Tabular Project Proposal: Predicting Legendary Pokemon Using Stats and Traits

By: Saige Young

Abstract: Proposed is a machine learning model to predict whether a Pokémon is considered “Legendary” based on its base stats, typing, and catch difficulty. Using data from the Kanto, Johto, and Hoenn regions, we aim to understand the traits that most commonly associate with legendary status. The project will involve feature analysis, classification modeling, and evaluation of prediction accuracy.

Introduction: The Pokémon franchise spans decades, with over 1,000 unique species that vary by stats, types, and rarity. Among these, "Legendary" Pokémon are considered more powerful, harder to find, and often play central roles in games and lore. This project explores how game-design attributes like stats and rarity contribute to a Pokémon being labeled “Legendary.”

We will use this domain as a backdrop to explore classification tasks in machine learning using a real, structured dataset that is both fun and rich in features.

Motivation: The Pokémon dataset is both accessible and interesting to a broad audience. Many players and fans recognize legendary Pokémon as exceptional — but what makes them legendary in terms of data?

By creating a predictive model, we can reverse-engineer the in-game logic and explore how quantifiable traits correlate with Legendary status. This is a compelling classification task and a strong application of exploratory data analysis, supervised learning, and model evaluation.

We have not found existing models that focus specifically on this problem, though the dataset has been used for other classification tasks (e.g., predicting Pokémon type or evolution stage).

Problem Formulation:

- **Dataset**:
  - Pokémon data from Kanto, Johto, and Hoenn (combined total of ~380 rows and 24 columns).
    Data includes:
  - Base stats (HP, Attack, Defense, etc.)
  - Typing (Type 1, Type 2)
  - Catch method and capture rate
  - Evolution stage
  - Physical traits like height, weight, color, and shape
    Type: Categorical + Numerical
    Quality: Clean and structured
- **Data processing**:
  - Encoding categorical variables (e.g., types, catch method, color)
  - Handling missing values if needed
  - Scaling numerical stats (optional)
- **Goal/task**:
  - Binary classification – predict whether a Pokémon is Legendary (Is Legendary as the target variable)
- **ML Algorithm**:
  - Baseline: Logistic Regression or Decision Tree
  - Extended: Random Forest or XGBoost
- **Libraries/tools**:
  - Python (Pandas, NumPy, Scikit-learn, Seaborn/Matplotlib for visualization)
- **Performance metrics**:
  - Accuracy
  - Precision, Recall, F1-Score
  - ROC-AUC
- **Expected performance**:
  - With strong features (like total base stats, capture rate, and evolution stage), we expect to reach ~85%+ accuracy.
- **Training & validation**:
  - Train-test split (e.g., 80-20)
  - Possibly cross-validation
- **Goal of the package**:
  - A simple Python-based tool or notebook that, given a Pokémon's traits, predicts whether it is Legendary
  - Could be adapted to show feature importance or visualize how different features affect the prediction
- **Workplan**:
  - Data cleaning, EDA, initial visualization --- Base model, feature engineering --- Advance model and tuning --- Final presentation