November 2nd, 2022

Nicholas Fallico - C964: Computer Science Capstone

Task 2 parts A, B, C, and D

Part A: Project Proposal for Business Executives

Letter of Transmittal

Nicholas Fallico
Founder of Machine Learning Solutions
123 Main Street
Chicago, IL
October 29th, 2022

Jeff Bridges
Fresh Anglers
Chief Executive Officer
141 W Jackson Blvd. Suite 2300
Chicago, IL, 60604

RE: Proposal for A Machine Learning Solution to Identify the Species of A Given Fish

Dear Mr. Bridges,

Machine Learning Solutions is an industry leader in providing data products and solutions to our business clients. We pride ourselves on identifying areas within a business that can benefit from our modern data product solutions. We identify such areas of opportunity, and we draw up the solution before contacting any new potential business partners. We are so confident in our work that we put in the time to draw up a solution prior to reaching out to you. As a result, this data product will benefit both your company and your customers.

I understand your business model targets the novice angler. Your business prides itself on helping and educating the new fishermen of today. I have been known to enjoy some leisurely fishing since I was a child. After all these years, I still consider myself a novice because I never put in the time to learn more than casting the line and trying to reel them in. I am proposing a data product that will inform the novice fisherman about what fish they have caught. It will work by having the fisherman, or end user, input six distinct measurements of the fish they caught. I will create a machine-learning solution that will run on a command-line interface. If you wish, you can integrate it into your app or website. You can also set up a display booth in your stores to allow customers to try it out. You can incorporate it in many ways.

This data product will help your customers identify fish. Not only that, but it will also get them more involved in fishing. That is because, with every catch, they will take the time to measure the fish and input their measurements into your system. Having your customers use this data

product is beneficial for two reasons. The first reason is it ensures the customer is engaging with your business through the automated specifies finder. The second reason is that the more your customers learn about fishing, the more passionate they will become about it. These two benefits will ultimately make your customers want to go fishing more often. If this data product makes your customers want to fish more often, then the more they will continue to shop at your establishment. Moreover, you could place ads wherever you host the data product. It's a win-win for you and your customers.

This data product will take four weeks to complete. It will cost your company $40,000, guaranteed. Machine Learning Solutions will cover the excess if the project cost goes over $40,000. I know this product will benefit both you and your customers. I started this company 20 years ago. Before that, I was a software developer for ten years. My company is filled with people with decades of experience each. If anyone can create a data product to help you capitalize on your customers, it is us.

Best Regards,

Nicholas Fallico

Problem Summary

Machine Learning Solutions has 20 years of hard work and success, making us an industry leader. Typically, individuals who work for our company have more than 20 years of software development experience under their belts. The data product we will create for Fresh Anglers will allow an end user, such as the customer, to input six distinct measurements of a fish. Once the measurements have been inputted, the customer will receive a prediction answer as to the species of fish they have measured. This data product will allow your novice customers to become more knowledgeable about the fish they encounter. Because your company focuses on helping and educating new fishermen, this aligns perfectly with your business goals.

A command-line interface program will be delivered. As previously stated, the program will give an output of the name of a fish species based on six measurement inputs. Because we are creating just the machine learning program, you can integrate it into your website, mobile app, or wherever your business sees fit.

Application Benefits

There are direct benefits to my company's proposed solution. This data product will allow your customers to educate themselves about the fish they encounter. Educating your customers will help them become better fishermen. It will also garner respect and customer loyalty to your brand. As such, you are securing returning customers for your business.

Additionally, an automated fish species identifier is desperately needed because there is nothing else out there like it. The market currently has no fish-identifying application. Your company would be the first! This product will help retain current customers, and it will also help bring in new customers when people hear about your new product. This great product will encourage existing customers to tell their friends about it. Word-of-mouth is like free advertising.

There are also indirect benefits to the solution. From one fisherman to another, I'm sure you know fishing can be competitive. It can be true even when it is just you and some buddies fishing in a nearby pond. Fishermen who have access to this data product will be tempted to measure every fish they catch. When a group of people are measuring their fish catches, it is safe to assume they will compare measurements. The fish species categorizing solution will indirectly bring out the competitiveness of people. This competitiveness can cause people to become more involved in fishing. As such, you are more likely to create more returning customers.

Application Description

The data product will use a machine learning prediction model. Our company will use machine learning to create a data product that will predict an output based on given inputs. This will be done through what is called "training." Machine learning training is akin to giving a person flashcards, except, in this case, there will be many, many flashcards, and the computer's memory is perfect. Once the training is complete, you can give the machine learning data product some inputs. Then it will match the input values to the closest ones from the original data. Based on the most immediate data values that match, it will use the output of the initial data values to give a prediction. The result is called a prediction because it compares the new inputs with the original training data inputs to make a guess.

The application will be trained with data on seven species of fish. Once the training is completed, the solution will be able to output the name of a fish species based on six input measurements from the end user. This data solution solves the problem for novice fisherman who doesn't know how to identify the fish they have caught. Currently, if a novice person catches a fish, they would have to guess or use a web search to try and figure out what fish they have caught. There is no definitive way for them to know the species of their fish on their own. The data product solves that problem by giving them the name of the species of their fish based on the measurements of the fish.

Data Description

The data I used is from a public dataset found on kaggle.com. It does not contain personal or classified information. For the purpose of this project, I am hosting the raw data on my GitHub. You will see the GitHub link used in the python code. The data type is nominal and structured. Each row represents data for one fish. The data could be considered a flat-file database structure because it is stored and presented in a way that mimics a relational database system, except that it does not contain any keys or indices.

The dependent variable is the first column. This column contains the species of fish. The independent variables are the following six columns. These columns are as follows: 'Weight (in grams),' 'Vertical Length in CM', 'Diagonal Length in CM', 'Cross Length in C,' 'Height in CM,' 'Diagonal Width in CM.' There are no outliers in this dataset. The limitation of this set is its size. There are 159 lines of data, or 159 individual fish, recorded. This is small compared to typical machine learning solutions, where they used thousands of data records to train their predictive model.

Objectives and Hypothesis

The objective of this project will be to create a data product for Fresh Anglers that uses machine learning to predict the species of a given fish. The objective will be considered completed when these two goals are reached. First, my company hands Fresh Anglers a finished data product that can predict the species of a given fish based on the user input of six distinct fish measurements. Second, we test the accuracy of the prediction model when the data product is hosted on their system. The desired accuracy for this prediction model will be 80% or higher.

Methodology

This project will use the Iterative Waterfall Methodology. The Iterative Waterfall Methodology differs from the standard Waterfall Methodology in that we can go back a phase if need be. I am not expecting the need to go back a phase, but it is on the table if needed. I chose this methodology because setting up a machine learning model and subsequently training it will be done step-by-step in a predetermined order. This project does not need an agile methodology because there will be no customer involvement. My team will create and test the solution. Again, I do not expect to need to go back a step, but if it were to happen, I would expect it to occur during the testing phase. For example, if we tested the trained solution and the accuracy was low, we might want to go back and determine where the issue lies or what else needs to be done.

Outline:

1) Determine the requirements. When it came to creating a solution to determine a fish's species, we thought of what requirements would be needed. We knew we needed more than one identifier from a fish to categorize it. Ultimately, we decided we will need six distinct measurements of a fish. Those measurements are: 'Weight (in grams)', 'Vertical Length in CM', 'Diagonal Length in CM', 'Cross Length in CM,' 'Height in CM,' and 'Diagonal Width in CM.' Once we decided on this, we searched for a dataset that met our requirements.

2) Design/Software Architecture. This program will be a command-line interface program. It will run inside an IDE. An IDE is short for Integrated Development Environment. An IDE is a program that lets us create other programs or software using code. The coding language that will be used for the project is called Python. The IDE that will be used to develop the data product will be IntelliJ CE 2022.2.2, build IC-222.4167.29. The machine that will be used to create the data product will be a 2020 MacBook Pro running Mac OS 12.5.1.

3) Implementation/Software: Our company will use Python to create the machine learning data product. We will also use the pandas library and the sklearn library. Pandas and sklearn are what are known as external libraries for the Python language. They will allow us to use functions not present in the default Python library of functions. For example, pandas will allow us to use data frame objects. Sklearn will allow us to use a logistic regression model to give us our predictions.

4) Verification/Deployment: Once our predictive model has completed training, we will test it for accuracy. We will consider the software creation process completed if the model has above 80% accuracy. If it is lower, we will go back to the implementation/software phase, determine where the issue lies, and fix it.

5) Maintenance: Maintenance for this data product will be minimal. It will need to be kept up-to-date as the Python language, and external libraries change with time. While there will be minimal updates to the language and libraries over time, I would not expect any significant major changes that would affect this program. I expect this program to run normally for at least five years without modification. Optionally, your company may choose to add more fish to the dataset. There are two benefits to this: adding more data to the species already in the dataset will increase the prediction model's accuracy. Additionally, you can add data for species of fish that will not be present when the data product is delivered. This second benefit will allow you to identify as many species of fish as you want.

Funding Requirements

We propose this project will be finished in four weeks. We expect the cost to be $40,000, and we will deliver this data product to your firm for that exact cost. Our company will not charge your firm extra if we run over the price. This project will be completed by one person. However, more will be added if the project runs behind. The software engineer's house will be

the physical environment where this product will be developed. It will be done on their MacBook. Our company has its employees working from home. The software environment will be Mac OS 12.5.1. The program that will be used to create the software is IntelliJ CE 2022.2.2, build IC-222.4167.29. This version of IntelliJ is free. The dataset we will obtain will also be public and free.

Data Precautions

Our company will be sure to use a public and free dataset. It will be obtained from kaggle.com. We will be sure that there will be no protections on the data. We will be sure to use data that will not contain private or personal information.

Developer's Expertise

My company has been in business for 20 years. I have 30 years of experience. Most of my employees have decades of experience. The firm specializes in machine learning solutions. Creating a fish identifier based on user input is a type of project we create regularly.

Part B: Project Proposal

Problem Statement

The problem novice anglers will face is the daunting task of identifying a fish they have caught. With 196 species of fish in Illinois alone, even experts will have difficulty identifying freshly caught fish ("Common Freshwater Fish of Illinois," 2003). Considering your company, Fresh Anglers, has stores all over the Midwest, the number of species your customers will encounter will surely be great. Therefore, they will need a tool to educate them about the types of fish they are catching.

Customer Summary

The customers for your company will be novice fishermen or anglers per the industry term for fishermen. The proposed data product is a machine-learning solution. We will leverage pandas and sklearn with Python to create a multinomial logistic regression prediction model. Once trained, the solution will quickly solve the problem of novice anglers who cannot identify the species of fish they have caught. The novice fisherman will be able to enter measurements of a fish, and the data product will tell them its predicted answer with a high degree of accuracy.

Existing System Analysis

Currently, there is no fish-identifying solution provided by your company. As a result, novice anglers must resort to web searching the traits of their fish to determine what kind of fish they have. This process will be shoddy at best. The proposed data product will make it easy for anyone to determine the fish they have based on their given inputs about the fish. This process will be easier, fun, and more efficient than using the internet to guess.

Data

The raw dataset will be a public dataset found on kaggle.com. We will look for a dataset that is a one-page, flat-file database. If need be, we will modify the dataset to fit our needs. For example, if there are outliers or rows of incomplete data, we will remove those. However, we will search for datasets with the least number of outliers or incomplete data.

Project Methodology

We will use the Iterative Waterfall Methodology to complete this project. This method will be good for us because we have an excellent understanding of the requirements and scope. We will also be constrained in time by four weeks. We also do not expect to retreat to previous phases during the process, but this methodology lets us do so if necessary. We also chose this methodology over an Agile methodology because we will not be working with any customers. Our company will create and test the data product. Once complete, we will submit the final project to your company.

Here is an outline of our methodology plan:

1) Determine the requirements. When creating a solution to determine a fish's species, we will consider what requirements will be needed. This is where we will search for a fish species dataset. We will search for datasets that are public, free, and contain no private or personal information. We will also search for datasets that have minimal outliers or incomplete data. Finally, we will search for a dataset containing the fish species as the dependent variable and the fish traits as the independent variables. Once we find the dataset that meets our requirements, we will modify the dataset to our needs, and then we can move on to the next phase.

2) Design/Software Architecture. This program will be a command-line interface program. It will be created and made to run inside an IDE. The IDE used to develop this data product will be IntelliJ CE 2022.2.2, build IC-222.4167.29. The machine used to create the product will be a 2020 MacBook Pro running Mac OS 12.5.1. We will design it to be small and efficient. We expect the program will be less than 100 lines of code.

3) Implementation/Software: Our company will use Python 3.9 to create the machine learning data product. We will also use pandas 1.5.1. From the pandas library, we will be using data frames. First, we will cast the raw data into a data frame. From there, we will create a data frame containing the list of dependent variables and a data frame containing the data for the independent variables. We will also use sklearn 1.1.1. From the sklearn library, we will use the built-in logistic regression model— the LogisticRegression() function. Additionally, we will use train_test_split() function, the fit() function, and the predict() function.

4) Verification/Deployment: We will use the sklearn train_test_split() to split the data between training and testing data. Into this function, we will input the dependent variable data frame, the independent variable data frame, and a test size of 0.3. A test size of 0.3 means the function will not train 30% of the data. This 30% of data will be saved for testing later. After this, we will use the sklearn fit() function to train the 70% of data set aside for training. Once our predictive model has completed training, we will test its accuracy by running the sklearn predict() function through the sklearn accuracy_model function. The predict function will output the name of its guessed fish, while the accuracy function will output how accurate the prediction is. If the model

has above an 80% accuracy we will consider the software completed. If it is lower, we will go back to the implementation/software phase and fix the issue.

5) Maintenance: Maintenance for the proposed data product will be minimal. It will need to be kept up-to-date as the Python language and external libraries change with time. Generally, updates of software languages or external libraries are to eliminate bugs. I would not expect a need to modify the code for at least a 5 years. If a function used in the program is deprecated, you will receive a notification in your IDE. It will tell you how to update it. Optionally, you can decided to add more fish to the dataset to increase accuracy. You can also decide to include species of fish that won't be currently present in the original dataset. In any case case, you will need to save a copy of the original datafile hosted on my GitHub onto your own servers. Then you can add more fish data to the dataset. At that point you will be able to retrain the predictive model simply by replacing the current url to my dataset with a url to your dataset and running the program.

Project Outcomes

The deliverable to your company, Fresh Anglers, will be a data product that runs in an IDE. It will run on a Mac OS 12.5.1 computer, with IntelliJ CE 2022.2.2, build IC-222.4167.29. It should work on a Windows 10 computer with a different IDE, but the data product was created on a computer with the Mac and IntelliJ specifications.

USER GUIDE:

A. Installing IntelliJ CE (Skip to section B if you have this IDE or wish to use your own)

1. Click the link to go to IntelliJ's website to download the product. Because my project uses a Mac I will give instructions as if you are using a Mac. The instructions will be similar enough if you wish to use a Windows machine. https://www.jetbrains.com/idea/download/#section=mac
2. On the right side of the page there is black button to download the Community version. On the right side there is a dropdown arrow. Selected this and it will show you two options: Intel or Apple Silicon. Select the one that describes your CPU. The download will start automatically. It will automatically go to your downloads folder.
3. Double click the .dmg file. A new window will pop up.
4. Drag the IntelliJ IDEA CE icon onto the Applications folder icon. It may ask for an admin password. Have that ready in case.
5. IntelliJ IDEA CE is now installed with the latest version. As of 10/29/22 this will work for the proposed data product.

6. Eject the IntelliJ .dmg file by dragging it to the trash.

B. Opening and Setting up the Project in IntelliJ

7. After opening my submitted .zip file you should have a folder titled, “C964 Machine Learning Capstone Nicholas Fallico”.
8. Open IntelliJ. You should see a window with projects to choose from. You should also see a button that says Open on the top right.
9. Select the Open button.
10. Navigate to where you have my folder located.
11. Click once on the folder, “C964 Machine Learning Capstone Nicholas Fallico”. You want to highlight it but not open it.
12. Once the folder is selected, click on the Open button on the bottom right of the finder window.
13. IntelliJ should open the project and display the Main.py file.
14. The program is complete. You do not need to do anything to the code.
15. You will probably see some errors in the import section at the top of the page. It likely says you don’t have Python, pandas, or sklearn installed.
16. You should see a message on the top right of the window asking to install Python. Click yes.
    1. If you do not see this message follow these steps to install Python.
    2. Selected IntelliJ IDEA in the top right in the toolbar.
    3. From there, navigate to, “Preferences…” Once selected a new window will appear.
    4. There is a list of options on the left side in the new window. There is an option that says, “Plugins.” Select this option.
    5. From there you can search for Python in the newly appeared search bar. You will find a search result, “Python Community Edition.” Select Install.
    6. After you install, click Apply on the bottom right of the window. IntelliJ may ask you to restart. Do so.
17. If you hover over the imports a message will pop up asking if you want to download the libraries. Click yes.

1. If you do not see these options to install the pandas library or the sklearn library, follow the previous steps to install Python to install the pandas and sklearn plugins.

C. Using the program

1. You will not need to do anything with the code. It is complete.
2. When you run the program it will ask you to input the six distinct measurements as independent variables. Then it will give a text output of the name of species the logistic regression model predicts.
3. To run the program click on the green hammer button at the top of the window, right of center.
4. Then click the green play button to the right of the hammer.
5. In the run module in your IDE you will see the welcome statement and directions to input your six measurements.
6. The instructions will dictate to enter your six measurements on one line, each separated by a space.
7. Once you have entered your six measurements hit the enter key on your keyboard.
8. The program will then give you a text output of the name of the species it has predicted.
9. The program has now run and completed its task. You may exit the IDE.
10. If you wish to run the program again, refollow these steps under letter C.

Implementation Plan

We will be using the Iterative Waterfall Method to design and create the data product solution. As such, there will be no phases of rollout. We will not be working side-by-side with the customer or end user as one would in an Agile methodology. Our firm will create the data product from start to finish, and then deliver the completed product to your company, Fresh

Anglers. We are confident in this method because we will not be using complex data along with the fact that the programming functions we are using come from built-in functions of common, external libraries for Python.

There will be a few dependencies for this project. The first two dependencies will be the external Python libraries known as pandas and sklearn. The Python language itself will be another dependency, as well as the IDE suggested to use, IntelliJ. These technologies will be updated as time goes on. The maintenance will be minimal, but as these technologies are updated, the data product may need to be updated with it. This will be an easy task for your IT department.

The fish dataset will not need to be updated or maintained. The dataset is currently hosted on my GitHub url, but your IT department should make a copy of it as a CSV file. Then they should host it internally. Once that is completed, there will be little maintenance. That is, unless your company wishes to add more fish to the dataset. You may choose to do so should your company want to add more species of fish to the model predictor. If your company chooses to add more fish to the dataset, this can be done by anyone who knows how to type on a CSV file.

Evaluation Plan

Stage 1, Requirements: Our company knows we want to make a logistic regression model to identify a certain species of fish based on input measurements. The first requirement is to find a dataset that could be used for this. Once we have found a dataset that can be used we will verify the data works for our purposes by manually checking if the data has these three main components: 1) Each row of data represents one fish, 2) there is a column with the species data (this will be our dependent variable), and 3) there are columns of measurement data (these will be our independent variables).

Stage 2, Design: There will be no formal drawn design because the program will be extremely simple, and it will not have a GUI. The verification for this stage is verifying that we have modern hardware, and up-to-date software to create the Python program.

Stage 3, Implementation of Software: This will be the coding stage. First, the data will be extracted from my GitHub url into a data frame. Then we will print the data frame. If the data frame matches the data from the original dataset, then it will have passed verification. Second, the data will be split into two data frames; one for the dependent variable, and one for the independent variables. We will verify the data frames by printing the two new data frames. If they match the data from the original dataset, then they will have passed verification.

Next we will use the sklearn library to call its Logistic Regression function, its Train, Test Split function, its Fit function that will be used to train the data, and lastly, its predict function. These four functions will be used to initialize and train our logistic regression model. To verify if

these called functions gave our desired output we have to go to stage four. Stage four exists specifically to verify the logistic regression model works.

Stage 4, Verification/Deployment: This will be the stage to verify the multinomial logistic regression model that was implemented in the previous stage. To verify the multinomial logistic regression model is working we will use a print statement along with the sklearn library function Accuracy Score. Assuming there are no errors in the code, we will expect to receive an accuracy score of high value. If we do not receive an accuracy score of 80% or higher, we will go back a stage to reconfigure the model.

At this stage if the accuracy of the model is 80% or higher, then the project will be considered complete. The data product will then be delivered to Fresh Anglers.

Stage 5, Maintenance: Once Fresh Anglers places the data product on their own servers, we will verify it works by testing it seven times. That is one test of accuracy per each species of fish the data product can predict.

The validation in place upon completion of the project will have two steps: 1) verifying the accuracy on our own systems is above 80%, and 2) verifying the accuracy with the data product hosted on Fresh Anglers' server by testing it seven times, once for each species of fish in the dataset.

Resources and Costs

Hardware cost: $0 — we already own hardware.

Software cost: $0 — we will be using the free version IntelliJ. Additionally, there will be no extra cost to use Python, pandas, or sklearn.

Data cost: $0 — we will be obtaining a free and public dataset

The full $40,000 is the for labor costs. This project will have one software developer because it is not a difficult task.

Deployment cost: $0 — it does not cost money to give the Python project file to the customer, Fresh Anglers.

Environment costs: $0 — no cost to use to use the free version of IntelliJ, along with the hardware we already own. This data product will take up zero server space as it will be saved locally on the machine of the developer.

Hosting costs: $0 — our company will not be hosting this data product. Once the data product is handed over to Fresh Anglers, they will decide where and how to host.

Maintenance: $0 — the maintenance will be handled by Fresh Anglers once the data product is handed over. There will be no cost to us.

Timeline and Milestones

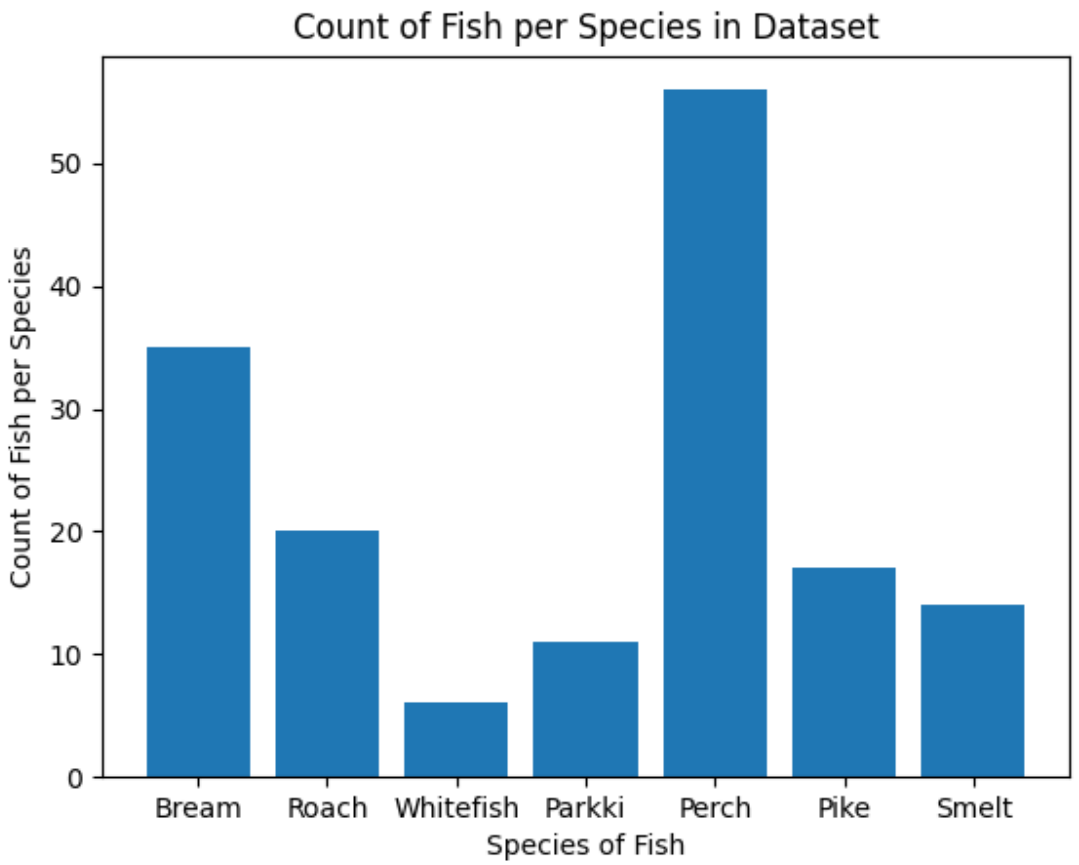
Task	Start Date	End Date	Cost
Determine requirements	11/1/2022	11/8/2022	$1,000
Find, verify, and validate data from kaggle.com	11/8/2022	11/15/2022	$2,000
Code the Python data product	11/15/2022	11/22/2022	$36,000
Test the data product	11/22/2022	11/29/2022	$1,000
Deliver product to Fresh Anglers	11/29/2022	11/29/2022	$0
		Total Cost	$40,000

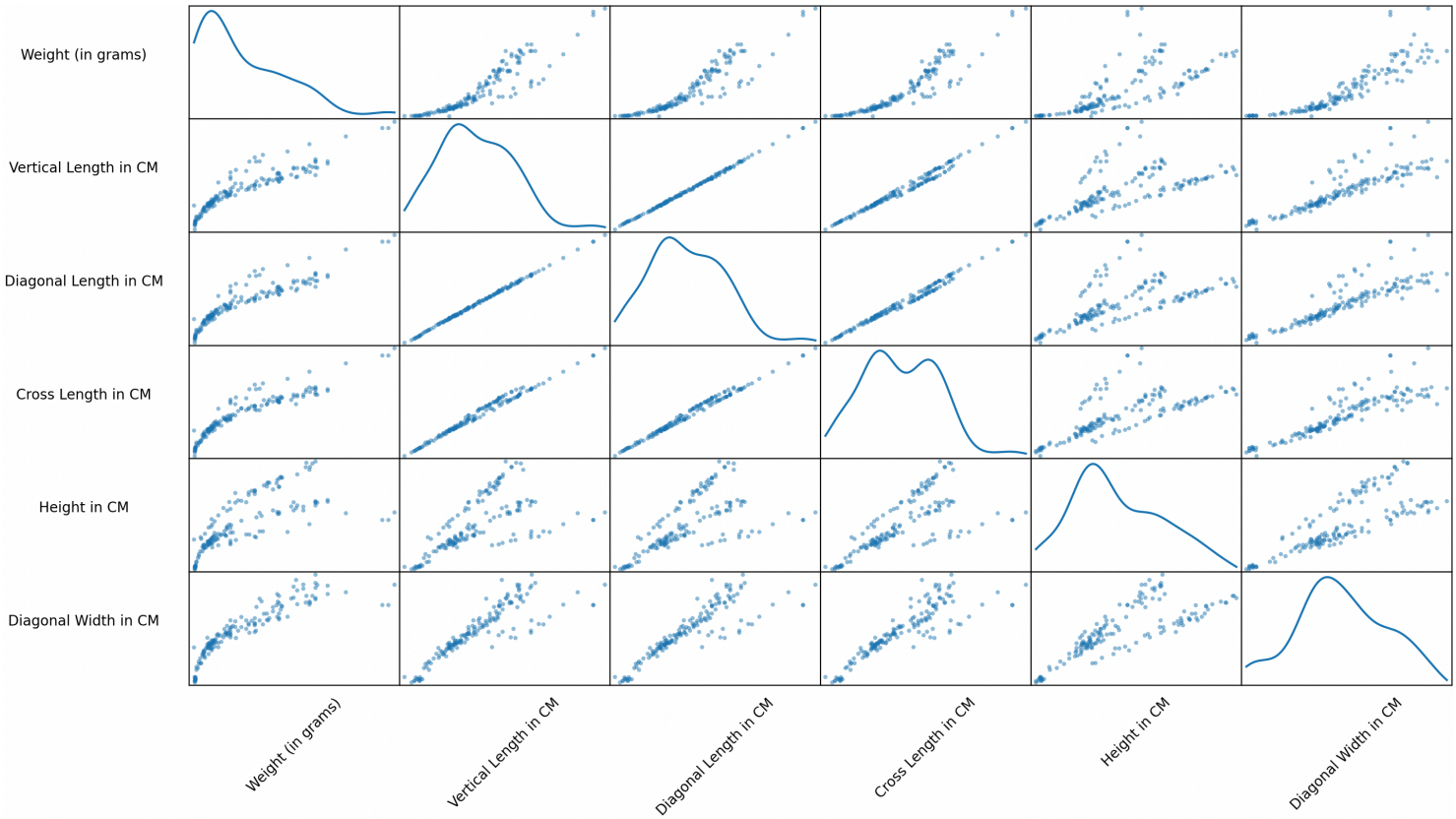
Part C: Application

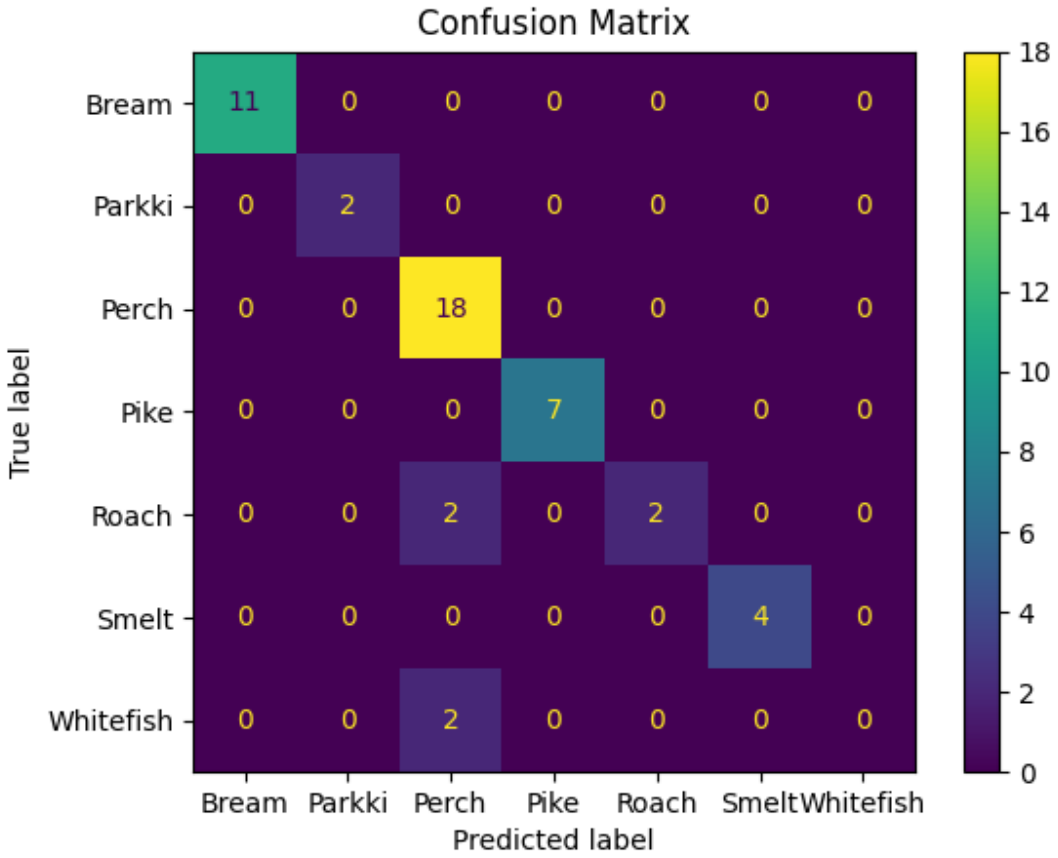
Part C is your submitted application. The document only needs to include a list of any submitted files or links.

Your submitted *application* (called a data product in the task directions) must include the following features:

- **Three visualizations (images). Static images are permissible.**




Weight (in grams)
Vertical Length in CM
Diagonal Length in CM
Cross Length in CM
Height in CM
Diagonal Width in CM
Weight (in grams)
Vertical Length in CM
Diagonal Length in CM
Cross Length in CM
Height in CM
Diagonal Width in CM



- **A *Descriptive method* = anything that describes the data.**
    - **Images can double count as your visualization and descriptive method.**
    - **Ex. mean, median, bar plot, scatterplot, k-means clustering, etc.**

  See images above.

- **A *Non-descriptive method* = anything that infers from the data, i.e., makes predictions or prescriptions.**
    - **Ex. classification models, regression, image recognition, etc.**

  I used a multinomial logistic regression model to make predictions.

- **An application of "machine learning" in the non-descriptive OR descriptive method (most data analysis algorithms are acceptable -including regression).**

  Same as previous answer.

- **An interactive "dashboard."**

    - **The application must be usable for solving the proposed problem. Any method enabling the user to interact is acceptable, including the command line. A GUI is *not* required.**

Running the program in the IDE will give a command-line interface input system.

- **A "user-friendly" interface.**
    - **Following the "user guide" of part D, the evaluator can successfully run your application as described on their machine.**
- **Security appropriate to your application's needs.**

In terms of security, this program is made to run on a local machine. It is much more secure than a program hosted online.

Part D: Post-implementation Report

A Business (or Organization) Vision

The problem Fresh Anglers' customers faced is the difficult task of identifying a fish once caught. My data product solved the problem by using multinomial logistic regression to predict the species of a given fish based on user input. The software is extremely easy to use. After the end user has entered the six required, distinct measurements of their fish, the data product tells them the predicted species with high accuracy.

Datasets

The values from the raw data were not changed. I did, however, import the data into three pandas data frames. The steps for this were:

1) Found the fish dataset on kaggle.com. Here is the source: https://www.kaggle.com/datasets/aungpyaeap/fish-market
2) I copied the dataset.
3) I created a blank file on Github.
4) I pasted the dataset onto the blank file.
5) I cut the header row of data and saved the file.
6) In IntelliJ, I used pandas to create a data frame from the url of my GitHub. Then I pasted the cut header row as names for the data frame. This way the header is not considered data by the data frame. This makes the data frame easier to work with.
7) I created a second data frame for the column of species of fish. These are the dependent variables.
8) I created a third data frame for the remaining 6 columns of fish data. These are the independent variables.

Examples:

Species	Weight (in grams)	Vertical Length in CM	Diagonal Length in CM	Cross Length in CM	Height in CM	Diagonal Width in CM
Bream	242	23.2	25.4	30	11.52	4.02
Smelt	7.5	10	10.5	11.6	1.972	1.16
Pike	456	40	42.5	45.5	7.28	4.3225

I have hosted the dataset here: https://raw.githubusercontent.com/FallicoFunctions/Fish_Dataset/main/Fish%20Dataset

Data Product Code

The data values from the original dataset were not changed, however the data was used to create three pandas data frames. The first data frame held all of the original data, the second data frame held only the dependent variables, and the third data frame held the independent variables. The original data frame was not used after the second and third ones were created. The two data frames for the dependent and independent variables were used for training and testing.

It was necessary to have one data frame containing the dependent variables and one data frame containing the independent variables. That is because the multinomial logistic regression model needed them split up for training, predicting, and determining accuracy.

For the non-descriptive portion, I used a multinomial logistic regression model. This model is a built-in function of the sklearn library. The code to setup this model is: linear_model.LogisticRegression(). This model works well when there are multiple independent variables used to determine one dependent variable. In the case of this project, there were six independent variables, the six different measurements of a fish, used to determine one dependent variable, the species of the fish.

Prior to training the model, the dependent variables were stored in their own pandas data frame, and the dependent variables were stored in their own pandas data frame. Then, the data is sent through the train, test, split sklearn function. The code for this is train_test_split(). I used a 30% test size. This function splits the data into the training data and the testing data. The testing data is used to determine the accuracy after the model has been trained.

The model is trained using the fit function from the sklearn library. The code for this function is fit(). After the Python code is run, the predictive model is considered trained. After the data was trained I needed to obtain predictions from the model. To obtain the predictions from the model, I needed to assign the predictions to a variable. This is done using the predict function from the sklearn library. The code for this is predict().

After all this had been completed, I was able to print the accuracy score. I was able to do so using the accuracy score function from the sklearn library. The code for this is accuracy_score(). When printed, the score is shown as a decimal, i.e .912233.

At this point, the model is considered trained and tested. It is tested because I partitioned 30% of the original dataset for the purpose of testing. Aside from that, I also reran the program 30 times to ensure it worked as expected, and that the accuracy score was over 80% each time. The program was successful.

Objective (or Hypothesis) Verification

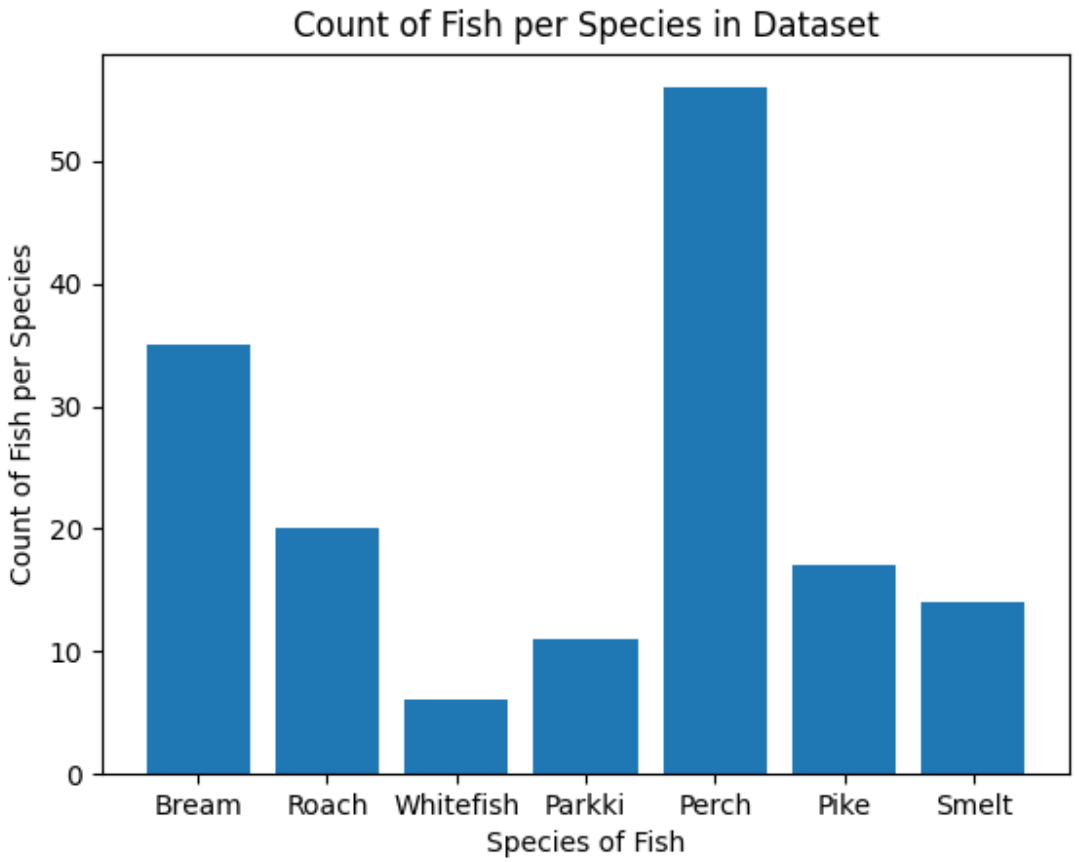
The project's objective was to create a machine learning data product that can predict a species of fish based on inputs of said fish by the end user. Another part of the goal for the objective was to be consistently over 80% accurate.

The objective was met because I created a data product that predicts a species of fish based on inputs entered by the end user. The accuracy of the data product was consistently over 80%.

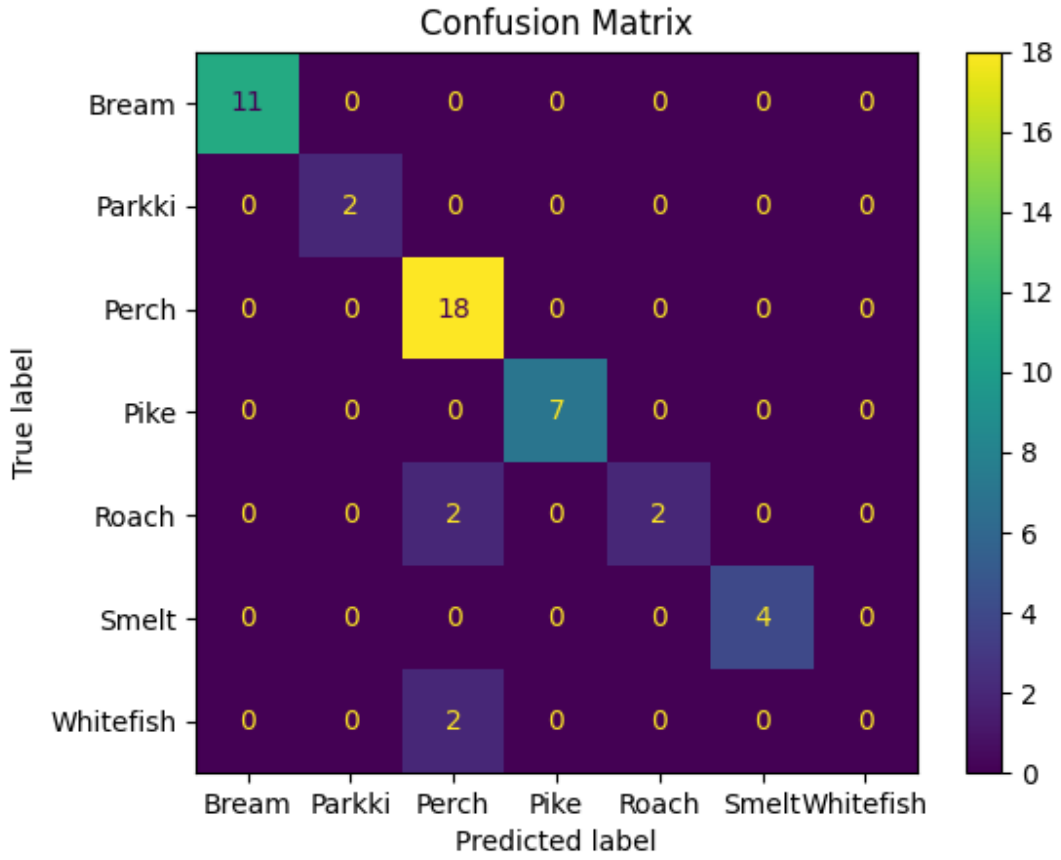
Effective Visualization and Reporting

After I settled on the dataset I decided to use, I manually checked all of the data. It was not difficult to do considering there were 159 lines of data. I ensured that for every dependent variable (the species of fish), there were six dependent variables (the six measurements of the fish). I concluded the dataset was complete and valuable for my project objective.

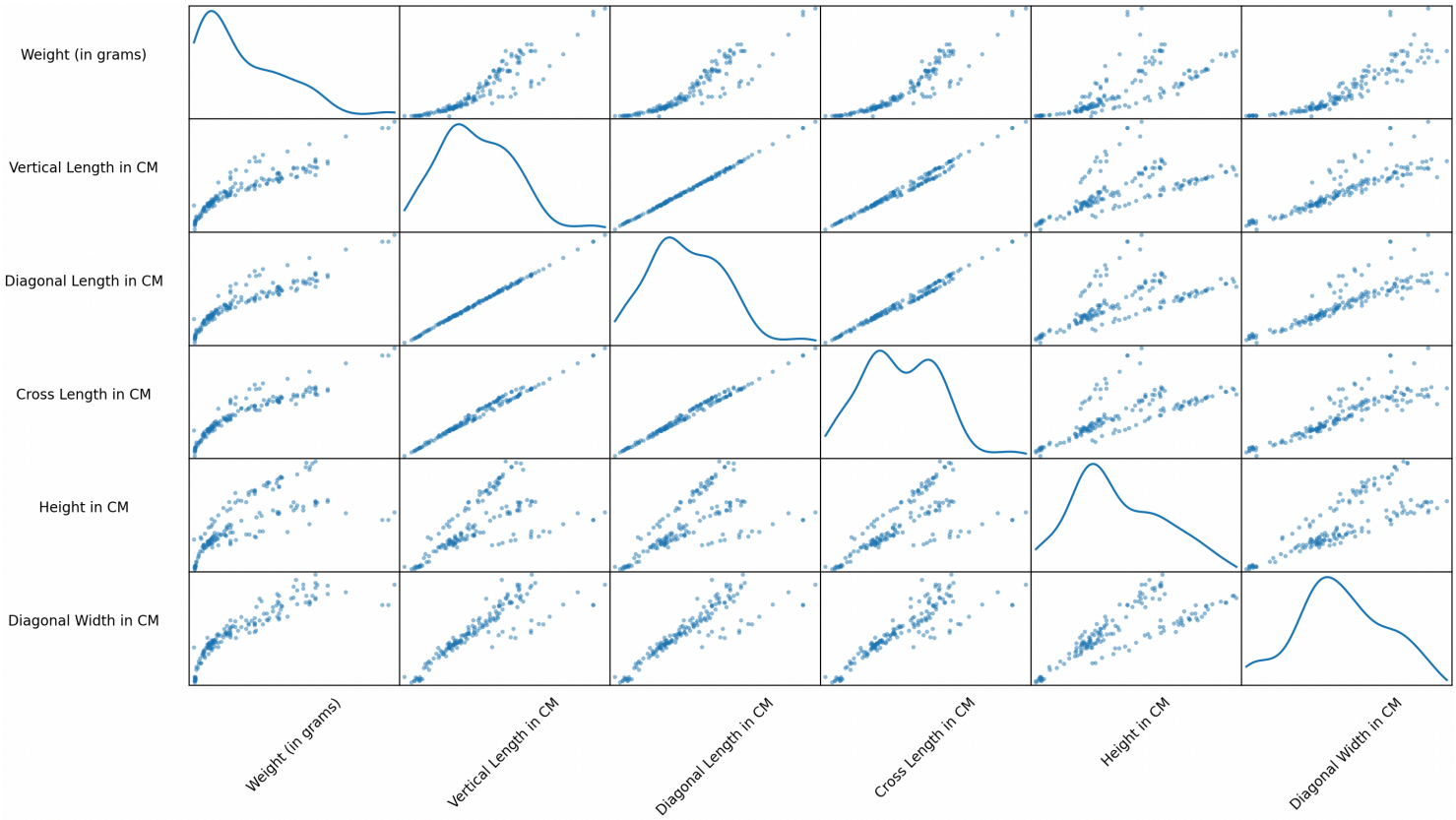
In the image below, there is a bar chart representing the seven different species of fish in the dataset, and the amount of individual fish data for each species. This helped me determine the breakdown of the number of fish by species. This was beneficial to know in case the prediction model skewed toward any one species, or if one species was not being predicted correctly enough.



The next graphic I have is a confusion matrix. The confusion matrix was useful in helping me determine how well my multinomial logistic regression model predicted. When a confusion matrix shows all values in a diagonal line from the top left to the bottom right, the prediction model is considered highly accurate. This confusion matrix shows 48 predictions. Four of them are incorrect. Twice the model predicted Perch when the fish species was Whitefish. Also twice, the model predicted Perch when the fish species was Roach. In this confusion matrix we can see that the prediction model never guessed Whitefish as the species. Looking back at the bar chart of number of fish per species, we can see there are less than 10 data entries for classifying Whitefish. In a future project, I would ensure the gathering of more data for the categories that are lacking. An even amount of data for each species of fish would ensure a greater accuracy.



Lastly, the below image is a scatter matrix. When reviewed, you can see a correlation amongst the measurements in a positive manner. That is to say in layman's terms, all measurements of a fish will be larger as the size of a fish increases. This seems fairly obvious; one would not expect a fish with low weight and length to have a fin larger than a fish with a larger weight and length.



Accuracy Analysis

30% of the original dataset was partitioned for testing purposes. I had made the decision that the program would be successful if the accuracy was over 80%. I ran the program 30 times to ensure accuracy. Each time I ran the program, the accuracy score was over 90%.

I used metrics.accuracy_score() from the sklearn library to gauge the accuracy of my machine learning data product. I made a personal goal of having an accuracy score of 80% or high. On the first try, the accuracy was 97%. Each try after that, the score has always been above 90%.

Application Testing

Luckily, testing was very easy. I wanted the accuracy to be over 80%. On the first try, the accuracy was 97%. To confirm this, I continued to test the prediction model. Each time I ran the prediction model the results of accuracy are over 90%. I ran the model 30 times. It is because of this that I had no need to modify my program.

Application Files

In the zip file for my submission, there is a folder name, "C964 Machine Learning Capstone Nicholas Fallico". Keep this folder in a place you will remember. This folder has all the files needed to run in IntelliJ.

You may run this Python program in the IDE of your choice but my instructions will be specific to IntelliJ IDEA CE.

Within IntelliJ you will need to install Python, the pandas library, and the sklearn library. Follow the user guide to do this.

User Guide

USER GUIDE:

A. Installing IntelliJ CE (Skip to section B if you have this IDE or wish to use your own)

11. Click the link to go to IntelliJ's website to download the product. Because my project uses a Mac I will give instructions as if you are using a Mac. The instructions will be similar enough if you wish to use a Windows machine. https://www.jetbrains.com/idea/download/#section=mac

12. On the right side of the page there is black button to download the Community version. On the right side there is a dropdown arrow. Selected this and it will show

you two options: Intel or Apple Silicon. Select the one that describes your CPU. The download will start automatically. It will automatically go to your downloads folder.

13. Double click the .dmg file. A new window will pop up.
14. Drag the IntelliJ IDEA CE icon onto the Applications folder icon. It may ask for an admin password. Have that ready in case.
15. IntelliJ IDEA CE is now installed with the latest version. As of 10/29/22 this will work for my data product.
16. Eject the IntelliJ .dmg file by dragging it to the trash.

B. Opening and Setting up the Project in IntelliJ

17. After opening my submitted .zip file you should have a folder titled, "C964 Machine Learning Capstone Nicholas Fallico".
18. Open IntelliJ. You should see a window with projects to choose from. You should also see a button that says Open on the top right.
19. Select the Open button.
20. Navigate to where you have my folder located.
21. Click once on the folder, "C964 Machine Learning Capstone Nicholas Fallico". You want to highlight it but not open it.
22. Once the folder is selected, click on the Open button on the bottom right of the finder window.
23. IntelliJ should open the project and display the Main.py file.
24. The program is complete. You do not need to do anything to the code.
25. You will probably see some errors in the import section at the top of the page. It likely says you don't have Python, pandas, or sklearn installed.
26. You should see a message on the top right of the window asking to install Python. Click yes.
    1. If you do not see this message follow these steps to install Python.
    2. Selected IntelliJ IDEA in the top right in the toolbar.
    3. From there, navigate to, "Preferences…" Once selected a new window will appear.

4. There is a list of options on the left side in the new window. There is an option that says, “Plugins.” Select this option.
5. From there you can search for Python in the newly appeared search bar. You will find a search result, “Python Community Edition.” Select Install.
6. After you install, click Apply on the bottom right of the window. IntelliJ may ask you to restart. Do so.

27. If you hover over the imports a message will pop up asking if you want to download the libraries. Click yes.
    1. If you do not see these options to install the pandas library or the sklearn library, follow the previous steps to install Python to install the pandas and sklearn plugins.

C. Using the program

1. You will not need to do anything with the code. It is complete.
2. When you run the program it will ask you to input the six distinct measurements as independent variables. Then it will give a text output of the name of species the logistic regression model predicts.
3. To run the program click on the green hammer button at the top of the window, right of center.
4. Then click the green play button to the right of the hammer.
5. In the run module in your IDE you will see the welcome statement and directions to input your six measurements.
6. The instructions will dictate to enter your six measurements on one line, each separated by a space.
7. Once you have entered your six measurements hit the enter key on your keyboard.
8. The program will then give you a text output of the name of the species it has predicted.
9. The program has now run and completed its task. You may exit the IDE.
10. If you wish to run the program again, refollow these steps under letter C.

Summation of Learning Experience

The WGU class Introduction to Artificial Intelligence – C951 prepared me for this project. The written report for that class is similar to the one for this capstone project. In addition to that class, the webinars by Jim Ashe were extremely helpful with the coding portion of this capstone. This experience showed me that machine learning is a lot simpler than I assumed it to be. I would not feel worried if I had to do create another machine learning prediction model.

References

Common Freshwater Fish of Illinois. (2003) Retrieved from https://www2.illinois.gov/dnr/publications/documents/00000317.pdf