Group Mini-Project

Part of your final assessment in PIC16A is via a group mini-project. In this project, you and your partners will perform and communicate about a complete data analysis of a real data set.

Timeline

You’ll have three scheduled discussion periods in Weeks 7 and 8 to work on this project (see the Schedule). In total, the project is intended to require approximately 10 hours of effort from each of you (that is, ~30 total person-hours for a three-person group, or ~20 total hours for a two-person group). You are expected to coordinate with each other to schedule additional meetings or delegate responsibilities as appropriate to complete the project. Projects are due on the last day of instruction, Friday, 12/11, and are submitted as a group.

There will also be homework assignments due during this time. Some of the homework problems will be specifically related to project tasks, and you are especially encouraged to work on these problems with your group members.

Your primary resource in this project should be your group members. However, if you are all stuck on a particular aspect, you are welcome to ask for help from your peers on Campuswire.

Project Components

Your project should be contained and submitted in a single Jupyter notebook. We will give you a template. You’ll perform components of the project in different notebooks over the course of the scheduled Discussion periods and homework assignments; your group is responsible for moving the required code into your final project file.

Your project should contain the following sections:

Group Contributions Statement: Briefly describe which group members contributed to which parts of the project (parts listed below). Here is an example of an acceptable Group Contributions statement:

“All three of us wrote the data acquisition and preparation, Xenith led the exploratory analysis, Rodrigo and Essun led on modeling, all three of us wrote the discussion.”

Exploratory Analysis: compute summary statistics and construct visualizations about the relationships between variables. You should explain how your analysis supports your modeling decisions below. Your exploratory analysis should include at least 3 figures and at least 1 displayed table. May include data loading, preparation and cleaning. 2-person groups need include only 2 figures and 1 table.

Modeling: deploy at least three machine learning models and evaluate their performance. Must include (cross)-validation and evaluation on unseen testing data. May include feature engineering. In this section, you should show how you systematically select your features to achieve optimal predictive accuracy. As part of your model evaluation, you are required to use your model(s) to make predictions on unseen (testing) data. You are required to use multinomial logistic regression and decision tree models. You must also select one additional model to deploy. Possibilities include but are not limited to random forests; support vector machines; nearest-neighbor classifiers; and neural networks. You will need to learn how to import and use the corresponding model. In your submission, you should describe briefly, for each model, how it works and why it is suitable for the task at hand. Your explanations don't need to be detailed – give some intuition for someone who has never seen that model before. 2-person groups need include and discuss only the two required models.

Discussion: describe the performance of your models, and state which combination of model and measurements you recommend. Discuss how the model could be improved if more or different data were available. Describe possible dangers associated with interpreting or using the model.

Project Evaluation

The project is worth 10% of your final grade, and will be graded out of 10 points. All members of the team will receive the same project grade, unless I either (a) receive private communication from a team-member that raises concerns about the team's collaboration or (b) the contributions statement suggests a highly inequitable division of labor.

Here's what I'm looking for:

- 2 point for a clear and equitable Group Contributions statement.
- 3 points for Exploratory Analysis: A helpful exploratory analysis, including at least three figures and at least one table that support your decisions in the Modeling section. 2-person groups need include only 2 figures and 1 table.
- 3 points for Modeling: A full modeling pipeline, in which you deploy three or more machine learning models, describe how they work at an intuitive level, train them on data, and systematically evaluate their performance. Your evaluation methods should include both cross-validation to select a candidate model and evaluation on unseen testing data to evaluate its overall predictive power. 2-person groups need include only the two required models.
- 2 points for Writing: Clear expository writing of the analysis, including narrative text walking the reader through the analysis; description of how the models work; docstrings and comments; and a critical discussion in the final section. I will have slightly lower expectations for the writing component for 2-person groups, but not much lower.

Your work on the project during scheduled discussion sections will additionally count toward your participation grade, as in any other discussion section.

Project Task

An important task in the ecology of the Antarctic is to catalog the many different species of penguins in that area. Determining the species of a penguin often requires a combination of biological expertise and many precise measurements, which can be difficult to obtain.

In the dystopian future, there are too many penguins:

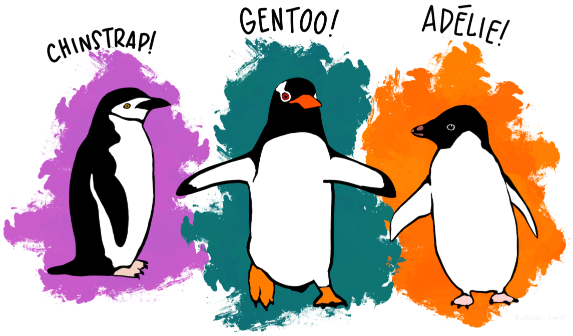
Too Many Penguins!.

Because there are so many, we can’t take many detailed measurements on all of them! In order to classify the species of penguins in large volume, we need to figure out which measurements are most important for distinguishing penguin species.

Your goal in this mini project is to determine a small set of measurements that are highly predictive of a penguin’s species. Options include the island on which the penguin was encountered, the length and depth of the culmen (bill), the length of the flipper, the body mass, and the sex of the penguin. That is, you should systematically determine which of these features is most predictive of the penguin’s species, using at least three distinct machine learning algorithms and evaluating the results. You are required to test decision trees and multinomial logistic regression. You may optionally use any additional algorithms that you would like.

Part of the point of this project is to use a small number of measurements (i.e. columns in the data). Submissions that use a large number of measurements may receive only partial credit, even if they achieve very good predictive accuracy.

For training and evaluating your models, we will use the Palmer Penguins data set. The Palmer Penguins data set was collected by collected by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network. The CSV data contains measurements on three penguin species: Chinstrap, Gentoo, and Adelie.



Three stylized penguins, one each of the species Adelie, Gentoo, and Chinstrap, with labels above their heads and patches of color behind them. Illustrations of the penguin species in the Palmer Penguins data set, by Allison Horst.

In data science, it is important to develop a thorough understanding of the data that you are attempting to model. I recommend consulting *these* (https://www.youtube.com/watch?v=ZbASA6fZaRI) *helpful* (https://www.youtube.com/watch?v=M5UlTRrVaTk) *videos* (https://www.youtube.com/watch?v=RoTVc32TLx8) to build your knowledge of penguin behaviors outside their natural habitats.

An Example of Good Data Writing

Here is a *very nice analysis* (https://humansofdata.atlan.com/2016/07/machine-learning-python/) of the famous Titanic survivors data that exemplifies the stages of data analysis and good expository writing. Your projects are not required to be as detailed as shown here, but this analysis would be an excellent place from which to draw inspiration. Please notice:

Explanations: The ratio of words to code and figures is high.

Iteration: The writers often start with one machine learning model, and then improve it (or compare it to competitors) as they go.

Discussion: The author critically considers what has been learned from the model; where the model might fail; and what this implies about the data.

You can also steal some good tricks here with pandas, matplotlib, and scikit-learn!

In [ ]: