Herald College, Kathmandu

Concepts and Technologies of AI

5CS037

SDG Goal 3 — Good Health and Well-Being: Health Impact Classification

Module: 5CS037 — Concepts and Technologies of AI

Assessment: Final Portfolio Project

Student Name: Ishuv Giri

Student ID: 2461987

Institution: Herald College Kathmandu (University of Wolverhampton)

Date: February 2026

Contents

1. Introduction

1.1 Background

Vehicle emissions are the primary contributors to air pollution, the largest environmental health risk, and are killing millions of people every year around the globe. One of the United Nations Sustainable Development Goal 3 (SDG 3) is Good Health and Well-Being (SDG 3), which is to achieve healthy lives and well-being of all people of all ages. Vehicle emissions (such as PM2.5) as well as the nitrogen oxides (NOx) and carbon dioxide (CO2) have a direct effect on respiratory health, cardiovascular disease, and the general well-being of the population. In this project, machine learning classification is used to classify vehicles by the level of health impact with regards to fuel type, engine specifications, and other vehicle characteristics. This research will add to SDG 3 by creating predictive models that determine the health outcomes of various types of vehicles and providing insights with data that can be used to improve transportation policies, emission control, and awareness of the population about their health.

1.2 Aim and Objectives

Purpose: To develop and validate machine learning classification models that can predict the type of health impact of cars (Low, Medium, High) based on their specification and the type of fuel. Objectives: Analyze the data on vehicle fuel and emission intensively using the exploratory data analysis. Define a sensible target variable (Health Impact) through the mapping of the type of fuels to the type of health impact. Build a Multi-Layer Perceptron (MLP) classifier neural network. Compare two classical machine learning models Random Forest Classifier and Logistic Regression. Gridsearchcv Stratified cross-validation Optimise model hyperparameters. Use feature selection methods to describe majority of the discriminative variables. Compare the performance of the respective models and determine the most performing model.

1.3 SDG 3 Relevance

This project has a contribution towards SDG 3 Good Health and Well-Being by:

Re-classification of vehicles based on their possible health effect based on the amount of emissions of the various fuel types.

Provision of the means of analysis of the identification of high-emission vehicle types that pose the biggest threat to the health of the population.

Encouraging policy proposals of cleaner transportation, which are evidence-based. Developing awareness regarding the health impact of fuel choices in the motor industry.

2. Dataset Description and Preprocessing

2.1 Dataset Overview

The dataset used in this project is fullGas.csv, containing vehicle listing data with fuel type, engine specifications, mileage, pricing, and other attributes.

Property	Details
File	fullGas.csv
Original Rows	40,465
Columns	25
Key Features	Fuel_Type, Engine_Size_cc, Power_hp, CO2_Emissions, Mileage_km, Year, Price
Target Variable	Health_Impact — Engineered from Fuel_Type

2.2 Data Cleaning

The preprocessing procedures that were involved were: Duplicate Removal: Rows that indicated duplicates were removed to prevent the cases of data leakage and bias model training. Column Dropping: Imageurl, Make and Model columns have been dropped as they cannot be of help in predicting health impact. Numeric Conversion: Year, Mileagekm, Price, EngineSize, and CO2Emissions were converted to numeric where invalid values were converted to NaN. Missing Value control: Row containing missing values were dropped when the percentage of missingness of features was considered. Outlier Handling: To minimize the impact of the extreme values without removing data points, an interquartile range (IQR) capping of all the numeric columns was performed. Cleaned Final Data: 3,626 rows by 23 columns (after feature engineering).

2.3 Feature Engineering — Health Impact Target Variable

The core contribution of this project's feature engineering is the creation of the Health_Impact target variable by mapping vehicle fuel types to health impact categories based on established emission profiles:

Health Impact	Fuel Types	Rationale
Low	Electric, Hybrid, Electric/Gasoline, Electric/Diesel, Hydrogen	Zero or minimal direct tailpipe emissions
Medium	Gasoline, Petrol, LPG, CNG, Ethanol, Others	Moderate emission levels with cleaner combustion characteristics
High	Diesel	High particulate matter (PM2.5) and nitrogen oxide (NOx) emissions

Class Distribution After Mapping:

Class	Count	Percentage
Medium	1,920	52.9%
High	1,254	34.6%
Low	452	12.5%

The dataset exhibits class imbalance, with the "Low" class being underrepresented. This was addressed through stratified splitting and weighted evaluation metrics.

2.4 Categorical Encoding

Nine categorical columns were encoded using LabelEncoder:

- Body (9 unique values), Country (7), Condition (1), Fuel_Type (8), Drivetrain (3), Gearbox (3), Color (14), Upholstery (6),Seller (1).

The target variable was separately encoded: High → 0, Low → 1, Medium → 2.

3. Exploratory Data Analysis

3.1 Class Distribution Analysis

The target variable distribution was visualised using bar charts and pie charts. The analysis revealed:

- Medium impact vehicles dominate the dataset (52.9%), primarily comprising gasoline, LPG, and CNG vehicles.

- High impact vehicles (diesel) constitute 34.6% of the dataset.
- Low impact vehicles (electric and hybrid) represent only 12.5%, reflecting the real-world market share of electric vehicles.

The class imbalance ratio (min/max) indicated the need for stratified sampling in train-test splitting and cross-validation.

3.2 Correlation Analysis

3.3 Feature Distributions by Health Impact

Correlation heatmap of all the numeric features showed: HealthImpact was the most strongly correlated with FuelType (encoded), which is understandable since the target is by the fuel type. FuelConsumptionl and EngineSizecc showed significant associations with the target which showed the physical association between engine variables and emissions. The correlation of Powerhp and Mileagekm were moderate, which revealed that engine power and vehicle usage patterns are also applicable predictors. Some pairs of features were also inter-correlated (e.g. EngineSizecc and Power_hp), and this was observed in consideration of feature selection.

3.4 Box Plot Analysis

Box plots of main characteristics by the health impact category corroborated: Clarity of division of classes based on fuel related features. There is a lot of features overlap such as Year and Price and these two features are not as discriminative on their own. The existence of these inter-class differences proves the presence of viability of classification modelling.

3.5 Key EDA Insights

The most predictive variable is FuelType, in which the target variable is directly obtained. The engine specifications (EngineSizecc, Powerhp, FuelConsumptionl) have a high discriminative ability. Supplementary predictive data is provided by vehicle characteristics (Mileage_km, Year, Body type). The imbalance in the classes will require attention by use of stratification methods and weighted measures.

4. Model Development

Three classification models were developed: an MLP neural network, a Random Forest ensemble, and a Logistic Regression baseline.

4.1 MLP Classifier (Neural Network)

Architecture Justification:

The most predictive variable is FuelType, in which the target variable is directly obtained. The engine specifications (EngineSizecc, Powerhp, FuelConsumptionl) have a high discriminative ability. Supplementary predictive data is provided by vehicle characteristics (Mileage_km, Year, Body type). The imbalance in the classes will require attention by use of stratification methods and weighted measures.

Parameter	Value	Justification
Hidden Layers	(128, 64, 32)	Three-layer architecture for hierarchical feature learning
Activation	ReLU	Efficient training, avoids vanishing gradient
Solver	Adam	Adaptive learning rates for stable convergence
Learning Rate	0.001	Standard starting rate for Adam optimiser
Max Iterations	500	Sufficient for convergence with early stopping
Early Stopping	True	Prevents overfitting by monitoring validation loss
Batch Size	64	Balances gradient quality with training speed

Results:

Metric	Value
Accuracy	0.9931
Precision (weighted)	0.9932
Recall (weighted)	0.9931
F1-Score (weighted)	0.9931
Training Iterations	21

Per-Class Performance:

Class	Precision	Recall	F1-Score	Support
High	0.99	1.00	1.00	251
Low	0.98	0.99	0.98	91
Medium	1.00	0.99	0.99	384

The performance of the MLP was impressive (99.3% accuracy) and it converged very fast in only 21 iterations. Confusion matrix reflected only 5 misclassifications of 726 test samples, 1 Low sample, misclassified as Medium, 2 Medium samples, misclassified as High, and 2 Medium samples, misclassified as Low. The training loss curve was exponential and had gone almost to zero loss after the 15th iteration

4.2 Random Forest Classifier

Model Justification:

Random Forest has been chosen as the major ensemble model due to: It combines several decision trees in a process of bagging and offers strong forecasts. It works with non-linear relationships and mixed feature features. Importances of features allow interpretation. It does not overfit using ensemble averaging.

Shiniple Default: initial trees=100, maxdepth= 15, minsamplessplit=5, minsamplesleaf= 2.Initial Configuration: 100 trees, max_depth=15, min_samples_split=5, min_samples_leaf=2.

Results:

Metric	Value
Accuracy	1.0000
Precision (weighted)	1.0000
Recall (weighted)	1.0000
F1-Score (weighted)	1.0000

On the test set, the Random Forest reached perfect classification being able to predict all of 726 samples in the test set. There were no cases of misclassifications in the confusion matrix (251 High, 91 Low, 384 Medium: all assigned correctly). Importance Analysis of the features: The 5 features with the highest importance: FuelType (0.51) -The most significant predictor, which is directly associated with emission profiles. FuelConsumptionl (0.07) - Fuel consumption rate is associated with the rate of emission. EngineSizecc (0.06) -- Bigger engines are found to emit more. Powerhp (0.05) - Engine power is connected to strength of combustion. Mileagekm (0.04) - Vehicle characteristics influence the emission.

4.3 Logistic Regression

Model Justification:

The Logistic Regression was used as a linear baseline to: Use non-linear modelling or not based on the problem of classification. Give a comprehensible model whose classes have probability estimates. As a benchmark that is computationally efficient. Options: solver= lbfgs, C= 1.0, maxiter= 1000.

Configuration: solver='lbfgs', C=1.0, max_iter=1000.

Results:

Metric	Value
Accuracy	0.9780
Precision (weighted)	0.9779
Recall (weighted)	0.9780
F1-Score (weighted)	0.9776

Per-Class Performance:

Class	Precision	Recall	F1-Score	Support
High	0.98	1.00	0.99	251
Low	0.98	0.88	0.92	91
Medium	0.97	0.99	0.98	384

Logistic Regression performed well with the accuracy of 97.8, but it demonstrated its weaknesses mostly in the lowest class, which was Low as the recall decreased to 0.88. The misclassification as indicated in the confusion matrix was 16 misclassifications: 1 misclassification of Low sample as High, 10 misclassifications of Low sample as Medium, 3 misclassifications of Medium sample as High and 2 misclassifications of Medium sample as Low. This indicates that the linear decision boundaries have difficulty in distinguishing electric/hybrid vehicles completely with other categories due to the features at hand, as indicated by the lower recall with the Low class.

5. Hyperparameter Optimization

5.1 Random Forest — GridSearchCV

Hyperparameter Grid:

Parameter	Values Tested
n_estimators	[50, 100, 200]
max_depth	[10, 15, 20]
min_samples_split	[2, 5, 10]

Method: GridSearchCV with 5-fold Stratified Cross-Validation, accuracy scoring. Total: 27 combinations × 5 folds = 135 fits.

Best Parameters:

- n_estimators: 50
- max_depth: 15
- min_samples_split: 2

Best CV Accuracy: 0.9983

It is interesting to note that the optimal setup used 50 trees (less than the original 100) indicating that the problem of classifying samples is rather simple when using an ensemble approach. The shallow depth of 15 and little split requirements was enough to bring perfect classification close to perfection.

5.2 Logistic Regression — GridSearchCV

Hyperparameter Grid:

Parameter	Values Tested
C	[0.01, 0.1, 1, 10, 100]
solver	['lbfgs', 'liblinear']
penalty	['l2']

Method: GridSearchCV with 5-fold Stratified Cross-Validation, accuracy scoring. Total: 10 combinations × 5 folds = 50 fits.

Best Parameters:

- C: 10
- solver: lbfgs
- penalty: l2

Best CV Accuracy: 0.9755

The optimal regularisation strength (C=10) suggests that less regularisation is beneficial, allowing the model more flexibility to fit the data. The lbfgs solver was preferred for its efficiency with multinomial classification.

6. Feature Selection

6.1 Method

SelectKBest was used to apply two filter-based feature selection approaches on the entire 22 features to select the best 11 features (about 50% of the total): fclassif (mutual information (mutual_info_classif)): Compares the difference between the attributes of the class means and the intraclass difference. mutualinfoclassif: This model is an experiment that compiles linear and non-linear statistical relationships between features and the target class.

6.2 Results

Top Features by mutual information (mutual_info_classif):

Rank	Feature	F-Score
1	Fuel_Type	~30,000
2	Condition	~1,000
3	Fuel_Consumption_l	~500
4	Mileage_km	~200
5	Doors	~100

Top Features by Mutual Information:

Rank	Feature	MI-Score
1	Fuel_Type	0.98
2	Engine_Size_cc	0.55
3	Power_hp	0.28
4	Fuel_Consumption_l	0.22
5	Mileage_km	0.15

6.3 Analysis

In both mutual information and mutual information analysis both methods affirmed FuelType as the most significantly important feature having the highest scores. The mutual information analysis also identified EngineSizecc and Powerhp to be a most informative non-linear predictor. Finally chosen features (MI): Body, Mileagekm, Price, Year, FuelType, FuelConsumptionl, Gears, Powerhp, EngineSizecc, Seats, Doors (11 features) The final model comparison was done with these 11 features which reduced the dimensions of the model by 50 percent and the most discriminative variables were preserved.

7. Final Model Comparison

All three models were retrained using the 11 selected features and optimal hyperparameters from GridSearchCV.

7.1 Comparison Table

Model	Features	CV Accuracy	Test Accuracy	Precision	Recall	F1-Score
MLP Classifier	11	0.9900	0.9931	0.9932	0.9931	0.9931
Random Forest (Tuned)	11	0.9986	1.0000	1.0000	1.0000	1.0000
Logistic Regression (Tuned)	11	0.9772	0.9807	0.9807	0.9807	0.9805

7.2 Analysis

On the test set, the best result was achieved at 100% accuracy by the Random Forest, and the cross-validation finding of 99.86 indicates the model is a good generalizer. The 50 decision trees with a depth of 15 were sufficient to separate the three categories of health impacts in a clean manner. - The MLP was also very close to perfect performance, 99.31% accuracy, 5 errors. It intersected in 21 epochs, implying that the neural net discovered good boundaries of decision-making in a short amount of time. The limited errors occur around the Low/Medium boundary, hence that is where the model fails. - Logistic Regression was very good but slightly worse and reaching 98.07% accuracy. The 16 errors primarily fell into the Low category (recall = 0.88) which indicates that a linear model cannot separate fully the electric/hybrid vehicles with Medium-impact ones based on the available features. This is not surprising since there are overlapping

specs of fuel type. All the three models had an accuracy of over 97.8% and as such the task per se is clear-cut in separability. The overvalued FuelType, however, engine specifications and consumption data are valuable supplementary clues. The ranking of the performances (Random Forest > MLP > Logistic Regression) indicates the capability of the models to deal with the non-linear interactions. The hard cases remain at the boundaries of similar vehicle profiles of other fading types of fuels.

8. Conclusion and Reflection

8.1 Summary of Findings

We used ML on SDG3 - Good Health and Well-Being by training classifiers to infer the type of health impact of a vehicle based on the specifications and type of fuel used.Three models were developed and compared:

- 3 layer neural net (128-64-32), usage of relu and early stopping, 99.3 percent accuracy, and fast converting.

- Random Forest Classifier 50 tuned trees through GridSearchCV, which with a 100 percent test accuracy.

- Logistic Regression: a linear tuned model, 98.1% accuracy with explainable probability results.

8.2 Key Insights

Top driver is Fuel Type that has a mutual information of 0.98, which basically defines the emission profile.

- Engine details (EngineSizecc, Powerhp, FuelConsumptionl) provide good second level discrimination, which associates design with emissions.

- The categories are distinguishable in our feature space, according to all the models with accuracy greater than 97.8%.

- The selection of features was the task of feature selection which drew 11 vital features through mutual information in order to maintain the model lean.

- CV stratified in tuning hyper-parameters provided sound generalisable cross-class performance.

8.3 Limitations

- The simplified proxies of the health-impact classes are the real impact that varies according to driving patterns, standards of emissions, age and maintenance. The fact that the absolute accuracy of Random forest is so high may be indicative of leaks based on FuelType, which in fact does belong to target. Although the lower recall of the Logistic Regression implies that the

model is learning actual interactions. - The findings might not be extrapolated to other areas - the composition of vehicles in different parts of the world and the fuel type distribution is different. - The Low class (12.5%) is skewed, which may not be very reliable with electric/hybrid cars

8.4 Future Work

- Direct impact assessment of PM2.5, NOx and CO to directly assess health impacts as opposed to proxy classification. - Use finer granularity by trying deeper networks or gradient-boosting. - Use geo-demography to make assessments more geo-specific. - Create a consumer and policymaker friendly tool indicating vehicles based on Level of risk of emission. - Full lifecycle (manufacturing, fuel production, end-of-life) to get a complete picture. - Reduce the imbalance in the classes using SMOTE or weighted training

8.5 SDG 3 — Good Health and Well-Being Contribution

In this project, AI can be used to support SDG3 by: - Bringing to the attention high-emission cars categories that are the most dangerous in the health perspective in terms of a higher level of particulate and NOx. - Artificial intelligence (AI) making transport policy more supportive and based on evidence. - Increasing consumer awareness of the health implications of a variety of fuels. - Empowering with precision such action as emission units, scrappage, and encouraging low-emission options. - Pitching in to air quality in cities through analytical schemes prepared to smart-city architects.

9. References

1. World Health Organization (2022) Ambient (outdoor) air pollution. Available at: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (Accessed: February 2026).
2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011) 'Scikit-learn: Machine Learning in Python', Journal of Machine Learning Research, 12, pp. 2825–2830.
3. Breiman, L. (2001) 'Random forests', Machine Learning, 45(1), pp. 5–32.
4. Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning. Cambridge, MA: MIT Press.
5. United Nations (2015) Transforming Our World: The 2030 Agenda for Sustainable Development. Available at: https://sdgs.un.org/goals/goal3 (Accessed: February 2026).
6. Health Effects Institute (2020) State of Global Air 2020. Boston, MA: Health Effects Institute.
7. European Environment Agency (2021) Air quality in Europe — 2021 report. Luxembourg: Publications Office of the European Union.
8. Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression. 3rd edn. Hoboken, NJ: John Wiley & Sons.

SDG 3 — Good Health and Well-Being: "Ensure healthy lives and promote well-being for all at all ages."