**Design and Methodology for Prediction of Demand for Public Transportation Services Using Data Analytics.**

**Abstract:**

In this paper we present a design and methodology for conducting experiments on different data analytics techniques to predict the demand for public transportation services. The experiment utilizes three datasets; demand\_data, demographic\_data and weather\_data. Our proposed methodology includes steps such as data preparation, preprocessing, outlier detection feature standardization, encoding of categorical variables and exploration of multiple machine learning (ML) and time series analysis methods. Our ultimate goal is to identify the accurate predictive model that can improve the efficiency and reliability of public transportation services.

**Introduction:**

Efficient public transportation services are crucial for planning and mobility management. Accurately predicting transportation demand can help optimize resource allocation and enhance passenger satisfaction. The focus of this paper is to leverage data analytics techniques in order to develop a predictive model, for public transportation demand.

**Data Preparation and Cleaning:**

The first step involves gathering and cleaning the datasets; demand\_data, demographic\_data and weather\_data. Cleaning the data includes handling missing values removing duplicates and ensuring that the data is consistent.

**Inclusive Preprocessing Methodology:**

To ensure predictions we adopt a comprehensive preprocessing approach. This approach entails exploring and understanding the relationships between these three datasets. We link demographic\_data and weather\_data to demand\_data based on attributes. This linking provides context to enhance our predictive model.

**Outlier Detection and Treatment:**

Outliers can have an impact on the performance of predictive models. Therefore we employ outlier detection techniques like Z score and interquartile range to identify and address any outliers in the datasets. We. Correct or remove these outliers depending on their impact.

**Standardization of Numerical Features:**

Standardizing numerical features is essential to ensure that all features are on a scale. This prevents any bias towards features during modeling particularly for models that rely on distance based calculations, like linear regression.

**Encoding Categorical Variables:**

Categorical variables in the dataset are transformed using a technique called label encoding. This conversion helps to represent values in a numerical format enabling machine learning algorithms to process them effectively.

**We explore machine learning methods for predicting transportation demand**

1. **Linear Regression** - Linear regression serves as a baseline model establishing a relationship between input features and the target variable. It quantifies the impact of each feature on the prediction.