# Machine Learning based Vehicle Performance Analyzer

## Problem Statement:

To predict the performance of the car to improve certain behaviors of the vehicle using various machine learning algorithms.

## Introduction:

The potential for processing car sensing data has increased in recent years due to the development of new technologies. Having this type of data is important, for instance, to analyze the way drivers behave when sitting behind steering wheel. Very little has been done to analyze car usage patterns based on car engine sensor data, and, therefore, it has not been explored to its full potential by considering all sensors within a car engine. Aiming to bridge this gap, the use of Machine Learning techniques (supervised and unsupervised) on automotive engine sensor data to discover drivers' usage patterns, and to perform classification through a distributed online sensing platform. Such platform can be useful used in different domains, such as fleet management, insurance market, fuel consumption optimization, CO2 emission reduction, among others.

## Literature Survey:

### Machine Learning Based Real-Time Vehicle Data Analysis for Safe Driving Modeling

In the paper "Machine Learning Based Real-Time Vehicle Data Analysis for Safe Driving Modeling" Machine learning approach to analyze and predict the vehicle performance in real time. The focus is on analyzing the data which is collected from the vehicle using the OBD-II scanner and eventually providing the driver's safety solutions The meta features of the vehicle are analyzed in the cloud and are then shared to the concerned parties. The proposed system consists of an OBD-II scanner and a mini dash cam which continuously send data to the cloud server where data analysis is done.

**Multivariate Linear Regression Model:**

It is used when we want to predict the value of a variable based on the value of two or more different variables. The variable we want to predict is called the Dependent Variable, while those used to calculate the dependent variable are termed as Independent Variables.

Features such as fuel efficiency, average speed value, maximum speed value, fourth section speed value, interval driving distance, driving time value during green zone, traveling time value, emergency accelerated value, emergency decelerated value, fourth rpm time value and fifth rpm time value are used for training the model.

The real time data obtained is normalized using Min-Max normalization technique and they hypothesize an outcome called Economic Driving Index (ECN_DRVG_INDX) and another outcome called Safe Driving Index (SFTY_DRVG_INDX). The results have proven to be approximately 80% fitting the given features.

Normalization: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$

Hypothesis:

$$h_1(x) = \sum_{i=1}^{i=5} x_i * \beta_i + bias$$

$$h_2(x) = \sum_{i=1}^{i=4} x_i * \beta_i + bias$$

## A Machine Learning Approach Based on Automotive Engine Data Clustering for Driver Usage Profiling Classification:

The paper “A Machine Learning Approach Based on Automotive Engine Data Clustering for Driver Usage Profiling Classification” proposes the use of Machine Learning techniques (supervised and unsupervised) on automotive engine sensor data to discover drivers’ usage patterns, and to perform classification through a distributed online sensing platform and that such platform can be useful used in different domains, such as fleet management, insurance market, fuel consumption optimization, CO2 emission reduction, among others.

As automotive engine data has no class label we use the following Machine Learning models used for clustering and class labels:

### K means:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

### Expectation-Maximization:

The expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables. It does this by first estimating the values for the latent variables, then optimizing the model, then repeating these two steps until convergence. It is an effective and general approach and is most used for density estimation with missing data, such as clustering algorithms like the Gaussian Mixture Model.

### Hierarchical Clustering:

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster. In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

Machine learning algorithms for Classification:

### Decision Tree:

The decision tree and its variants are the other learning algorithms that divide the input space into regions and has separate parameters for each region. They are classified as non-parametric supervised learning method which is widely used in classification and regression, as well as in representing decisions and decision making. The structure of a decision tree is a tree like flowchart, in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Besides, the paths from root to leaf represent classification rules.

### KNN:

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

**Multilayer Perceptron:**

A multilayer perceptron is a fully connected class of feedforward artificial neural network. it uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

**Naive Bayes**

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

**Random Forest**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

**Support Vector Mechanism:**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

## Driving Behavior Analysis Using Machine and Deep Learning Methods for Continuous Streams of Vehicular Data:

The paper "Driving Behavior Analysis Using Machine and Deep Learning Methods for Continuous Streams of Vehicular Data" authored by Nikolaos Peppes, Theodoros Alexakis, Evgenia Adamopoulou, Konstantinos Demestichas aims to combine well-known machine and deep learning algorithms together with open-source-based tools to gather, store, process, analyze and correlate different data flows originating from vehicles.

Machine Leaning Algorithms for Classification:

**Support Vector Mechanisms (SVM):**

Support vector machines is a supervised machine learning algorithm used for both classification and regression. SVM classifies data points based on the hyperplane in an N – dimensional space.

The separation function in support vector classification is a linear combination of kernels linked to the support vector.

**Decision Tree-Based Algorithms:**

The decision tree and its variants are the other learning algorithms that divide the input space into regions and has separate parameters for each region. They are classified as nonparametric supervised learning method which is widely used in classification and regression, as well as in representing decisions and decision making. The structure of a decision tree is a treelike flowchart, in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. Besides, the paths from root to leaf represent classification rules. Three decision tree-based models, including decision tree (DT), extra trees (ExT), and random forest, were evaluated in relation to various learning methods.

**Random Forest**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Deep Learning Model:

**RNN-based algorithms:**

RNN-based models have been used widely nowadays due to its robustness and capability to handle nonlinear data even with its typically structured, single hidden layer, or advanced structured, multiple hidden layers. RNN includes three layers: input, hidden, and output layers. In case of increasing complexity of the problem, the number of layers will rise, and the computational resources will consequently also rise. Here, both the mentioned structures of the RNN-based models were utilized for predicting the Driving Behavioral Analysis.

**Multilayer Perceptron:**

A multilayer perceptron is a fully connected class of feedforward artificial neural network. it uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

**REFERENCES:**

- Barreto, Cephas & Xavier-Júnior, João & Canuto, Anne & Silva, Ivanovitch. (2018). A Machine Learning Approach Based on Automotive Engine Data Clustering for Driver Usage Profiling Classification. 10.5753/eniac.2018.4414. Daud, M.K.; Nafees, M.; Ali, S.; Rizwan, M.; Bajwa, R.A.; Shakoor, M.B.; Arshad, M.U.; Chatha, S.A.S.; Deeba, F.; Murad, W.; et al. Drinking water quality status and contamination in Pakistan. BioMed Res. Int. 2017, 2017, 7908183.

- Yadav, Pamul & Jung, Sangsu & Singh, Dhananjay. (2019). Machine learning based real-time vehicle data analysis for safe driving modeling. SAC '19: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. 1355-1358. 10.1145/3297280.3297584.

- Peppes, Nikolaos, Theodoros Alexakis, Evgenia Adamopoulou, and Konstantinos Demestichas. 2021. "Driving Behaviour Analysis Using Machine and Deep Learning Methods for Continuous Streams of Vehicular Data" *Sensors* 21, no. 14: 4704. https://doi.org/10.3390/s21144704

- Al-Sultan, Saif & Al-Bayatti, Ali & Zedan, Hussein. (2013). Context-Aware Driver Behavior Detection System in Intelligent Transportation Systems. IEEE Transactions on Vehicular Technology. 62. 4264 - 4275. 10.1109/TVT.2013.2263400.