Abbey Allen
Ben Shaw

Project 7

Introduction

This report aims to provide a brief overview of two datasets, namely, penguin characteristics and forest fires in Portugal. The penguin characteristics dataset contains information on different characteristics of penguin species, such as bill length, flipper length, body mass, sex, and species. The second dataset pertains to forest fires in Portugal and includes information on factors such as temperature, humidity, wind speed, and rain.

The purpose of analyzing these datasets was to aid those in environmental positions to accurately categorize penguins by species characteristics and to determine the likelihood of a forest fire occurring. To achieve this, logistic regression and Support Vector Machines (SVM) were employed as the analysis techniques. The report will discuss the datasets, the analysis techniques used, and the results obtained, highlighting the importance of the analysis in aiding conservation efforts and preventing forest fires.

Github: https://github.com/ablibranix/project7
Presentation: Project 7

Dataset

Penguin Dataset

The penguin data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network. The data contains 345 rows, with 3 different species of penguins, namely Adelie, Gentoo, and Chinstrap. In this project, we utilize the simple version of the data, the rows of which have the following attributes: species, native island (near Antarctica), bill length, bill depth, flipper length, body mass, sex, and year the observation was made (2007 - 2009). 146 of the rows are claimed by the Adelie species, 119 of the rows are claimed by the Gentoo species, and the remaining 68 of the rows are claimed by the Chinstrap species.

Forest Fire Dataset

This dataset contains information for forest fires in the Montesinho Natural Park in northern Portugal. It consists of 517 instances of forest fires from January 2000 through December 2003 with each fire containing 13 attributes: X, Y, month, day, FFMC, DMC, DC, ISI, temperature, humidity, wind, rain, and burned area in hectares. The X and Y attributes coordinate to a

cross-section of the park from the paper “A Data Mining Approach to Predict Forest Fires using Meteorological Data” by Paulo Cortez and Anibal Morais. The FFMC, DMC, DC and ISI are codes used by the Fire Weather Index with Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC) and Drought Code (DC) relating to fuel moisture codes and Initial Spread Index (ISI) being a fire behavior index. For total burned area, anything below 1 hectare is labeled as 0. Key attributes used for this project were area and FFMC. Other attributes were used to visualize the data over the course of time.

Analysis Technique

Penguin Technique

Two rows of data contained missing values: these rows were removed from consideration. Additionally, we wish to predict the species using only the bill length and bill width column. We used 80% of the data for training, and 20% for validation. We created an additional attribute of the data, namely “target,” which was simply an integer representation of the “species” column. We also standardized the data, calculating the mean and standard deviation from the training dataset. Our evaluation metric is the accuracy, and we employ two predictive models, namely logistic regression and a support vector machine with a linear kernel.

Forest Fire Technique

Support Vector Machine (SVM) and Logistic Regression were applied to our dataset of forest fires in Portugal. Both techniques are well-suited for input data classification. SVM can identify complex decision boundaries, while Logistic Regression estimates the probability of an input belonging to a particular class. The dataset contains multiple features, such as temperature, humidity, wind, and rain, which these algorithms can handle. SVM and Logistic Regression have been widely used in environmental studies and are effective in similar tasks, making them a popular choice for this domain.

Results

Penguin Results

The logistic regression model classified the data with a validation accuracy of 0.925. Meanwhile, the linear support vector machine classified the data with a validation accuracy of 0.94. Figure 1 depicts the decision boundaries for logistic regression, while figure 2 depicts the decision boundaries for the linear support vector machine. It is noteworthy that the decision region for the Chinstrap class is more narrow in the SVM model than with the logistic regression model.

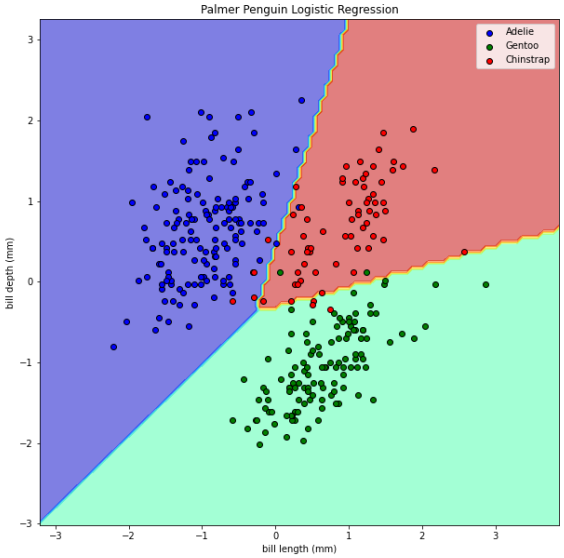


Figure 1: The penguin dataset by class with decision boundaries learned by logistic regression. Note that the bill length and bill depth features have been standardized.

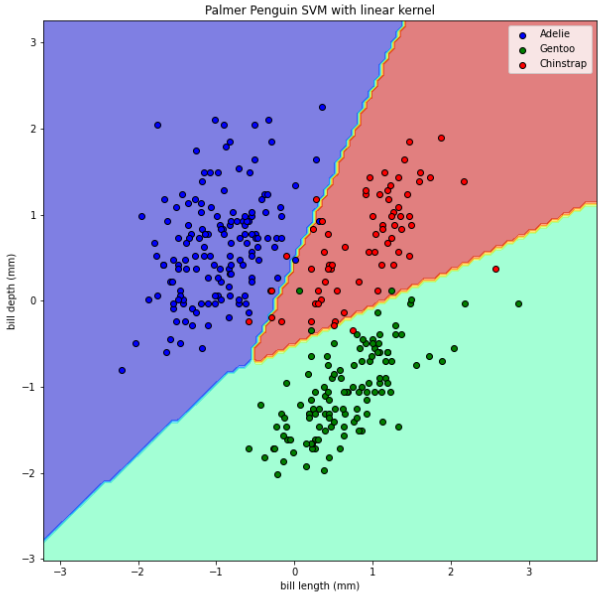
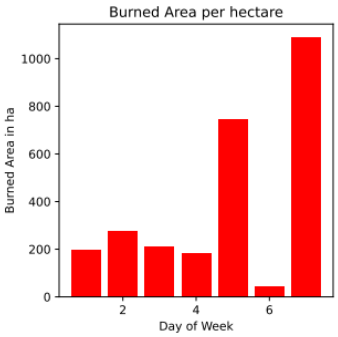


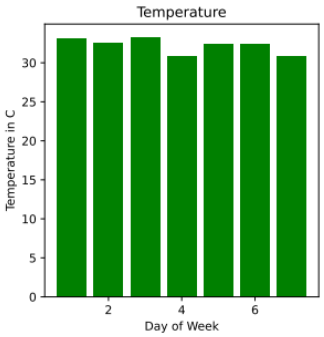
Figure 2: The penguin dataset by class with decision boundaries learned by a linear support vector machine. Note that the bill length and bill depth features have been standardized.

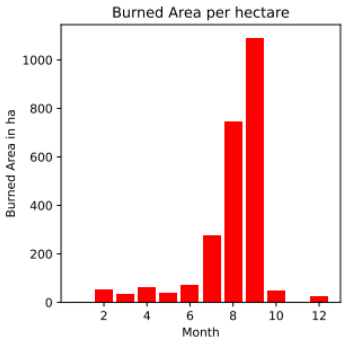
Our models suggest that the bill length and bill depth can be reasonably relied upon to discriminate between the three classes of penguins. This will allow researchers to use, in particular, outlying bills for penguins to predict the species of the penguin.

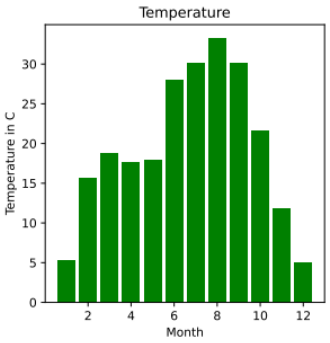
Forest Fire Results

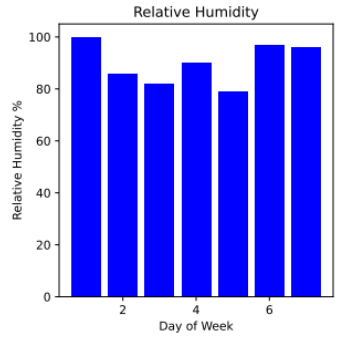
The analysis revealed that the burned area varied significantly by month and week. Specifically, higher areas were observed during weekends and in the fall months, particularly August and September. The number of fires was found to be positively correlated with the burned area, which is logical as more fires result in a larger area being burned. Additionally, we examined the relationship between wind, humidity, rain, and temperature with the occurrence of fires to identify potential attributes for logistic regression and SVM.

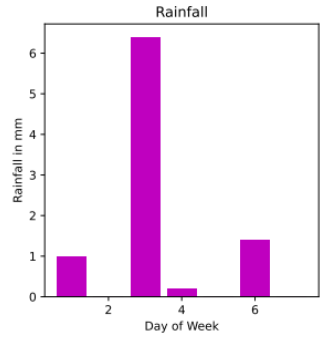


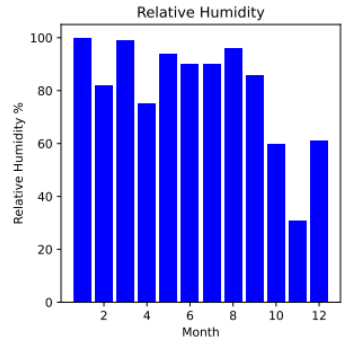


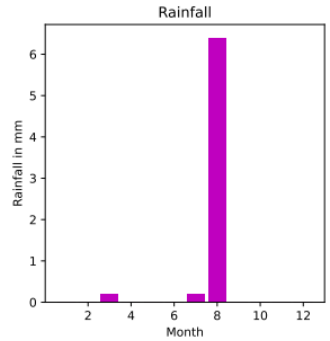


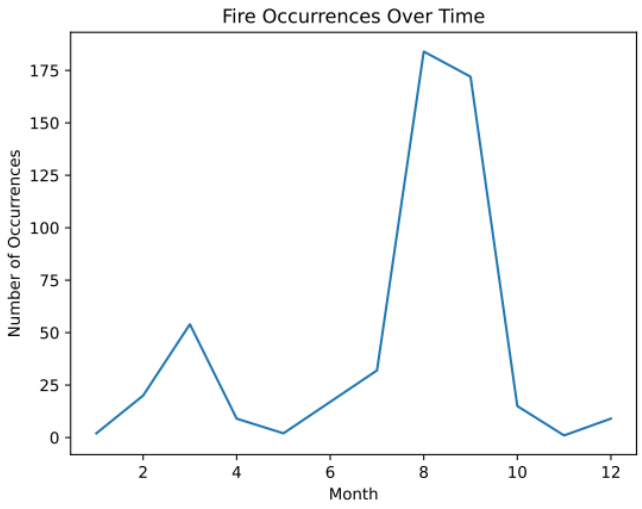


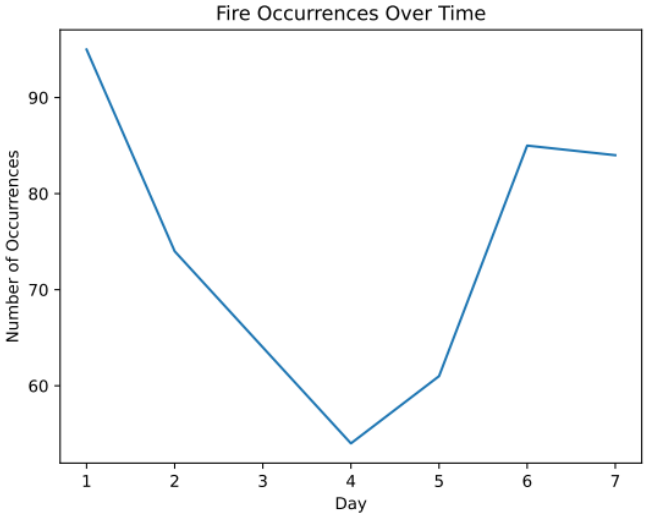




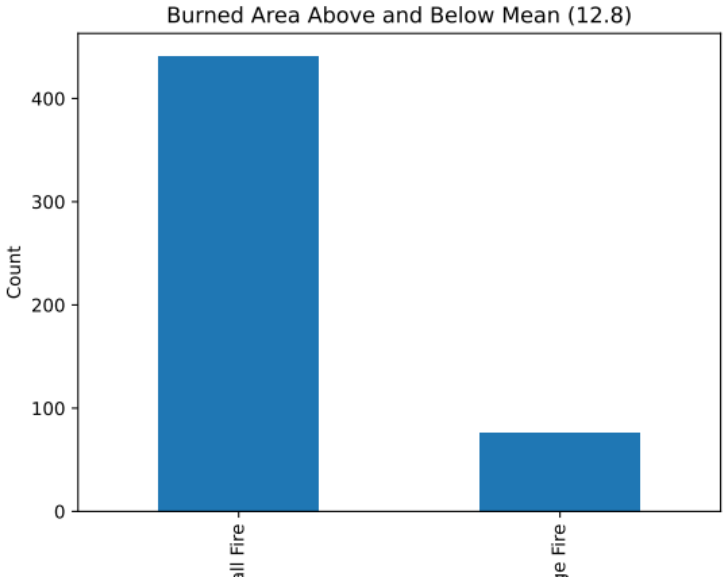








To further explore the data, we categorized fires as small or big, based on the burned area exceeding 12.8 hectares, and achieved an 85.7% accuracy when using the mean and a 51.9% accuracy when using the median. This information can assist the forest service in preventing fires or detecting them when they are still small. This approach is faster than traditional manual look-up methods and can leverage the existing probability of fires. The results were consistent when applying SVM using the same attributes and target value.



Technical

For the penguin dataset, we used a simplified version of the original data. The features of the original data are much less interpretable and require domain knowledge to interpret. For the purposes of visualization, only two features were used: the bill length and the bill width. However, when the body mass and flipper length features are also used, an accuracy score of 1.0 can be obtained where only about 50% of the data is used: this dataset is perfectly linearly separable, it seems. This accuracy score applies to both the logistic regression model and the linear SVM: the linear separability of this dataset gives justification for the use of models which impose linear decision boundaries, such as logistic regression and the linear SVM.

We also defend the case for using accuracy as our evaluation metric. Even though the Chinstrap species class is a minority class–accounting for approximately 20% of the data–we do not believe that other metrics offer substantially better insight.

For the Forest Fires dataset, we did two things with the month and day columns. First, we assigned numerical values using a dictionary for charts and graphs. Second, we performed one-hot-encoding for logistic regression and SMV. We also standardized all the features being used for the logistic regression and SVM to help with uniformity. Lastly, we changed our area column from numerical to binary to determine a threshold for our logistic regression.

SVM and logistic regression were required so we were tasked with using those techniques. We thought they are effective for this because they can handle multiple input features and have been widely used in environmental studies in the past.

The analysis process was not as complex as we initially thought it would be. Once the dataset was loaded, we first wanted to chart some attributes over time since the dataset is information over multiple years and wanted to know which attributes would be valuable. After reading the paper written on the dataset, we were able to figure out the key variables for our Logistic Regression and SVM functions. Since our area was quantitative, we changed it to a binary variable based on the mean and median of the column as a whole. We then assigned those values as our thresholds and performed 2 instances of logistic regression and SVM. We don't think we would have changed much with our process except maybe adding a nice chart similar to the one in the penguin dataset showing our test and training data but was not able to.