PreMiEr
Precision Microbiome Engineering
NSF Engineering Research Center

The Hitchhiker's Guide to Data Science

Overview
This is an exploratory activity where students use AI to build a machine learning model. Students will preprocess data, apply regression analysis, and visualize their results using Python. This lesson integrates concepts from statistics, programming, and data science to provide hands-on experience with machine learning techniques.

Key Search Words
Statistics, Python, Programming, Data Science, Data Analysis, Machine Learning, Mathematics, Biology, Bioinformatics, Regression Analysis, Data Visualization

Learning Objectives
Activity Learning Outcomes:
- Preprocessing Data:
  - Learners will prepare data for analysis by locating and filling missing data values.
- Regression Analysis:
  - Learners will construct a correlation heatmap and interpret correlation coefficients.
  - Learners will compare pairwise relationships and perform linear regressions.
- Visualization:
  - Learners will construct plots using Python visualization libraries (e.g., seaborn, matplotlib).

Curriculum Alignment
Statistical Methods I (MAT-152) Learning Objectives:
- Organize, display, calculate, and interpret descriptive statistics.
- Perform regression analysis.

Classroom time required
Prior to Lesson (night before) 30 minutes:
- Learners should create an OpenAI account (openai.com).
- Visit Kaggle.com and explore datasets of interest.
- (Optional) Upload a dataset to GPT3.5 and ask it for analysis.
- (Optional) Create a google account specifically for this activity.
- Explore google collab by asking GPT3.5 to write you a "Hello World" script for collab.

Classroom Time Required 1 – 1.5 hours:
- Engage: 5 – 10 minutes
- Explore: 30 – 45 minutes
- Explain: 10 minutes
- Elaborate: 20 – 25 minutes
- Evaluate: 10 minutes

Post Activity Assessment 5 – 10 minutes:
- Low-stakes Quiz
- Class Presentation
- Written Reflection
- Discussion Posts

Materials & Technology
- Computers with internet access and
- Google, OpenAI Accounts (Kaggle optional)
- Projector for in class demonstrations

Safety

Facilitators should be aware of potential privacy considerations when utilizing generative AI in the classroom. UC Berkeley lists these prohibited uses:

"At present, any use of generative AI tools should be with the assumption that no personal, confidential, proprietary, or otherwise sensitive information may be entered into models or prompts."

Students should avoid use of non-public datasets for this activity.

Teacher Preparation for Activity

Facilitators should take 30 – 45 minutes, prior to the activity, to familiarize themselves with prompts and with the data analyst capabilities of generative AI (GPT, Gemini, etc.).

Test Run:

- Create an OpenAI account and select "Explore GPTs". Scroll down the page and select Data Analyst – by ChatGPT. Note: This lesson was finalized on 7/17/24.
- Select a dataset from Kaggle.com (or use a built-in python dataset) and share it with Data Analyst GPT.
- Try this prompt: "Provide an overview of the data."
- Followed by: "Please guide me through any analysis a data scientist would conduct before implementing machine learning. Pause after each step so that I may ask any questions necessary for my understanding."
- Create a google account (if desired) and launch google collab.
- Copy the code generated by Data Analyst GPT into google collab.
- Use this article to upload your data into collab:
  - https://www.geeksforgeeks.org/ways-to-import-csv-files-in-google-colab/
- Explore python visualizations with assistance from generative AI.

Helpful Hints:

- Never hesitate to reinitialize the chat and clear all memory.
- Your first prompt can make or break your experience (especially with free/limited accounts)
- Request for code comments and an explanation for steps taken.

Experienced Python Users:

- You may opt to run/modify any code generated locally, but the activity was designed to remove any requirement for local disk space.

New Python Users:

- Prior to the test run detailed above, create a "Hello World" script in collab using this tutorial:
  - https://www.learnpython.org/en/Hello,_World!

Student Preparation for Activity

Before class:

- Learners should create an OpenAI account (openai.com).
- Visit Kaggle.com and explore datasets of interest.
- (Optional) Upload a dataset to GPT3.5 and ask it for analysis.
- (Optional) Create a google account specifically for this activity.
- Explore google collab by asking GPT3.5 to write you a "Hello World" script for collab.

Practical Tips for Students:

Anonymize Data: If you need to discuss sensitive topics, try to anonymize the data as much as possible.
Review Outputs: Always review the outputs generated by AI tools critically and verify important information from reliable sources.
Limit Usage: Use AI tools as a supplement to your learning or work, not as a replacement for critical thinking and personal judgment.

Procedure

Engage: 5 – 10 minutes

- Begin the lesson with one (or both) videos:
  - *"Why is the gut microbiome important?"* | https://www.youtube.com/watch?v=AsyzqhFKLoI
  - *"How can we solve the antibiotic resistance crisis?"* | https://www.youtube.com/watch?v=ZvhFeGEDFC8
- If audio is unavailable, engage students with this article:

- Why Choose a Data Science Career in 2024 | https://www.ptechpartners.com/2024/01/17/why-choose-a-data-science-career-in-2024/

- Discussion Questions
  - What are some reasons you would or wouldn't pursue a career in data science?
  - Are you concerned about antibiotic resistance?
  - Do you know anyone personally who suffers from autoimmune disorders?
  - What impacts would a major outbreak of a resistant bacteria have on everyday life?
  - How do you think AI can help solve the problem(s) of 1) data scientists, 2) antibiotic resistance etc.

  Note: A discussion forum can be used for an asynchronous online class format.

Explore: 30 – 45 minutes

- Introduce python by having students launch a web browser and navigate to: https://colab.research.google.com/
  - Have students code the following line: *print("Hello World!")*
  - Instruct students to select the "run" button (looks like a stereo play button) or in the Runtime menu select Run All.
    - If any errors are encountered, allow other students to offer suggestions for resolution. Intervene if problem persists.
- Students should create a new tab in their browser and navigate to: https://chatgpt.com/
  - Login with the previously created account (or create an account).
  - Start a new chat and enter the following prompt: "Write a simple python script that prints Hello World!"
  - Copy python code generated and paste underneath previously written code (if no errors are present).
  - Run the code which should generate two "Hello World!" statements.

- Introduce students to the UrbanRuralChina dataset
  - The study demonstrates the relationship between urbanization and impacts on the microbiome in both China and the western world. This basically means scientists can sample a person's microbiome and likely determine if that person lives in an urban or rural setting.
  - Have students select, "Explore GPTs" (in the ChatGPT tab) and select Data Analyst by ChatGPT.
  - Provide students with their objectives:
    - Load Data into google collab:

```
url = 'https://raw.githubusercontent.com/kwinglee/UrbanRuralChina/
51dd745ff22c335698343e2167644bea3a37b3b5/16SrRNA/inputData/
abundantOTU/abundantOTULogNormal.txt'
data = pd.read_csv(url, delimiter="\t") (Note: data and df are used by GPT)
```

    - Locate missing data (if any). Print the number of missing values:

```
# Check for missing values
missing_values = data.isnull().sum()
print(missing_values)
```

    - Construct a Correlation Heat Map.

```
# Calculate the correlation matrix (ex. df.corr())
corr_matrix = data.iloc[:, 2:].corr() # must slice out categorical
# Utilize seaborn to create the heat map
plt.figure(figsize=(12, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show() (Note: A local environment may be required or premium GPT)
```

    - Compare pairwise relationships using linear regressions.

```
# Select two highly correlated variables (example: column '1' and '10')
x = data['1'].values.reshape(-1, 1)
y = data['10'].values.reshape(-1, 1)
# Perform linear regression
regressor = LinearRegression()
regressor.fit(x, y) (Note: Students can use heat map to select variables)
```

    - Plot the regression line on a scatter plot.

```
plt.scatter(x, y, color='blue')
plt.plot(x, regressor.predict(x), color='red')
plt.xlabel('Variable 1')
plt.ylabel('Variable 10')
plt.title('Linear Regression between Variable 1 and Variable 10')
plt.show()
```

Explain: 10 minutes

Discussion Questions

- *"How many missing values did you (all) encounter when preprocessing data?"*
- *"Why do you think it is important to check for missing values?"*
- *"Any struggles loading data?"*
- *"What prompts worked well for you?"*
- *"What prompts required modification?"*
- *"Did you encounter any python errors? How did you resolve them?"*
- *"Has anyone written python scripts before today's activity?"*
- *"What are some ideas you have for data analysis with AI?"*

Elaborate: 20 – 25 minutes

- Present students with the final goal of the activity: **use AI to build a machine learning model that classifies the samples as either urban or rural. Generate a confusion matrix and plot the ROC curves.**

Helpful prompts:

- "What are the basic steps required to create a random forest machine learning model to classify urban/rural samples in this data set and plot ROC curves?"
- "How can I model this data with machine learning?"
- "Can machine learning be used to classify this data?"
- "I'm looking to implement a random forest machine learning algorithm to classify samples as urban or rural. Can you plot a ROC curve using the confusion matrix?"
- "I have a text file on GitHub and am interested in machine learning. I've heard of random forest models and wondered if you would help guide me through the process?"

Helpful code snippets:

- Read-in Data to Collab Directly from GitHub

```
# Step 1: Define the GitHub raw URL
url =
'https://raw.githubusercontent.com/kwinglee/UrbanRuralChina/51dd745ff22c335698343e2
167644bea3a37b3b5/16SrRNA/inputData/abundantOTU/abundantOTULogNormal.txt'

# Step 2: Read the file into a DataFrame
china = pd.read_csv(url, delimiter="\t")
```

- Python Library Installation

```
# Step 1: Install necessary libraries (if not already installed)
!pip install pandas scikit-learn matplotlib seaborn io
```

- Adjust window on ROC curve plot

```
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([-0.005, 1]) # Modify these two lines
plt.ylim([0, 1.01])   # in order to adjust window
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
```

Note: Students may encounter computational limits if using a free account. If students are in groups of 2 or 3, this can be mitigated by sharing the computational load.

Evaluate: 10 minutes

This is another great time to have a discussion or have student write a reflection. Some things to note:

- It's not unlikely that ChatGPT goes "rogue" and doesn't provide an experience that leads to a workable model.
- Students may produce the visualizations but struggle to interpret them and don't have the skills to reproduce the code without assistance.
- Computational limits on free accounts are imminent with this activity. You can likely avoid by skipping a correlation matrix.
- How do we know that the response from AI is correct? What problems could that pose for programmers or researchers looking to integrate generative AI in their workflows?
- Generative AI can extend a person's capabilities but is not a substitute for skill mastery.

Differentiation

Differentiated Instruction:

- Offer additional materials or extension activities for advanced learners to delve deeper into the topics covered in the lesson. Good options for these are modifying plot names/colors, extending the activity by having students calculate other metrics like accuracy, precision, recall etc.
- Provide flexibility in the level of complexity for the hands-on activities, allowing students to choose the level that best suits their abilities.
- Advanced students can download a different dataset and attempt to analyze with generative AI.
- Create a small group of 5 – 7 students and lead the activity for those who would benefit.

Collaborative Learning:

- Structure activities that promote collaboration, such as group discussions or peer-to-peer assistance with the activity.
- Provide opportunities for students to share their challenges, insights and discoveries.

Scaffolding:

- Break down complex tasks, such as loading and manipulating the dataset in collab, into step-by-step instructions with clear guidance. Sample code is provided, but a paid subscription is recommended for the facilitator.
- Provide prompts or to assist students in organizing their analysis and interpretation of the microbiome data.
- Gradually release responsibility to students by allowing them to independently explore additional features or analyze other datasets.

Assessment/Check for Understanding

Assessment recommendations:

- Low-stakes Quiz
  - If using a quiz, tie the SLOs to the activity outcomes.
    - **Preprocessing**, why are missing data values a problem? (open ended)
    - **Regression**, what do the positive numbers mean on the correlation matrix? The negatives?
    - **Visualization**, what is a python library that is used for visualization?
- Class Presentation
  - Make the focus sharing the story for the exploration.
- Written Reflection
  - Have students document their experiences and how their perspectives changed (or didn't).
- Discussion Posts
  - Perfect for synchronous/asynchronous online courses but can be used in-person as well.

Required resources

Google Collab

- This solution allows learners to write and execute python code from the web browser. This resource is free but requires a google account (also free) to execute code. Collab stores python notebooks in the Google Drive associated with the account in use.
- https://colab.research.google.com/

OpenAI

- OpenAI houses ChatGPT, the generative AI used for this activity. GPT3.5 is free but requires an account to make uploads. There are computational limits for the unpaid version of ChatGPT. This activity was replicated on GPT3.5, but computation limits are reached quickly if you instruct GPT3.5 to execute code and display plots. Premium features are available with the paid version ($20) like unlimited use, access to advanced and beta versions. Can be used from the browser.
- https://openai.com/

GitHub (UrbanRuralChina Data)

- This activity builds on a previous project that utilized data to explore the impacts of urbanization on the gut biome. The UrbanRuralChina project is housed on GitHub and the dataset used for the activity can be imported into google collab from the browser or downloaded and imported locally.
- https://github.com/kwinglee/UrbanRuralChina
- https://raw.githubusercontent.com/kwinglee/UrbanRuralChina/master/16SrRNA/inputData/abundantOTU/abundantOTULogNormal.txt (raw link to data)

Kaggle (Recommended)

- Kaggle is a web-based platform and online community for data scientists and machine learning practitioners. There are many public datasets to download and explore. This resource is free but requires an account to download datasets. It may be possible to import datasets into collab directly from Kaggle, but that is not resolved in this activity.
- https://www.kaggle.com/

Sources

Papers:

*Winglee, K., Howard, A.G., Sha, W. et al. Recent urbanization in China is correlated with a Westernized microbiome encoding increased virulence and antibiotic resistance genes. Microbiome **5**, 121 (2017). https://doi.org/10.1186/s40168-017-0338-7*

Articles:

UC Berkeley Office of Ethics. (n.d.). Appropriate use of generative AI tools. https://oercs.berkeley.edu/privacy/privacy-resources/appropriate-use-generative-ai-tools

AWS Security Blog. (2023, July 12). Securing generative AI: Data, compliance, and privacy considerations. https://aws.amazon.com/blogs/security/securing-generative-ai-data-compliance-and-privacy-considerations/

PTech Partners. (2024, January 17). Why choose a data science career in 2024. Retrieved from https://www.ptechpartners.com/2024/01/17/why-choose-a-data-science-career-in-2024/

GeeksforGeeks. (2022, February 16). Ways to import CSV files in Google Colab. https://www.geeksforgeeks.org/ways-to-import-csv-files-in-google-colab/

Reiter, R. (n.d.). Hello, world!. Learn Python - Free Interactive Python Tutorial. https://www.learnpython.org/en/Hello,_World!

Videos:

Microbiology Bytes. (2018, September 3). Why is the gut microbiome important? [Video]. YouTube. https://www.youtube.com/watch?v=AsyzqhFKLoI

World Health Organization. (2020, October 10). How can we solve the antibiotic resistance crisis? [Video]. YouTube. https://www.youtube.com/watch?v=ZvhFeGEDFC8

Appendices