**Predicting Kart Lap Times Using Machine Learning: Final Project Report**

**1️⃣ Challenge Overview: Karting Performance Prediction**

In motorsport, particularly karting, lap time is influenced by a complex interplay of driver characteristics, kart setup, and environmental conditions. Accurately predicting lap times helps teams:

* Fine-tune kart setups based on track conditions
* Customize driver training
* Optimize performance strategies

**Objective:**  
To develop a machine learning pipeline that predicts a driver’s **lap time in seconds** based on 15 measurable factors related to the driver, the kart, and environmental conditions.

**2️⃣ Why Machine Learning?**

Traditional statistical methods struggle to capture **nonlinear relationships** and **interdependencies** between multiple karting variables. Machine learning is well-suited for this because:

* It can learn **complex, high-dimensional mappings** from data.
* It allows us to **compare multiple model types** (linear, tree-based, kernel-based).
* With enough data, it can generalize well to new unseen conditions.

**Expected Outcome:**  
A model that accurately predicts lap time and reveals which features (e.g., reaction time, kart power) contribute most to lap performance.

**3️⃣ Data & Preprocessing**

The dataset consists of **10,000 karting samples** with the following features:

* **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)

**Preprocessing Steps:**

* Data cleaning (handled missing values)
* Exploratory visualizations (pair plots, heatmap)
* Feature scaling using StandardScaler
* Train/validation/test split (85/15 with 15% of training used for validation)
* Cross-validation strategy (KFold with 5 splits)

**4️⃣ Model Selection and Justification**

I evaluated the following machine learning models:

| **Model** | **Type** | **Why Included?** |
| --- | --- | --- |
| **K-Nearest Neighbors** | Instance-based | Simple baseline, good for non-parametric data |
| **Decision Tree** | Tree-based | Interpretable, handles non-linearity |
| **Support Vector Regressor (SVR)** | Kernel-based | Captures complex nonlinear relationships |
| **Tuned SVR** | Kernel-based | Hyperparameters optimized using Grid Search |
| **Tuned Decision Tree** | Tree-based | Tuned for depth and split size |

**5️⃣ Results & Evaluation**

The following metrics were used for model performance:

* **R² (Coefficient of Determination)**: Closer to 1 is better.
* **RMSE (Root Mean Squared Error)**: Lower is better.

**🏆 Final Leaderboard**

| **Model** | **CV R²** | **CV RMSE** | **Test R²** | **Test RMSE** |
| --- | --- | --- | --- | --- |
| 🔧 Tuned SVR | 0.9532 | – | 0.9566 | 4.90 |
| 🌳 Tuned Decision Tree | 0.9537 | – | 0.9563 | 4.91 |
| 📉 Decision Tree | 0.9472 | 5.56 | 0.9495 | 5.28 |
| ➰ Support Vector Regressor | 0.8941 | 7.88 | 0.9029 | 7.32 |
| 👥 K-Nearest Neighbors | 0.7834 | 11.27 | 0.7768 | 11.10 |

**✅ Best Model:**  
**Tuned SVR** achieved the highest R² and lowest RMSE across both cross-validation and test set, indicating excellent generalization.

**6️⃣ Visual Insights**

**Pairplot**

Shows relationships between selected features and lap time. Key trends:

* Lower kart power correlates with longer lap times.
* Higher track temperature slightly reduces lap times.

A screenshot of a graph

AI-generated content may be incorrect.A group of blue squares

AI-generated content may be incorrect.

**Correlation Heatmap**

Highlights that **KartPowerHP**, **TrackLength\_m**, and **FuelLoad\_kg** are strongly correlated with lap time.

A screenshot of a graph

AI-generated content may be incorrect.A close up of text

AI-generated content may be incorrect.

**Actual vs. Predicted**

For each model, a scatter plot visualizes how well the predicted lap times align with actual times.

**K-Nearest Neighbors:**

A graph with a red line and blue dots

AI-generated content may be incorrect.

**Decision Tree:**

A graph with a line drawn on it

AI-generated content may be incorrect.

**Support Vector Regressor (SVR)**

A graph with blue dots and red line

AI-generated content may be incorrect.

**Tuned SVR**

A graph with a line and a red line

AI-generated content may be incorrect.

**Tuned Decision Tree**

A graph with dotted lines and numbers

AI-generated content may be incorrect.

**📉 Residual Plot (Random Forest)**

The residuals are symmetrically scattered around zero, indicating no bias and well-fitted predictions.

A graph with blue dots

AI-generated content may be incorrect.

**📌 Feature Importance**

Extracted from the best Decision Tree, showing which features most influence lap time.

Top contributors:

* **TrackLength\_m**
* **KartPowerHP**
* **Training hours**
* **Elevation change**
* **Number of corners**

A graph with blue bars

AI-generated content may be incorrect.

**7️⃣ Conclusion & Next Steps**

* **Tuned SVR** is the most robust and accurate model for this prediction task.
* The approach demonstrates how ensemble and tree-based methods outperform simpler or linear models for real-world regression tasks.
* In future iterations, incorporating categorical data (e.g., driver ID, track type) and telemetry could enhance prediction accuracy.