EQ2

1. Using iris.csv data

My code:  
import pandas as pd

#using the csv

filename = r"C:\iResearch project\iRI\_2025\_AS\Data\coding practice\iris.csv"

names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']

dataset = pd.read\_csv(filename, names=names)

1. Research question, hypothesis, and testing hypothesis using ML:
   1. Research question: Which machine learning model provides the most accurate, interpretable, and reliable classification of urinary biomarkers for distinguishing pancreatic cancer from benign disease and normal controls?
   2. Hypothesis: 1D-CNN + LSTM will be the most effective ML model to classify the urinary biomarkers because in a prior study, it achieved a 97% accuracy score using the same dataset that I am using [1].
   3. I will use several ML algorithms and compare their performance. One of these algorithms will be the 1D-CNN + LSTM. Some others will include Random Forests, XGBoosts, Neural Networks Linear SVM, Gaussian SVM, Decision Tree. First, I will pre-process the dataset. Then I will split the dataset into 80% training and 20% testing/validation. Then I will use each model type to make predictions, and I will look at their predicted accuracy scores. Then I will use these models on the validation set. Finally, I will evaluate the performance of the models and be able to determine whether 1D-CNN + LSTM is the most effective strategy.
2. About the Dataset:

Link: <https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer>

* 590 datapoints
* Patients are organized into 1 (**healthy**), 2 (**non-cancerous/benign pancreas condition**), and 3 (**pancreatic cancer**)
* **What does plasma\_ca19\_9 mean?** - CA 19-9 is a protein that functions as a tumor marker, meaning it's a substance that can be produced by cancer cells or by normal cells in response to cancer within the body
* **Normal:** <37; not normal: >37
* **Tracking 4 urinary biomarkers:** creatinine, LYVE1, REG1B, and TFF1.
* **Creatinine:** protein indicating kidney function
* **LYVE1:** lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
* **REG1B:** a protein that may be associated with pancreas regeneration
* **TFF1:** trefoil factor 1, which may be related to regeneration and repair of the urinary tract

Sources:

1. Mohamed Esmail Karar, Nawal El-Fishawy, & Marwa Radad. (2023). Automated

classification of urine biomarkers to diagnose pancreatic cancer using 1-D convolutional neural networks. Journal of Biological Engineering, 17(1). https://doi.org/10.1186/s13036-023-00340-0

**Essential Question #2 - AS**

*JD’s AI/ML Laboratory*

**Due: 07/06/2025, 11:59 PM EST**

**General Instruction**: Please **answer all questions** and provide relevant support from literature using **APA references**. You can answer the questions using bullet points with citations (preferably from peer-review journals). Keep in mind that most science competitions focus on participants having a very strong grasp of the literature. I would highly recommend printing out the relevant journal articles and do a deep dive. Using a highlighter and writing notes on the journal article can help you narrow down important content for future reference. I will be thorough in my review of your response and expect all the answers to be well thought-out and clearly explained with supporting literature.

1. **For your coding exercise, use the iris.csv data file from your Data folder and not a url.**

My code:  
import pandas as pd

#using the csv

filename = r"C:\iResearch project\iRI\_2025\_AS\Data\coding practice\iris.csv"

names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']

dataset = pd.read\_csv(filename, names=names)

1. **What is your research question, Hypothesis and how you are going to test the hypothesis using ML**
   1. **Research question: Which machine learning model provides the most accurate, interpretable, and reliable classification of urinary biomarkers for distinguishing pancreatic cancer from benign disease and normal controls?**
   2. Hypothesis: 1D-CNN + LSTM will be the most effective ML model to classify the urinary biomarkers because in a prior study, it achieved a 97% accuracy score using the same dataset that I am using [1].
   3. I will use several ML algorithms and compare their performance. One of these algorithms will be the 1D-CNN + LSTM. Some others will include Random Forests, XGBoosts, Neural Networks Linear SVM, Gaussian SVM, Decision Tree. First, I will pre-process the dataset. Then I will split the dataset into 80% training and 20% testing/validation. Then I will use each model type to make predictions, and I will look at their predicted accuracy scores. Then I will use these models on the validation set. Finally, I will evaluate the performance of the models and be able to determine whether 1D-CNN + LSTM is the most effective strategy.
2. **Describe the dataset you are going to use for answering the research question.**

Link: <https://www.kaggle.com/datasets/johnjdavisiv/urinary-biomarkers-for-pancreatic-cancer>

* 1. **Title: Urinary biomarkers for pancreatic cancer**
  2. 590 datapoints
  3. Patients are organized into 1 (**healthy**), 2 (**non-cancerous/benign pancreas condition**), and 3 (**pancreatic cancer**)
  4. **What does plasma\_ca19\_9 mean?** - CA 19-9 is a protein that functions as a tumor marker, meaning it's a substance that can be produced by cancer cells or by normal cells in response to cancer within the body
  5. **Normal:** <37; not normal: >37
  6. **Tracking 4 urinary biomarkers:** creatinine, LYVE1, REG1B, and TFF1.
  7. **Creatinine:** protein indicating kidney function
  8. **LYVE1:** lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
  9. **REG1B:** a protein that may be associated with pancreas regeneration
  10. **TFF1:** trefoil factor 1, which may be related to regeneration and repair of the urinary tract

1. **List the datasets you have reviewed?**

For this topic, I have only reviewed the Kaggle dataset titled “Urinary biomarkers for pancreatic cancer.”

1. **Perform EDA for your dataset, including data visualization and generate summary statistics. Also include distribution of age across the three classes. Explore if there are any missing values.**

(see code on VS)

1. **In discussion with your mentor, create a workflow for your analysis and document how you would generate the analysis set.**

**Objective:**

Evaluate and compare multiple machine learning models for classifying urinary biomarkers into three diagnostic groups: healthy (1), benign pancreatic disease (2), and pancreatic cancer (3).

**Dataset:**

Use the provided urinary biomarker dataset.

Features: age, plasma\_CA19\_9, creatinine, LYVE1, REG1B, TFF1, REG1A, etc.

Target variable: diagnosis (1, 2, 3).

Handle missing values and outliers.

**Preprocessing:**

Clean data (drop or impute missing values).

Encode categorical variables (e.g., sex).

Normalize or standardize numeric features.

Exploratory Data Analysis (EDA):

Summary statistics and distributions.

Visualize biomarker distributions by class.

Check class balance and correlations.

**Model Selection:**

Train and evaluate the following models:

Logistic Regression

Random Forest

XGBoost

Support Vector Machine (SVM)

k-Nearest Neighbors (kNN)

Neural Network (e.g., MLPClassifier)

**Evaluation Criteria:**

Accuracy

F1-score (especially weighted or macro)

Confusion matrix analysis

ROC-AUC (if applicable)

Model interpretability (e.g., SHAP values, feature importance)

**Cross-Validation:**

Use stratified k-fold cross-validation (e.g., k=5 or 10) to ensure robustness.

**Interpretability Tools:**

Feature importance (tree-based models)

SHAP or LIME for model-agnostic explanation

Result Comparison:

Summarize metrics across all models.

Rank models based on a balance of performance and interpretability.

**Conclusion:**

Identify the top-performing model.

Discuss trade-offs between accuracy and interpretability.

Highlight clinical implications and limitations.

1. **Optional for this week: Train a machine learning model that would help you answer your scientific question.**