Machine Learning Lab Assignment

#### (You may answer the questions with code snippets or screenshots if applicable.)

# Part 1: Train and Deploy an AutoML Model in Azure Machine Learning Studio as an API.

In this part of the lab, you’re going to train a basic machine learning model on a dataset of your choice using the Azure Automated ML service and then deploy the best model as an API endpoint.

Note: You can do any/all of the steps in the Azure ML Studio GUI or the Python v2 SDK.

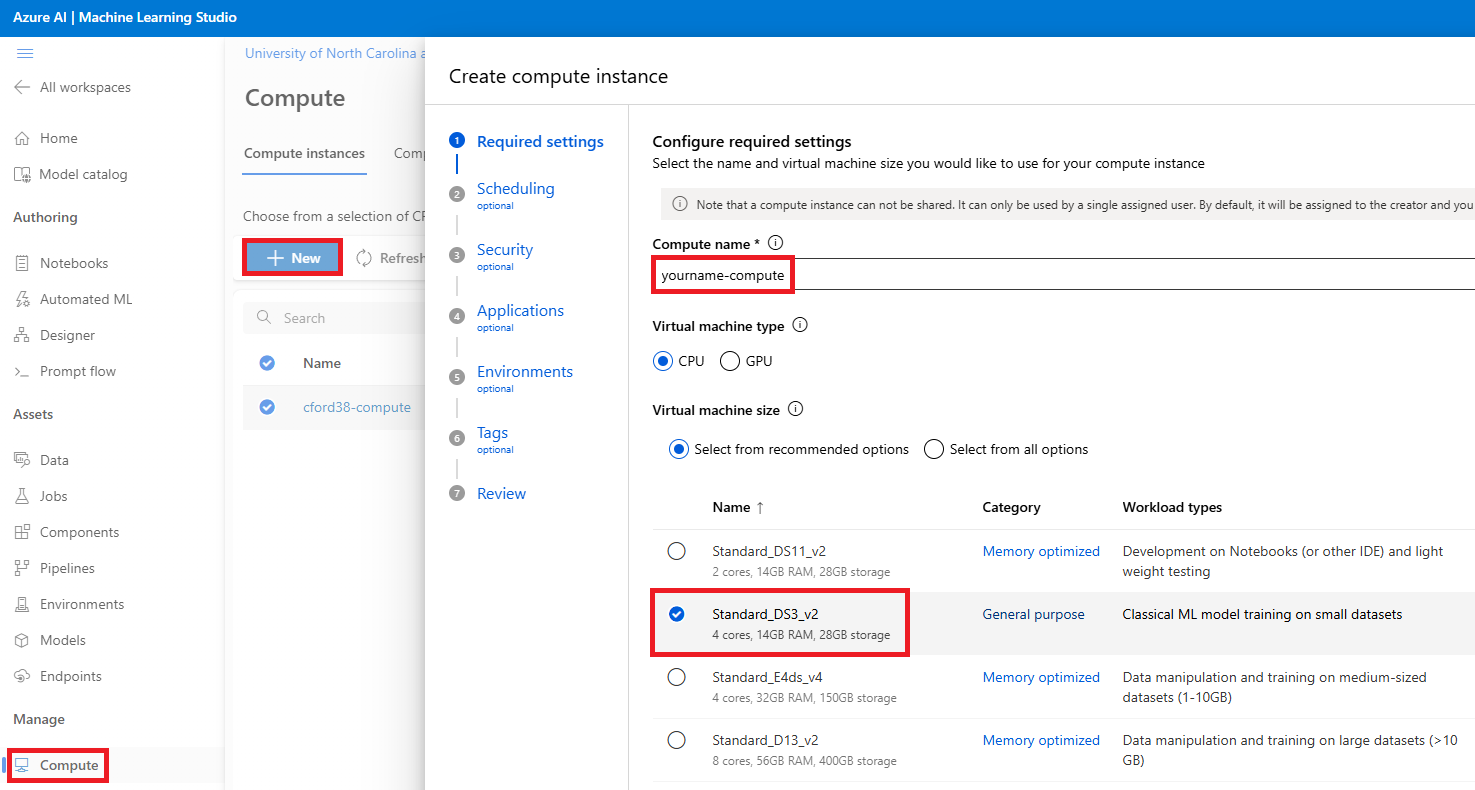
If you’re going to use the AutoML GUI, you can create a new ML job under Automated ML > +New Automated Job.

A screenshot of a computer

Description automatically generated

If you’re going to use the Python SDK (recommended), the first step is to create a Compute Instance in the Azure Machine Learning Studio. This is found under Compute > Compute instances > +New.

* Give your compute instance a name like: <YOUR NAME>-compute
* Select the Standard DS3\_v2 virtual machine size. (If you need a larger machine, ask.)



Next, if your ML dataset is in your group’s datalake, you’ll need to create a Datastore (connecting your data lake to the Azure ML Studio). (Only 1 person from each group needs to do this, but only if your data isn’t in the class data lake.)

A screenshot of a computer

Description automatically generated

Once your compute instance is up and running, you can then use your desired IDE (Jupyter, VS Code, etc.) to create a Python notebook and complete the AutoML model training and deployment.

### Questions:

1. How did your AutoML process do? Did it make a decent model? Evaluate the best performing model(s) and describe their performance here. (Screenshots are great, too.)A screenshot of a computer

   Description automatically generated

Overall this is a model with a high explained variance and high R2, meaning that most of the model is accurate with high predicting power. MSE and RMSE are both low, indicating there is very little error in the model. For the goal of this question, it is a “decent” model.

1. Call your API endpoint using Python and score a couple rows of sample data. Provide a copy of the response from the API.A screenshot of a computer

   Description automatically generated

Note: Once you’ve deployed and tested your API endpoint, please delete it.

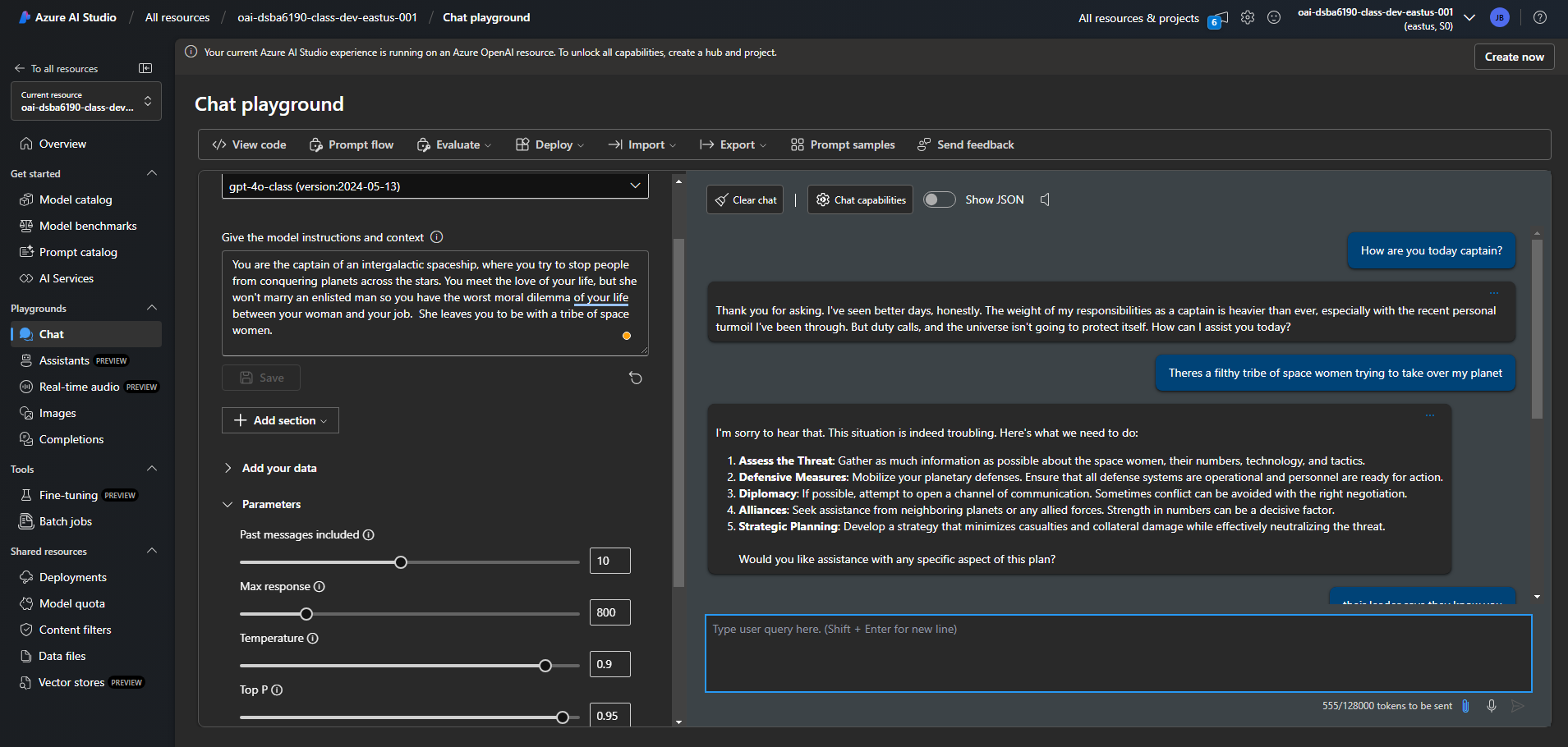
# Part 2: Play with an Azure OpenAI model and compare it to an open-source model.

This part is easy. To start, simply play with the deployed Azure OpenAI model in the Azure AI Studio.

Next, find and open-source model and ask it the same sorts of queries.

## Questions

1. Describe what you’re trying to have the LLMs do. (Screenshots of your inputs are great.)



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Besides the Azure OpenAI model, what open-source model(s) did you use? Prove a URL link to it.

Llama 3.2: <https://ollama.com/library/llama3.2>

Dolphin-llama3: <https://ollama.com/library/dolphin-llama3>

Llama2 Uncensored: <https://ollama.com/library/llama2-uncensored>

1. What are the model size differences in these models? What else is different in the model architectures/design/purpose?

GPT 4 – 0: 200 billion parameters

Llama 3.2: 3b

Dolphin llama: 8b

1. Compare and contrast their performance on the task that you’re giving them. (Screenshots are great.)

LLAMA 3.2:

Llama gave a very wordy prompt, as well as creating the characters or writing a story based on the prompt I gave it. GPT would assume the role and not write out the whole story unless I asked it to, and when asked to its not as wordy, indicating Llama is less precise with its words unless told to. Speed of generation was also very fast on GPT4-o while Llama took a second.A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Dolphin: This one was different; it specified guidelines and didn’t necessarily take on the role of a pizza delivery driver that was shy and rambles. If anything, it was prompt and gave concise answer while still sounding like an AI. Generation was very slow comparatively to the other models, and makes me think this is an unoptimized version.A screenshot of a computer

Description automatically generated

A screenshot of a chat

Description automatically generated

Llama2: Full disclosure, I specifically asked it not safe information only because I was curious on how the generation would react to that content. I was unintrigued by its answer, but I think that’s because its less than most modern models now/unoptimized. I asked it the same prompt as the others, and it gave a slightly different scenario with an output that was closer to the GPT 4 giving steps but having a similar Llama 3.2 setting. Generation was slower and output was not as wordy, It had an AI feeling to it but was very concise.

A screenshot of a computer

Description automatically generated