Assignment Cover Sheet

To be completed **electronically** by the student and submitted with each piece of work. Please upload this completed cover sheet via Turnitin

Assignment Title: Experiment Portfolio – 40%

Tutor: Usman Ahmad

Student Name: **Naomi Chellsea Espiritu**

Student Number: **639876**

Date of Submission: **30th October 2025**

Details of your submission:

This is my Assessment 1 project, which contains two Kaggle links to my notebooks (one for regression and one for classification tasks), along with one Kaggle certificate.

Github Link : https://github.com/naomichellsea/Assessment-1---Machine-Learning

In submitting this assignment, I am confirming that I have read and understood the regulations for assessment, and I am aware of the seriousness with which the University regards unfair practice. Please see the universities Unfair Practice Policy for details.

Signed: 

Date: **30th October 2025**

Title of Experiment 1:

In bike rental prediction exercise 1, I am satisfied about its final outcome. The whole aim was to predict the total daily rentals. Achieving a test R^2 score of 0.871 confirms that the developed model captures a significant amount of underlying variance in rental demand.

Strength:

One of the projects most successful parts was the deliberate feature engineering. Recognizing the critical influence of time and extracting both the year and month from the date column says a lot. The subsequent analysis confirmed the year columns' powerful predictive utility. Furthermore, the inclusion of the Ridge Regression model was a powerful choice for me as it established a robust linear baseline of R^2 = 0.826, allowing for a quantifiable measure of the performance gain achieved by more complex non-linear Decision Tree model.

The evaluation process also provided helpful insights, specifically through the residual plot analysis. This confirmed that the model's errors are generally unbiased. They cluster well around zero. However, it also highlighted a subtle systematic weakness, the model consistently displays a tendency to under predict demand during the absolute highest peak rental days. This indicates that an area where the model needs to be more sensitive to extreme values.

Limitations:

The learning curve in this exercise involved diagnosing the overfitting tendency in the Decision Tree. Although the test R^2 is strong, the clear gap between the high training performance and the testing score signals a potential breakdown. Despite implementing a maximum depth constraint (max_depth = 10), the model retained the capacity to memorize patterns.

To improve effectively, the logical next step is implementing dedicated hyperparameter tuning. A structured approach such as using Grid or Randomized Search, would be necessary to systematically find optimal values for parameters such as min_samples_leaf.

The strategic goal in this is to enforce greater model simplicity, thereby reducing variance and ensuring the model generalizes more reliably to data.

Kaggle Link	https://www.kaggle.com/code/naomiespiritu/final-exercise-1-regression-task-assessment-1

Title of Experiment 2:

This classification task involved building a model for clinical diagnoses, a high stakes scenario where the priority is strictly placed on minimizing False Negatives. The initial methodology was ideal, correctly implementing Stratified Sampling and using the class_weight='balanced' hyperparameter across all models to address potential class imbalance and prioritize the detection of positive cases.

Diagnosis:

The most significant takeaway from my exercise was the critical need to accurately diagnose the model's performance problem. Across the Logistic Regression, Random Fores,t and SVM models, the top F-1 score barely exceeded 0.508. This low score indicated that the models were only marginally better than random guessing. Consequently, the initial hypothesis of overfitting was dismissed in favor of diagnosing underfitting. The models were simply too weak to capture the complexity of the data.

This failure was visually confirmed by the ROC and Precision-Recall curves, which remained clustered very close to the baseline of random chance. With this, the curves demonstrated that the algorithms could not establish sufficient power to reliably distinguish between the two classes.

Analysis:

My analysis includes the likely source of the undersitting lies in the preprocessing stage. Using One-Hot Encoding appears this process diluted the signal from the genuinely important numerical predictions like Tumor_Size by introducing too much noise from weakly correlated encoded features. The current feature set as processed, does not provide a clear enough signal for the models to learn from effectively.

To achieve a successful performance in the future, the strat must rotate dramatically towards feature selection and algorithm improvement. These include:

1. *Feature Pruning: to apply advanced techniques such as recursive elimination to aggressively remove non-contributing features*
2. *Algorithm Upgrade: Transitioning to more powerful collection methods, such as Gradient Boosting. These models excel at identifying and weighing complex which is essential to overcome the current underfitting problem.*

In conclusion, this exercise reinforced that in predictive modeling, the quality signal is an absolute necessity for any algorithm's success.

Kaggle Link	https://www.kaggle.com/code/naomiespiritu/final-exercise-2-classification-task-a1

LinkedIn/Kaggle Course Completion Certificate (Optional)

kaggle

CERTIFICATE OF COMPLETION

Naomi Espiritu

HAS SUCCESSFULLY COMPLETED THE COURSE

Intro to Machine Learning

ON OCTOBER 23, 2025

DAN BECKER, KAGGLE INSTRUCTOR

ALEXIS COOK, HEAD OF KAGGLE LEARN