Distinction Task -I (Supervised learning)

Step-1

This task is designed to assess the Distinction level expectations.

Step-2

Your tutor will then review your submission and will give you feedback. If your submission is incomplete the tutor will ask you to include missing parts. Tutors can also ask follow-up questions, either to clarify something that you have submitted or to assess your understanding of certain topics.

Feedback and submission deadlines

Feedback deadline: Tuesday 20 May (No submission before this date means no feedback!), and only **one round of feedback** will be provided for this task.

Submission deadline: Before creating and submitting portfolio.

Required documents

1. Submit a report (pdf format) in Ontrack (https://ontrack.deakin.edu.au)
2. Complete the problem credit task and submit your code file (.ipynb) separately in the OnTrack (https://ontrack.deakin.edu.au).

Background

The growing trend of internet-connected appliances, or IoT devices, in our homes brings both convenience and security risks. These devices can be vulnerable to attacks that steal our data, disrupt their functionality, or even cause physical damage. To combat this, it's crucial to classify these attacks and implement effective defenses. Based on this scenario, the goal is to develop a prediction model that minimizes a total attack related damage and save IoT enables devices from intruder.

Datasets Description

The attributes are as follows: target, then anonymized operational data. In total there are 84 attributes. Missing values are denoted by "na".

Datasets: “Dataset4.csv”

Evidence of Learning

Q1. Develop and evaluate multiple supervised machine learning models to predict the `target` variable.

a. Justify the choice of models and preprocessing strategies used to prepare the data for training and evaluation.
b. Compare the models' performance using appropriate evaluation metrics and discuss any tuning or optimization decisions made.
c. Have you optimized any hyper-parameters for each ML model? What are they? Why have you done that? Explain.
a. Reflect on model performance and behavior (e.g., generalization, complexity, or interpretability), and explain which model you would recommend and why.

Q2. Identify and interpret the most important features for predicting the `target` variable.

a. Select and apply appropriate techniques for determining feature importance, explaining why these methods are suitable for your task.
b. Interpret and compare the insights gained from your selected methods, highlighting any consistent or unexpected results.
c. Discuss how these findings might influence model selection, feature engineering, or deployment in a real-world setting.

Q3. Evaluate the impact of ensemble learning techniques on classification performance.

a. Design and implement a set of ensemble models suited to this task. At least one model should combine diverse base learners (e.g., using a Voting Classifier). Justify your model choices and explain how they aggregate predictions.
b. Assess how ensemble models compare to individual models in terms of predictive performance, robustness, and generalizability.
c. Recommend an approach for real-world deployment, considering trade-offs between accuracy, interpretability, computational cost, and maintainability.

Q4. Critically evaluate the use of Support Vector Machines (SVMs) for multiclass classification in this dataset.

a. Explain how SVMs can be applied to multiclass problems and select a suitable strategy to do so. Justify your approach in the context of this dataset.
b. Evaluate the effectiveness of your chosen strategy in terms of predictive performance and scalability.
c. Discuss any limitations you observed and explore possible improvements or alternative approaches better suited to the task.