**Datasheet for "Predicting Kart Lap Times Using Machine Learning"**

**Basic Information**

* **Dataset Name**: Synthetic Kart Lap Time Dataset
* **Project Title**: Predicting Kart Lap Times Using Machine Learning
* **Authors**: Himabindu Kummeta
* **Affiliation**: Imperial AI/ML Course
* **Date**15/06/2025

**Motivation**

* **Purpose**: This dataset was synthesized using ChatGPT to support a machine learning pipeline designed to predict kart lap times based on measurable features related to the driver, the kart, and environmental conditions.
* **Intended Use**: Educational purposes, model benchmarking, and demonstration of regression-based performance prediction in motorsports.
* **Target Users**: ML students, motorsport analysts, and enthusiasts interested in predictive modeling.
* **Tasks Supported**: Regression – specifically predicting continuous lap time values in seconds.

**Composition**

* **Instances**: 10,000 samples
* **Features**: 15 input features + 1 target
  + **Driver-related**: Age, Weight, Aggressiveness, Reaction Time, Training Hours
  + **Kart-specific**: Horsepower, Gear Ratio, Fuel Load, Tire Pressure
  + **Track-related**: Length, Number of Corners, Elevation Change
  + **Environmental**: Temperature, Humidity, Wind Speed
  + **Target**: LapTimeSeconds (continuous)
* **Type of Data**: Fully synthetic based on one drivers kart data(generated via ChatGPT)
* **Labels**: Regression target (LapTimeSeconds)
* **Missing Values**: Simulated dataset — missing data was synthetically introduced and cleaned during preprocessing.

**Collection and Preprocessing**

* **Data Source**: Generated using ChatGPT, informed by domain knowledge of motorsport dynamics.
* **Timeframe of Collection**: N/A (synthetic)
* **Annotation**: Not applicable
* **Preprocessing**:
  + StandardScaler applied for feature scaling
  + Train-validation-test split (85/15, with 15% of training set reserved for validation)
  + Missing values handled during synthetic data generation
  + 5-fold cross-validation for model evaluation

**Dataset Statistics**

* **Descriptive Statistics**: Not explicitly included but inferred through visualizations
* **Key Correlations**:
  + TrackLength\_m, KartPowerHP, and FuelLoad\_kg strongly correlated with LapTimeSeconds
* **Visualizations Used**:
  + Pairplot
  + Correlation heatmap
  + Actual vs. Predicted scatter plots
  + Residual plots
* **Issues**: As a synthetic dataset, real-world variation or noise may not be fully represented. Assumes realistic but simplified distributions.

**Ethical Considerations**

* **Sensitive Data**: No personal identifiable information or real-world data is included.
* **License**: Intended for educational and non-commercial use. Reuse allowed with attribution.
* **Ethical Risks**: Low, due to synthetic nature. Should not be used as a proxy for real driver or kart telemetry without domain validation.

**Usage & Evaluation**

* **Published Work**: This dataset is part of a student project and not used in peer-reviewed research.
* **Recommended Metrics**:
  + R² (Coefficient of Determination)
  + RMSE (Root Mean Squared Error)
* **Not Recommended For**:
  + Real-world safety-critical decision-making
  + Commercial karting applications without further validation

**Model Summary**

* **Models Evaluated**:
  + K-Nearest Neighbors
  + Decision Tree (and Tuned)
  + Support Vector Regressor (and Tuned)
* **Best Performing Model**: Tuned SVR
  + **Test R²**: 0.9566
  + **Test RMSE**: 4.90
* **Feature Importance** (from Decision Tree):
  + Track Length, Kart Power, Training Hours, Elevation Change, Number of Corners

**Maintenance & Updates**

* **Maintained By**: Himabindu Kummeta
* **Update Plan**: No updates planned due to synthetic and static nature