CP322 Machine Learning - Assignment 2
Due Date: Nov 4, 2024 at 11:59 PM

About Submission

When writing and submitting your assignments follow these requirements:

- You are expected to submit a single Notebook file for this assignment.
- Extensions for assignment are only granted for medical reasons with a doctor's note. Assignments submitted within 48 hours after the deadline will have their grade reduced by 50%. Submissions beyond 48 hours post-deadline cannot be accepted and will receive a grade of 0.
- Your assignment should be submitted online through the MyLearningSpace website. Email submission is not accepted.
- Please document your program carefully.

Objective

Pokémon, originally known as "Pocket Monsters," is a series of Japanese role-playing video games (RPGs) created by Satoshi Tajiri for the Game Boy. The first games, Pokémon Red and Green, were released in Japan on February 27, 1996. In these games, players catch, train, and trade 151 creatures to become Pokémon Masters. The goal is to collect as many Pokémon as possible by catching them in the wild, trading, and evolving species. The series has since expanded to include various spin-offs and genres, but the core gameplay remains centered around capturing and battling Pokémon.

This dataset includes data on 898 Pokémon, with 1072 entries considering alternate forms. It encompasses various attributes such as the Pokémon's number, name, type, stat total, and basic stats, which include HP, Attack, Defense, Special Attack, Special Defense, and Speed. It also includes data on each Pokémon's generation and whether it is classified as legendary. This assignment involves training various machine learning models to predict whether a Pokemon is legendary and to estimate the combat power of a Pokemon. Combat Power (CP) in Pokémon refers to the overall measured strength of any Pokémon. While we lack a specific CP calculation function, we believe it is closely tied to specific factors. This dataset contains the following features.

- **Number**: The ID for each Pokémon.
- **Name**: The name of each Pokémon.

- **Type 1**: The primary type of the Pokémon, determining weaknesses and resistances to attacks.
- **Type 2**: A secondary type, for Pokémon with dual types.
- **Total**: The overall CP power, serving as a general indicator of a Pokémon's strength.
- **HP**: Hit Points or health, indicating how much damage a Pokémon can withstand before fainting.
- **Attack**: The base modifier for normal physical attacks (e.g., Scratch, Punch).
- **Defense**: The base resistance against normal physical attacks.
- **SP Atk**: Special Attack, the base modifier for special attacks (e.g., Fire Blast, Bubble Beam).
- **SP Def**: Special Defense, the base resistance against special attacks.
- **Speed**: Determines the order of Pokémon's actions within each round of combat.
- **Generation**: Indicates the generation of Pokémon games in which the Pokémon was first introduced.
- **Legendary**: Indicates whether a Pokémon is of legendary status, reflecting its rarity and power.

1 Data Preparation: 3 Points

1. Load the Pokemon dataset, display the first 10 rows, and describe the dataset.
2. Check for and report any missing values in the dataset. If any feature has missing values, report the name of the feature and remove that feature from the dataset.
3. Convert the **remaining** categorical variables to numerical using one-hot encoding. For both classification and regression tasks, split the dataset randomly into 80% for training and 20% for testing, ensuring to set the 'random_state' to 42 for reproducibility.

2 Legendary Prediction: 8 Points

1. Implement a logistic regression classifier to predict whether a Pokemon is legendary. Evaluate the model using precision, recall, and F-measure. Set 'max_iter=1000'.
2. Implement a KNN classifier to predict the legendary status of Pokemon. Use 10-fold cross-validation to determine the best number of neighbors, ensuring the use of Manhattan Distance as the distance metric. Examine the number of neighbours from 1 to 10, and report the optimal K value.
3. Discretize all continuous descriptive features into categorical features with 'KBinsDiscretizer(n_bins=3, encode='ordinal', strategy='uniform')' before training a 'CategoricalNB' Naive Bayesian model. Evaluate the classifier's performance using precision, recall, and F-measure.

4. Plot ROC curves for the logistic linear model, KNN with the optimal K, and Categorical Naive Bayesian models developed above to compare their performances. Report which model yields the best performance.

3 Combat Power Prediction: 4 Points

1. Use logistic regression to predict the CP power (shown as 'total') of a Pokemon, ensuring that **only** continuous descriptive features are considered. Set max_iter=1000. Evaluate the model using Mean Squared Error (MSE).

2. Implement Lasso and Ridge regression model to predict CP power using only continuous features. Report their performance using MSE.