Case Study Rubric

Case Study Title: *Tomato Trouble: Using Machine Learning to Combat Crop Disease*
Due: TBD
Submission Format: Upload PDF and link to GitHub repo on Canvas

Why am I doing this? This case study is an opportunity to bring together your skills in machine learning, model evaluation, computer vision, and data storytelling. More than just classification, your work could help lay the groundwork for accessible tech that supports sustainable agriculture.

What am I going to do?

At this point, you have likely accumulated a variety of technical and conceptual skills in data science. You will now have the opportunity to combine and express these skills in an independently-driven case study. You will ultimately provide a deliverable that covers all requirements, including significant results and conclusions. The deliverable will include: You will develop a machine learning pipeline to classify images of tomato leaf diseases. The PlantVillage dataset contains labeled images for several common diseases, and your job is to design a prototype tool capable of identifying them.

Your final submission must include:

- **Written Portion (PDF)** – outlining the problem, your approach, results, reflection, and references.
- **GitHub Repository** – containing all code, data, models, and visualizations necessary to replicate your work.

The Github repository for this case study can be found at: https://github.com/ytkidanu/PlantVillage_CS3-DS4002-SPR25. In addition, the original project's GitHub repository can be found at: https://github.com/ytkidanu/ImageProcessing. You will import data using TensorFlow and construct predictive models to classify plant diseases. Begin by exploring and preprocessing the image dataset. Evaluate whether class balancing is necessary and apply it if needed. Perform data augmentation to enhance model performance, except when training the Support Vector Machine (SVM) model, which should use the unaugmented data. Next, build and train your Convolutional Neural Network (CNN) models. Finally, evaluate model performance using a confusion matrix to identify where your model performs well and where it struggles.

How will I know I have succeeded?

You will meet expectations for this case study when you complete the following components according to the rubric below.

Category	Spec Details
Formatting	● Submit the written portion as a PDF file. ○ Plant_village_reflection.pdf ● Submit code created for all portions in a GitHub repository. ● The repository should be titled “CS-[insert first & last name]”. ● Scripts: All scripts used in the analysis. ● Outputs: All visualizations and result files. ● Data: ○ Initial: Raw dataset. ○ Final: Cleaned or transformed dataset. ○ Prepare a data appendix ● README.md: Overview of the project and how to run the code. ● LICENSE.md: Licensing information ● REFERENCES.md: All external references used, in IEEE citation format
Written Portion	● In a small paragraph, summarize the problem presented in the study as well as its importance. ● In a small paragraph, discuss your plan to meet the demands of the deliverable. ○ Include a simple graphic that outlines your analysis plan. Finally, discuss the results of your study in complete sentences as well as the significance of these results in the greater context. ○ Justify why these steps were taken ○ Provide a reflection on the entire process ■ Evaluate how you performed ■ Provide a suggestion on how to improve the case study
Code	● Include preprocessing for the PlantVillage tomato image dataset (resizing, normalization, augmentation). ● Implement and train four models: ResNet50, VGG19, InceptionV3 (transfer learning), and Support Vector Classifier (SVC). ● Evaluate each model using: Accuracy, Precision, Recall, F1-score, and Confusion Matrix. ○ Include visualizations: training/validation plots and sample image predictions. Write clean, organized, and well-commented code. ● Visualizations should go in outputs folder ● All code should go in the Scripts folder
References	- include all external references in IEEE citation style. - Place references on a separate page at the end of the PDF. Cite any tools, datasets, libraries, or external content.