based on morphological features.

## Validation

For the validation of the SVM classification model, we can review the accuracy (correct predictions/total predictions) using metrics. The classification report function from the Scikit Learn library was one metric used for measuring the accuracy of the supervised classification method. This class gave additional information such as precision, recall, f1-score, support, macro average, and weighted average values that may be of use in the future.

A screenshot of a computer

Description automatically generated

Another metric used was a confusion matrix to again calculate the accuracy by comparing the actual data (testDFY) with the predicted data (predictDFY) while also providing a visualization of the results.

A screenshot of a computer program

Description automatically generated

## Visualizations

I used a pair plot from the Seaborn Library to visualize the relationships between the morphological features of the different penguin species.

A screenshot of a graph

Description automatically generated

After reviewing the morphological features, I realized the biggest difference in data of morphological features and their relationships was each species’ culmen length. A violin plot was implemented to better visualize this difference. A violin plot will help display the distribution and summary statistics, can handle uneven sample sizes, compares across each species, and is aesthetically pleasing.

A diagram of different species

Description automatically generated

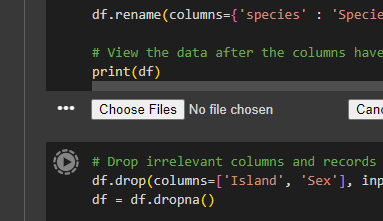
As mentioned in the validation section, a confusion matrix was created to help measure the accuracy of the machine-learning model. A confusion matrix heatmap is a useful tool when evaluating and visualizing performance.

A screenshot of a graph

Description automatically generated

## User Guide

### Guide for Google Colab with a Google Account

1. Click on the link to open the notebook: <https://colab.research.google.com/drive/1IPjVnJLq8EBdTYjeJpKc3wFE4FYmgbGB?usp=sharing>
2. Under the Runtime tab, select “Run all” or press Ctrl+F9 to run all cells.
3. After the first cell begins running it will prompt you to upload the CSV file, select “Choose Files”. 
4. Locate and select the “penguins\_size.csv” file that was provided and click open.
5. The remaining cells will execute, and you may review data or use the data inputs at the bottom to predict a species.

A screenshot of a computer program

Description automatically generated

### Guide for local python IDE

1. Download the following libraries using ‘pip install’ in the command prompt of your IDE:
   * Notebook
   * Pandas
   * Matplotlib
   * Seaborn
   * Scikit-learn
   * Ipywidgets
2. Type ‘Jupyter Notebook’ to open Jupyter Notebook in your default browser.
3. Locate the following files and ensure they are in the same directory:
   * penguins\_size.csv
   * prgnotebook.ipynb
4. Restart the kernel and run all cells by selecting the button (as shown below) towards the top. A screenshot of a computer

   Description automatically generated
5. All cells will execute and you may review data or use the data inputs at the bottom to predict a species.

