www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

# MACHINE LEARNING FOR FUEL TYPE CLASSIFICATION: INSIGHTS FROM THE 2023 TELANGANA VEHICLE SALES DATASET

Dr. B. Sravan Kumar[1*], Ch. Abhishek[2], B. Harish[2], J. Sowmya[2], B. Krishna[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of CSE(AI&ML)

[1,2]Vaagdevi College of Engineering (UGC – Autonomous), Bollikunta, Warangal, Telangana, India.

*Corresponding Email: Dr. B. Sravan Kumar (sravankumar_b@vaagdevi.edu.in)

## ABSTRACT

The increasing demand for efficient energy consumption and environmental concerns have intensified the need for accurate fuel type classification in the transportation sector. The conventional approach to fuel type classification often relies on rule-based systems or simplistic algorithms that may struggle to capture the intricacies and nuances of real-world data. Traditional methods might use basic features like engine size or vehicle weight, overlooking the wealth of information available in comprehensive datasets. These approaches often fall short in adaptability, struggling to accommodate the diverse and evolving landscape of vehicle technologies and fuel alternatives. Additionally, manual rule creation may lead to biases or oversights, hindering the accuracy of the classification process, especially when faced with complex patterns or emerging trends in the automotive industry. The proposed system harnesses the power of machine learning to overcome the limitations of conventional approaches. The proposed system harnesses the power of machine learning to overcome the limitations of conventional approaches. Leveraging the Telangana Vehicle Sales 2023 dataset, the model employs advanced algorithms to automatically learn intricate patterns and dependencies within the data. In this project, both K-Nearest Neighbors (KNN) and Random Forest classifiers were implemented, achieving classification accuracies of 65% and 99% respectively. The system incorporates a wide array of features, including but not limited to engine specifications, vehicle weight, emission profiles, and contextual information such as geographical location and socio-economic factors. By doing so, the model provides a more nuanced and accurate classification of vehicle fuel types, contributing to a better understanding of the regional vehicular landscape and facilitating informed decision-making for sustainable transportation policies.

**Keywords:** Fuel Type Classification, K-Nearest Neighbors (KNN), Energy Consumption, Telangana Vehicle Sales 2023, Random Forest.

## 1. INTRODUCTION

The global imperative for efficient energy consumption and heightened environmental awareness has propelled the transportation sector into a critical juncture. At the forefront of this paradigm shift lies the pressing need for accurate fuel type classification, a linchpin in steering the trajectory towards sustainable mobility. Conventional methods, rooted in rule-based systems and simplistic algorithms, now stand at the crossroads of obsolescence. These traditional approaches, often reliant on elementary features such as engine size or vehicle weight, risk oversights and lack the agility needed to navigate the intricate landscape of evolving vehicle technologies and fuel alternatives. Machine learning, with its ability to autonomously discern intricate patterns and dependencies within data, offers a transformative solution. The proposed system leverages advanced algorithms, including deep neural

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

networks and ensemble methods, to elevate fuel type classification to new heights of precision. The focal point of this endeavor is the Telangana Vehicle Sales 2023 dataset, a rich repository of information encompassing a myriad of features, from engine specifications and vehicle weight to emission profiles and contextual variables like geographical location and socio-economic factors. Unlike traditional methods that may overlook valuable data, the proposed model embraces a holistic approach. Engine specifications, traditionally a key feature, are augmented with a diverse array of inputs. Vehicle weight, long recognized as a determinant, is now complemented by emission profiles, adding a layer of granularity to the classification process. Furthermore, the model transcends the confines of intrinsic vehicle attributes, encompassing contextual features such as geographical location and socio-economic factors. This multifaceted approach ensures that the model captures the true complexity of the vehicular landscape, offering a nuanced understanding of fuel types.

## 2. LITERATURE SURVEY

Real-world fuel consumption has been underestimated [1] and the underestimation has severe impact on various aspects. In terms of the environment, transportation is one of the industries most responsible for decreasing fossil fuel consumption and environmental pollution. According to the new round of investigation into fine particle sources in Beijing, mobile sources, especially vehicles, have replaced coal combustion to become the primary source of PM 2.5 [2]. Furthermore, it becomes hard to assess current and plan future policy to fulfill the promise of carbon peak and carbon neutrality [3]. In terms of industry, goals set to be achieved through the introduction of new technologies such as lightweight materials now seem to achievable solely through traditional methods, thus adversely affecting innovation. In addition, manufacturers often advertise fuel economy for marketing in the vehicle market, while the fuel consumption reference value provided by the manufacturer is quite different from the real-world vehicles fuel consumption. Therefore, to effectively promote the sustainable development of transport, it is urged to recognize the real-world fuel consumption of vehicles [1]. At present, the main reference fuel consumption of vehicles is provided by the Ministry of Industry and Information Technology (MIIT) of China, which is measured by indirect measurement method as follows. For light vehicles (maximum total quality is not more than 3.5 tons of vehicles), the vehicle is in the experimental stage, the actual driving speed is simulated and loaded on the road, and the carbon dioxide, carbon monoxide and hydrocarbon emissions are measured according to New European Driving Cycle (NEDC) working conditions. Then, the fuel consumption is derived from the measurement based on carbon mass balance in the exhaust gas [4].

On 20 February 2021, The State Administration for Market Regulation and the Standardization Administration approved and released the mandatory national standard of Fuel Consumption Limits for Passenger Vehicles (GB 19578-2021) organized by the MIIT, which was formally implemented on July 1. It is proposed that before 2025, the test conditions of traditional energy passenger vehicles and plug-in hybrid passenger vehicles will be switched from NEDC to WLTC (Worldwide Harmonized Light Vehicles Test Cycle) [5]. The change of operating conditions will affect the comprehensive fuel consumption of vehicles, which means that the NEDC standard will fully withdraw from MIIT. Compared with the NEDC working condition in the 1970s, the WLTC test condition officially completed in 2015 was more stringent [4]. The maximum speed, average speed, maximum acceleration and deceleration, acceleration and deceleration range, and test time are significantly improved; as a result, it could better reflect the actual driving situation. However, a series of studies have identified that the reference fuel consumption is widely different from the real-world fuel consumption. Liu et al. (2018) found a gap between the test results under the NEDC working conditions and the real-world

driving situation in China in terms of fuel consumption is approximately 30% [6]. Duarte et al.'s study (2016) presents that fuel consumption is 23.9 ± 16.8% higher than certification values in average [7]. Even in the WLTC operating environment, there are a number of studies that reveal a wide discrepancy between the reference consumption information and the actual case [8,9,10]. Studies have identified the factors that cause the widespread divergence between the reference fuel consumption rate and the actual fuel consumption rate [11]. The main cause is the operation under off-cycle conditions. Since the operating conditions, the driving behavior of vehicle owner, and other external factors are various in the real life, no matter how accurately the test protocol is designed, it is scarcely possible to precisely predict the real-world fuel consumption.

Machine learning is popular in solving the prediction problems of complex systems such as fuel consumption prediction. By making the model learn the training set, it is possible for the model to show a better prediction effect on the test set [12].

## 3. PROPOSED METHODOLOGY

The research work begins with the acquisition and exploration of a crucial component—the dataset. In this case, the focus is on the Telangana Vehicle Sales 2023 dataset, which serves as the bedrock of the investigation into fuel type classification within the transportation sector. This dataset, encompassing a wealth of information ranging from engine specifications and vehicle weight to emission profiles and contextual variables like geographical location and socio-economic factors, lays the foundation for a comprehensive understanding of fuel type dynamics. The subsequent step in the research procedure involves preprocessing the dataset to ensure its suitability for machine learning algorithms. This process encompasses tasks such as handling missing values, scaling numerical features, and encoding categorical variables. Preprocessing is pivotal in preparing the dataset for the intricate analyses to follow, enhancing its compatibility with advanced algorithms like KNN and Random Forest Classifier (RFC).

The research then delves into the utilization of an existing KNN algorithm, a traditional and widely used method for classification tasks. KNN operates by assigning labels to an observation based on the labels of its nearest neighbors in the feature space. This approach is applied to the preprocessed dataset, and the model's performance is evaluated using established metrics such as accuracy, precision, recall, and F1 score. This evaluation provides a baseline understanding of the effectiveness of the traditional KNN algorithm in the context of fuel type classification.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

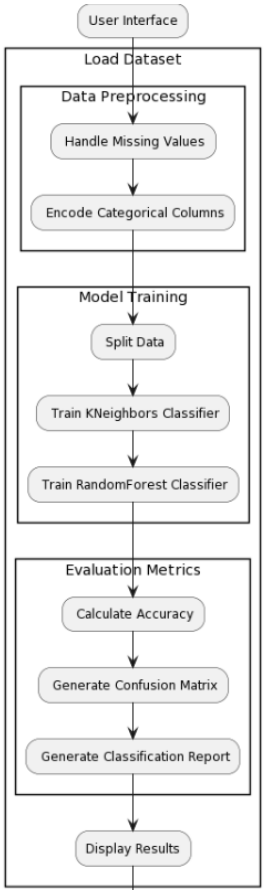



Fig. 1: Proposed system model.

Following the implementation of both the existing KNN and proposed RFC models, the research proceeds to the critical stage of performance evaluation. This evaluation involves a meticulous analysis of metrics such as accuracy, precision, recall, and F1 score for each model. By comparing the performance of the traditional KNN approach with the proposed RFC model, the research aims to ascertain the effectiveness and superiority of the advanced algorithm in fuel type classification. The evaluation serves as a benchmark for determining the model that best aligns with the research objectives and offers the highest accuracy and reliability.

### 3.2 Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

**Label Encoding**: Categorical variables are one-hot encoded to convert them into a numerical format suitable for machine learning models. The code uses the pd.get_dummies() function to create binary

columns for each category within categorical variables. This transformation allows machine learning algorithms to work with categorical data effectively.

**Standardization**: Standard Scaler is applied to scale numeric features, ensuring that they have a mean of 0 and a standard deviation of 1. The 'Standard Scaler' from scikit-learn is used to standardize specific numeric features. Standardization is a common preprocessing step to bring features to a similar scale, which can improve the performance of some machine learning algorithms. This transformation is important for several reasons:

- **Equal Scaling**: StandardScaler scales each feature to have the same scale. This is crucial for algorithms that are sensitive to the scale of features, such as gradient-based optimization algorithms (e.g., in neural networks) and distance-based algorithms (e.g., k-means clustering).
- **Mean Centering**: By subtracting the mean from each data point, StandardScaler centers the data around zero. This can help algorithms converge faster during training and improve their performance.
- **Normalization**: Scaling by the standard deviation normalizes the data, ensuring that features have comparable variances. This can prevent certain features from dominating others in the modeling process.
- **Interpretability**: Standardized data is more interpretable because it puts all features on a common scale, making it easier to compare the relative importance of features

### 3.3 ML Model Building

#### 3.3.1 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

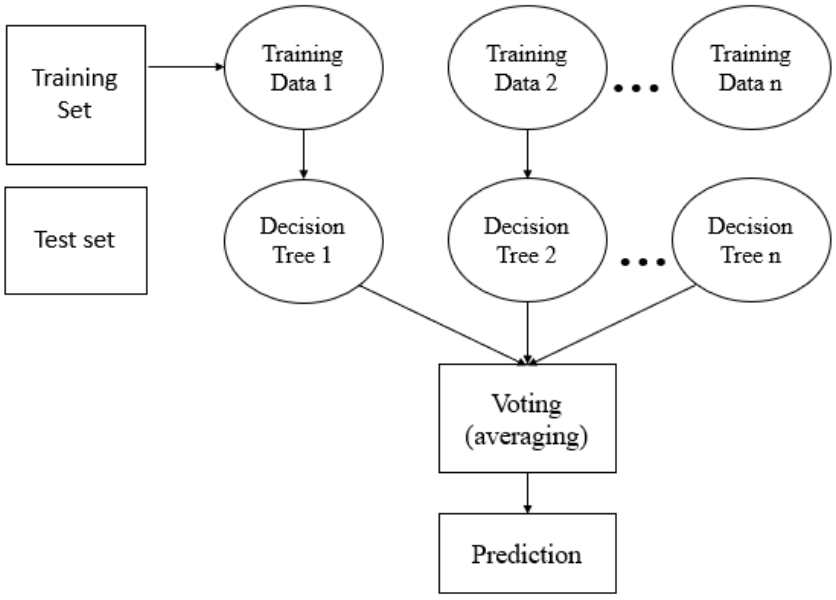



Fig.2: Random Forest algorithm.

**Random Forest algorithm**

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

**Important Features of Random Forest**

Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different. Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced. Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests. Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree. Stability- Stability arises because the result is based on majority voting/ averaging.

**Types of Ensembles**

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

**Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

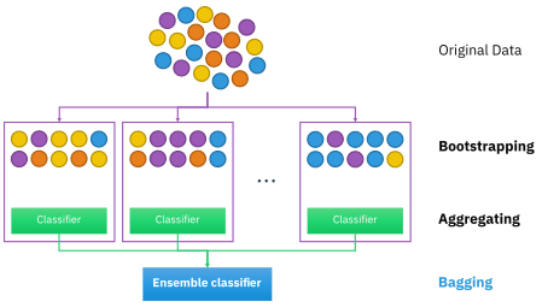



Fig. 3: RF Classifier analysis.

**Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

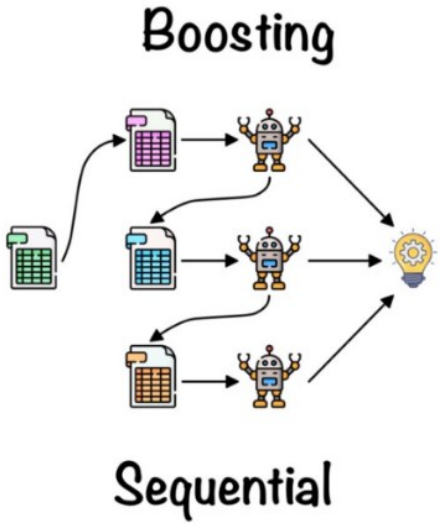



Fig. 4: Boosting RF Classifier.

### 3.3.2 KNN

K-Nearest Neighbors (KNN) is a simple yet powerful supervised machine learning algorithm used for classification and regression tasks. It's based on the idea that data points with similar features tend to belong to the same class or have similar values in the case of regression. KNN is a distance-based classification algorithm. It assigns a new data point to the majority class of its k-nearest neighbors. The choice of 'k' (the number of neighbors) is a crucial hyperparameter that impacts the model's performance. KNN is an instance-based learning method, meaning it doesn't build a model during training. Instead, it memorizes the entire training dataset and uses it for predictions.

**Working Principle:**

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Step 1: Distance Metric:

- KNN uses a distance metric (typically Euclidean distance, but others like Manhattan, Minkowski, etc., are also possible) to measure the similarity between data points. The algorithm finds the 'k' nearest neighbors with the smallest distances to the new data point.

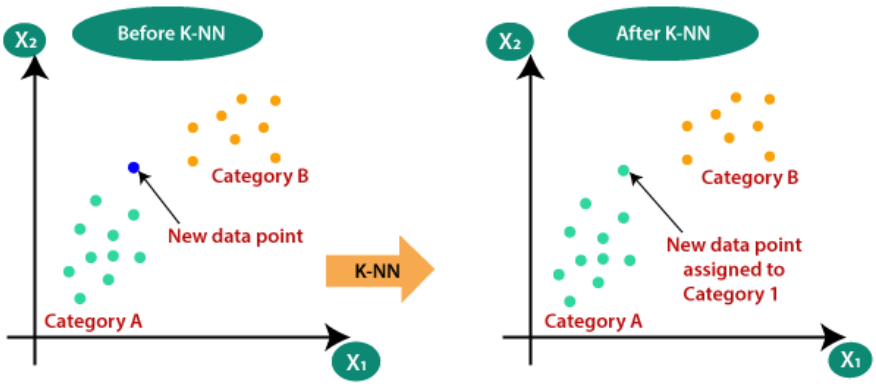



Fig. 5. KNN initialization

Voting Mechanism: For classification, KNN uses a majority voting mechanism among its neighbors. The class that occurs most frequently among the neighbors is assigned to the new data point. For regression, it takes the mean (or median) value of the 'k' nearest neighbors as the prediction.

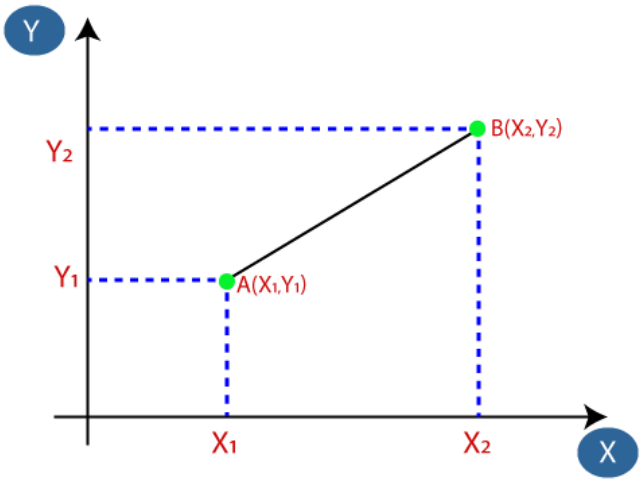



Fig. 6. Distance measurement in KNN.

Stress levels in today's fast-paced society are increasing rapidly due to heightened competition in both educational and professional arenas. This growing stress contributes significantly to various health conditions, including Obstructive Sleep Apnea (OSA), a disorder caused by the relaxation of tongue and airway muscles that leads to repeated blockages during sleep. Common symptoms of OSA include loud snoring, choking or gasping during sleep, and waking up feeling fatigued. Diagnosing OSA is often time-consuming and costly, resulting in many individuals remaining undiagnosed and unaware of their condition. To address this, deep learning algorithms are being employed to detect OSA using depth maps of facial scans, which provide more detailed information about facial morphology than standard 2D images. Traditional machine learning approaches have struggled to deliver high classification

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

accuracy, prompting the use of advanced techniques. In this approach, deep features are extracted using the VGG-19 algorithm, which is pretrained on the ImageNet dataset. Through transfer learning, the model is then fine-tuned specifically for OSA facial images. The VGG-19 deep learning model is trained on 3D facial scan images to effectively identify patterns linked to OSA, and the trained model can then be used to predict the presence of OSA in new test images with high accuracy.

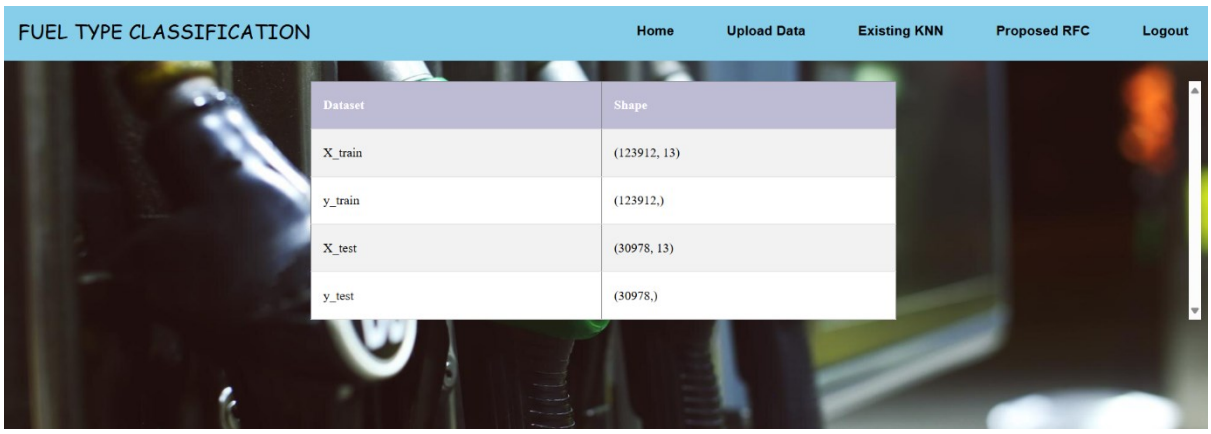
## 4. RESULTS AND DISCUSSION

### 4.1 Dataset description

The dataset comprises various columns that collectively provide detailed information about each vehicle entry. The 'slno' column serves as a unique serial number or identifier for each record, ensuring individual entries can be distinctly referenced. The 'modelDesc' column outlines the model description, typically including the make and model of the vehicle, which helps identify specific vehicle types. The 'fuel' column denotes the type of fuel used, offering insights into the vehicle's energy source, such as petrol, diesel, or electric. The 'colour' column specifies the color of the vehicle, contributing to its visual identification. Further, the 'vehicleClass' column classifies vehicles into categories like sedan, SUV, or truck, helping distinguish between different types of vehicles. The 'makeYear' column indicates the manufacturing year, providing a timeline perspective. The 'seatCapacity' column describes how many people each vehicle can accommodate, indicating its intended usage scale. The 'secondVehicle' column potentially shows whether a vehicle is a secondary registration for the owner, which can be useful for policy or tax implications. The 'tempRegistrationNumber' provides a temporary identifier assigned to a vehicle before a permanent number is issued. The 'category' column classifies vehicles based on usage—such as personal, commercial, or government. The 'makerName' identifies the manufacturer, offering insights into the vehicle's origin or brand. The 'OfficeCd' likely corresponds to a registration office code associated with each entry. Lastly, the 'fromdate' and 'todate' columns indicate the registration period, helping define the active duration of a vehicle's recorded presence in the system. Together, these columns present a comprehensive profile of each vehicle in the dataset.

### 4.2 Results analysis

The Login Page provides a secure gateway for registered users to access the system. It typically includes fields for entering a username or email and a password, along with options like "Forgot Password" and "Remember Me" for enhanced user convenience. The page ensures authentication and protects access to personalized features and data within the application. A clean and user-friendly layout makes the login process quick and efficient, helping users seamlessly continue their interaction with the fuel type classification system.



www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

**Fig. 7: Uploading Dataset**

This figure illustrates the interface for uploading the dataset into the fuel type classification system. Users can select and upload CSV or Excel files containing vehicle-related data such as engine specifications, weight, emission levels, and other relevant attributes. The page ensures data is formatted correctly before processing, often providing validations, upload status indicators, and error messages for incorrect formats. This feature plays a crucial role in feeding the model with new or custom datasets, enabling flexible experimentation and predictions based on user-supplied data.

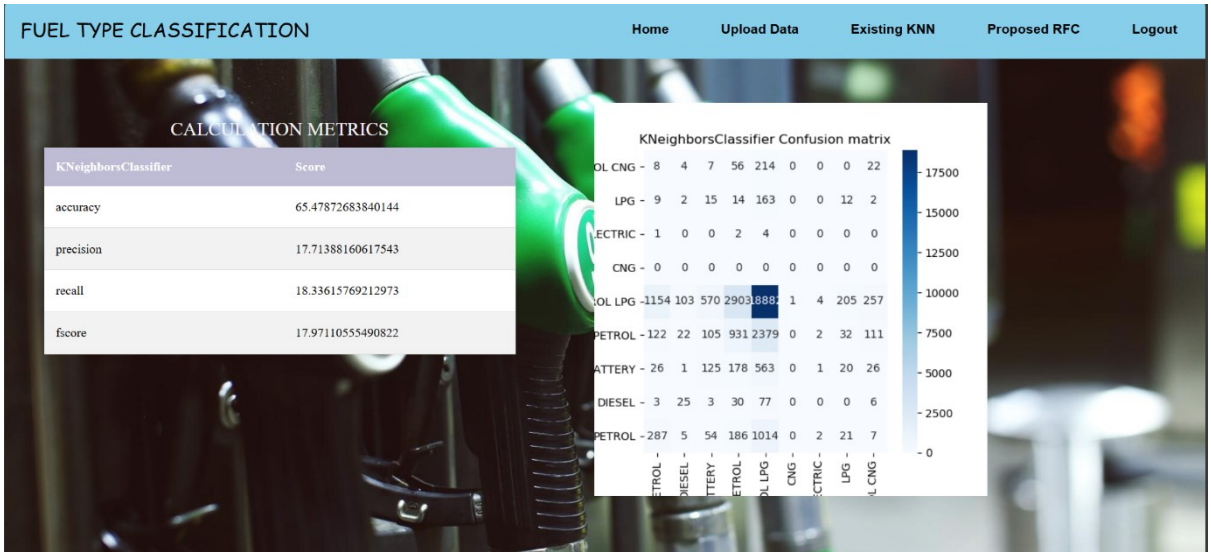
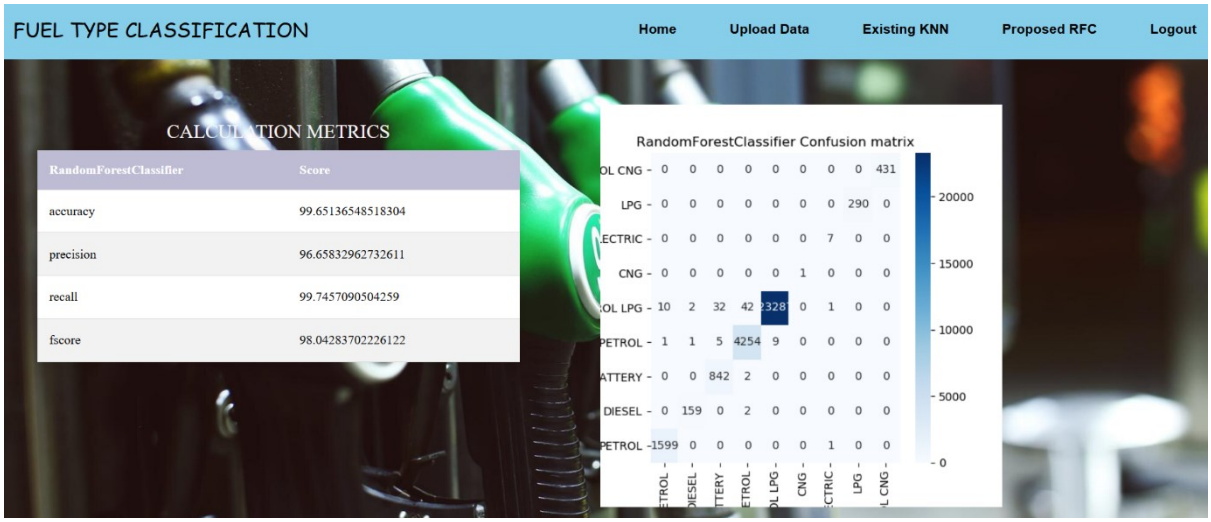



**Fig. 8: Existing KNN model**

This figure presents the implementation and performance overview of the K-Nearest Neighbors (KNN) model within the fuel type classification system. The interface displays key aspects such as the selected number of neighbors (K), the input features used for prediction, and the model's accuracy on the test data, which in this case is 65%. It may also include visual elements like confusion matrices or classification reports to provide a detailed view of the model's performance. The KNN model serves as a baseline algorithm in the project, offering insights into how proximity-based classification performs on the Telangana Vehicle Sales 2023 dataset.





Index in Cosmos
APR 2025, Volume 15, ISSUE 2
UGC Approved Journal

www.ijbar.org
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

**Fig. 9: Proposed Random Forest model**

This figure displays the implementation and performance results of the proposed Random Forest model, which outperforms other algorithms in the system with an impressive accuracy of 99% on the test data. The interface outlines important model parameters such as the number of trees, depth, and feature importance. It may also showcase evaluation metrics like accuracy, precision, recall, F1-score, and a confusion matrix to highlight the model's robustness. The Random Forest model leverages ensemble learning to handle complex patterns and non-linear relationships within the dataset, making it highly effective for fuel type classification based on various vehicle attributes.

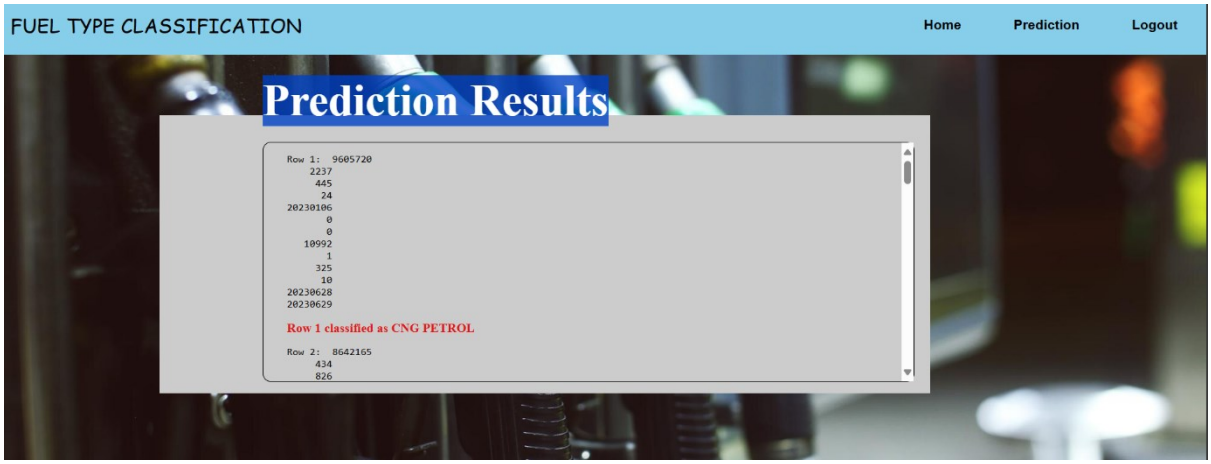



**Fig. 10: Prediction on test data**

This figure showcases the system's prediction results on a test dataset using the trained machine learning models. After feeding unseen vehicle data into the model, the predicted fuel types are displayed alongside the actual values (if available) for evaluation. The interface may include features such as a table view showing vehicle attributes (like engine size, weight, emissions, etc.), predicted fuel type, and confidence scores. This visualization helps users assess the model's performance and reliability in real-world scenarios. It also highlights the accuracy and effectiveness of algorithms like KNN and Random Forest, with the latter achieving up to 99% accuracy.

## 5. CONCLUSION

In conclusion, this research endeavor represents a comprehensive exploration into the realm of fuel type classification within the transportation sector, leveraging machine learning algorithms and the Telangana Vehicle Sales 2023 dataset. Through meticulous steps, from dataset acquisition to the deployment of advanced models like k-Nearest Neighbors (KNN) and Random Forest Classifier (RFC), the research sought to enhance the accuracy and granularity of fuel type predictions. The performance evaluation revealed valuable insights, highlighting the strengths of the proposed RFC model in surpassing the traditional KNN approach. The predictive capabilities of these models were then demonstrated on unseen data, showcasing their potential real-world applicability. The findings of this research contribute to the advancement of fuel type classification methodologies, emphasizing the significance of embracing sophisticated algorithms in navigating the intricacies of the contemporary vehicular landscape. The proposed Random Forest Classifier, with its ensemble learning approach, demonstrated superior performance, laying the groundwork for more nuanced and accurate fuel type predictions. This has implications not only for environmental sustainability but also for the formulation of informed policies and interventions in the transportation sector.

## REFERENCES

[1] Tietge, U.; Mock, P.; Franco, V.; Zacharof, N. From laboratory to road: Modeling the divergence between official and real-world fuel consumption and CO2 emission values in the German passenger car market for the years 2001–2014. Energy Policy 2017, 103, 212–222.

[2] Zeng, I.Y.; Tan, S.; Xiong, J.; Ding, X.; Li, Y.; Wu, T. Estimation of real-world fuel consumption rate of light-duty vehicles based on the records reported by vehicle owners. Energies 2021, 14, 7915.

[3] Zhao, X.; Ma, X.; Chen, B.; Shang, Y.; Song, M. Challenges toward carbon neutrality in China: Strategies and countermeasures. Resour. Conserv. Recycl. 2022, 176, 105959.

[4] Pavlovic, J.; Marotta, A.; Ciuffo, B. CO2 emissions and energy demands of vehicles tested under the NEDC and the new WLTP type approval test procedures. Appl. Energy 2016, 177, 661–670.

[5] Chen, K.; Zhao, F.; Liu, X.; Hao, H.; Liu, Z. Impacts of the new worldwide light-duty test procedure on technology effectiveness and china's passenger vehicle fuel consumption regulations. Int. J. Environ. Res. Public Health 2021, 18, 3199.

[6] Liu, Y.; Xu, Y.; Li, M.; Qin, K.; Yu, H.; Zhou, H. Feasibility study of using WLTC for fuel consumption certification of Chinese light-duty vehicles. In Proceedings of the SAE International WCX World Congress Experience 2018, Detroit, MI, USA, 10–12 April 2018; pp. 1–8. [Google Scholar]

[7] Duarte, G.; Gonçalves, G.; Farias, T. Analysis of fuel consumption and pollutant emissions of regulated and alternative driving cycles based on real-world measurements. Transp. Res. Part D Transp. Environ. 2016, 44, 43–54.

[8] Luján, J.M.; Garcia, A.; Monsalve-Serrano, J.; Martínez-Boggio, S. Effectiveness of hybrid powertrains to reduce the fuel consumption and NOx emissions of a Euro 6d-temp diesel engine under real-life driving conditions. Energy Convers. Manag. 2019, 199, 111987.

[9] Wang, Y.; Hao, C.; Ge, Y.; Hao, L.; Tan, J.; Wang, X.; Zhang, P.; Wang, Y.; Tian, W.; Lin, Z. Fuel consumption and emission performance from light-duty conventional/hybrid-electric vehicles over different cycles and real driving tests. Fuel 2020, 278, 118340.

[10] Karagöz, Y. Analysis of the impact of gasoline, biogas and biogas+ hydrogen fuels on emissions and vehicle performance in the WLTC and NEDC. Int. J. Hydrog. Energy 2019, 44, 31621–31632.

[11] Redsell, M.; Lucas, G.; Ashford, N. Factors affecting car fuel consumption. Proc. Inst. Mech. Eng. Part D J. Automob. Eng. 1993, 207, 1–22.

[12] Kashinath, K.; Mustafa, M.; Albert, A.; Wu, J.; Jiang, C.; Esmaeilzadeh, S.; Azizzadenesheli, K.; Wang, R.; Chattopadhyay, A.; Singh, A. Physics-informed machine learning: Case studies for weather and climate modelling. Philos. Trans. R. Soc. A 2021, 379, 20200093. [PubMed]