Trip Based Modeling of Fuel Consumption in Modern Fleet Vehicles Using Machine Learning

KEERTHANA.K
Department of Electronics and communication Engineering
Panimalar engineering College

LAVANYAA. R
Department of Electronics and Communication engineering
Panimalar engineering college

LIKITHA .K
Department of Electronics and communication engineering
Panimalar engineering college

EVANGELINE JINCY S.R
Department of Electronics and communication engineering
Panimalar engineering college

<u>Abstract</u>— Ability to model and predict the fuel consumption is vital in enhancing fuel economy of vehicles and preventing fraudulent activities in fleet management. Fuel consumption of a vehicle depends on several internal & external factors However, not all these factors may be measured or available for the fuel consumption analysis.
The main aim of the project is to build a Machine Learning algorithm to predict the fuel consumption of fleet vehicles based on the gas type. A web application is built which is integrated with the ML model.

Keywords— Pandas, numpy, Code, Datasets, Logistic regression, K-nearest neighbors, seaborn, Deep learning.

<u>INTRODUCTION :</u>

The fuel efficiency of heavy-duty trucks can be beneficial not only for the automotive and transportation industry but also for a country's economy and the global environment. The cost of fuel consumed contributes to approximately 30% of a heavy-duty truck's life cycle cost. Reduction in fuel consumption by just a few percent can significantly reduce costs for the transportation industry.

The effective and accurate estimation of fuel consumption (fuel consumed in L/km) can help to analyze emissions as well as prevent fuel-related fraud. As per Environmental Protection Agency (EPA) reports, 28% of total greenhouse gas emissions come from transportation (heavy-duty vehicles and passenger cars) .

The United States Environmental Protection Agency has introduced Corporate Average Fuel Economy (CAFÉ) standards enforcing automotive manufacturers to be compliant with standards to regulate fuel consumption . US EPA regulations enacting fuel economy improvements in freights released in 2016 target truck fuel efficiency, which is predicted to improve by 11–14% by 2021 . Most states have now mandated that truck fleets update their vehicle inventory with modern vehicles due to air quality regulations.

This system is utilized within conjunction with vehicle pace Also seven predictors inferred starting with way review to prepare a neural system model utilizing machine Taking in that predicts Normal fuel utilization done vehicles. The proposed model can be easily developed for each individual vehicle and fitted into one fleet to optimize fuel consumption over the entire fleet.

The model's predictions are comprehensive on fixed window sizes and on the distance traveled. Different window sizes are evaluated and the results mean that the 1km window can estimate the fuel consumption with a coefficient of 0.91 and it also means less than 4% peak to peak percentage error for routes that include

TECHNICAL BACKGROUND:

Bias and variance

In statistical ML applications for regression, the bias of a model is the difference between the estimated value and the true value of the parameter being estimated. This means that bias is a measure of the model's ability to give accurate estimations. High bias is related to underfitting [15]. Variance has to do with the stability of the model in response to new training examples. It can be described as the variation of estimations between different realizations of a model. Variance is small if the training set has a minor effect on the model's estimates. Variance does not measure if a model is correct or not, only if it is consistent. High variance is related to overfitting [15]. The total error of a model can be expressed as

Error = Bias + Variance.

The bias-variance tradeoff is the problem of simultaneously minimizing these two properties to achieve a low error [15]. Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. The model's ability to generalize can be evaluated by examining these two properties.

2. Mean squared error

The mean squared error (mse) of a model is the average of the squares of the prediction errors. The error in this case is defined as the difference between the estimate and the true value. The mse incorporates both the variance of the estimator and its bias

MSE(Variance) = VarianceEstimate + Bias(Estimate, TrueValue)

Thus the mse assesses the quality of an estimator in terms of its variation and degree of bias. The root mean squared error (rmse) is simply the square root of the mse. Using the rmse as a measure will give the same results as using the mse, but the rmse can be considered a more meaningful representation of the error. In this study rmse will be used to evaluate the different models.

Cook's distance

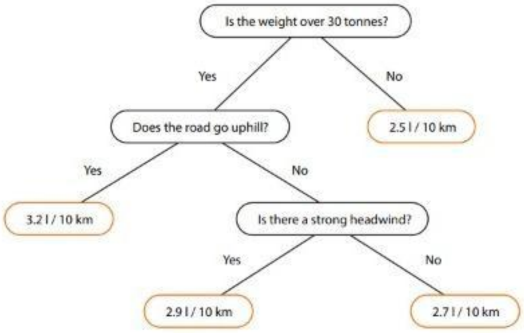
A common metric used to evaluate the influence of a single data point on a linear regression model is Cook's distance. Cook's distance, or Cook's D, is used to estimate the influence of a data point when performing least squares regression. The mathematical definition is given by (3.6) where Yˆ j is the prediction from the full regression model for observation j, Yˆ j(i) is the prediction for observation j from a regression model trained on data where observation i has been omitted, p is the number of fitted parameters in the model and MSE is the mean squared error of the model

$$Di = \frac{Pn\ j{=}1(Y\hat{}\ j - Y\hat{}\ j(i))\ 2}{p\ MSE}$$

In this study Cook's distance will be used to analyze the results of the linear regression fits.

Regression trees

Decision trees are a simple method of supervised learning in which the final model takes a vector of attributes as inputs and returns a single value, or decision, as output. In a decision tree, leaf nodes represent the decisions and branches represent conjunctions of attributes that lead to those decisions



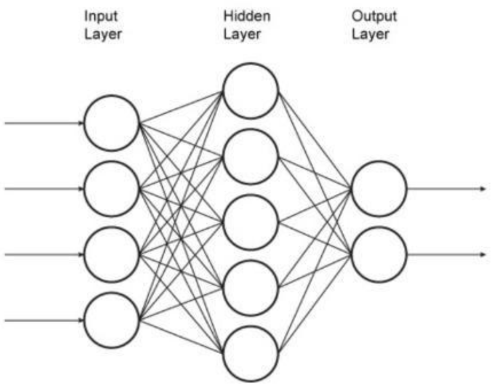
Regression trees are decision trees used for regression, that is their target variable can take continuous values. A regression tree is a tree of nodes where each leaf node has a linear function of some subset of numerical attributes, rather than a single value which is the case for classification trees. The order in which to place the nodes and which node to choose as the root is decided by examining the entropy and information gain of the attributes. Information gain is the expected reduction of entropy achieved after eliminating an attribute from the equation.

Random forests

A random forest for regression is an ensemble learning method where several regression trees are trained and which outputs the mean prediction of the individual trees. Random forests use a modified tree learning algorithm that selects a random subset of the attributes at each candidate split in the learning process. Random forests correct for the tendency of decision trees to overfit to training data [19]. Random forests uses two parameters for tuning a model fit. They are mtry and ntrees. mtry defines how many features to use in each tree and ntrees how many trees to train in total. The default mtry is usually set to the square root of the total number of features and ntrees is usually selected to be as high as possible while keeping training time reasonably short.

Artificial neural networks

Anns were first envisioned as a digital model of a brain, connecting many simple neurons into a network capable of solving complex problems. A neuron in ann terms is a node in the neural network. Roughly speaking one can say that it “fires” when a linear combination of its inputs exceed some threshold The activation function used in the node is most often either a hard threshold function, in which case the node is called a perceptron, or a logistic function. To form a network, the nodes of an ann are arranged in layers and connected by directed links where each link has an associated weight. The layers in between the input and output layers are referred to as hidden layers.



Support vector regression

Support vector machines (svm) is a very popular method for supervised learning and it is a good first try for problems where you do not have any specialized prior knowledge of the problem domain [17]. In its original formulation svms do classification of data points by a maximum margin decision boundary. In its original formulation svms do classification of data points by a maximum margin decision boundary. For example an svm might find the line between two clusters of data points that give the largest margin to the clusters [17]. To find such a decision boundary the svm finds so-called support vectors, which are the data points that lie on or inside the margin. Using the so called kernel trick and dual formulation of the svm optimization problem different kernels may be used to embed the input data in a higher dimensional space, producing non-linear classifiers and greatly expanding the hypothesis space

Evaluation

To determine which models are best suited to fit the data error metrics are computed of the models prediction on test data. The error metrics used are rmse and percent error

1. 2-way analysis of variance

The 2-way anova test is a statistical test for analysing the influence of two different indepent parameters on a single dependent variable [26]. 2-way anova is a parametric test that makes strong assumptions about the distribution of the data [26]. In this study 2-way anova is used to assess what effect the choice of model and the choice of sampling rate has on the prediction error.

2. Friedman test

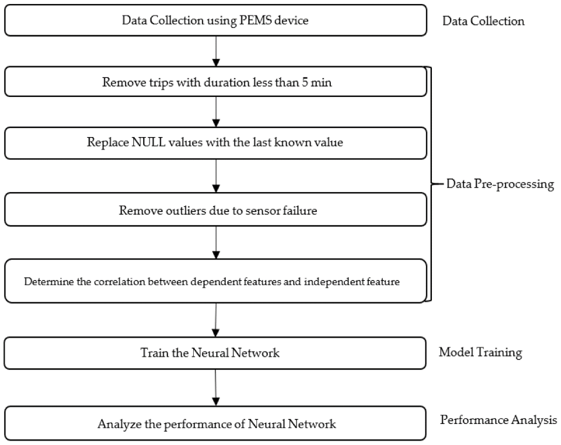
The Friedman test is a non-parametric statistical test that can be used for the same purpose as 2-way anova [27]. Unlike the 2-way anova test, the Friedman test does not make any assumptions about the data distribution [27]. In this study the Friedman test is used to assess what effect the choice of model and the choice of sampling rate has on the prediction error.

MODEL GRAPH & DESIGN:

METHODOLOGY:

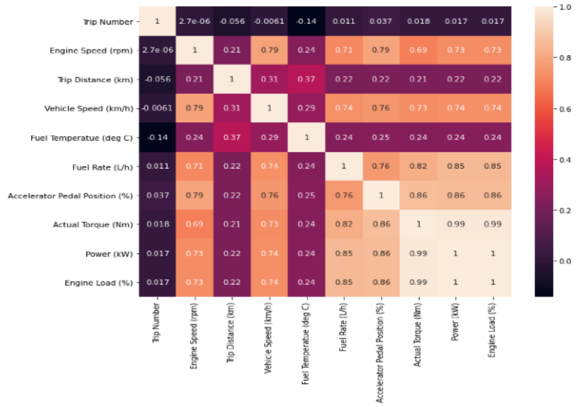
Regression analysis was performed using Machine Learning techniques such as Artificial Neural Network, Linear Regression, and Random Forest to estimate the fuel consumption of modern heavy-duty trucks using PEMS data. The preprocessed dataset, which related to a single vehicle, contained 672,658 rows of actual torque (Nm), vehicle speed (km/h), and engine speed (rpm), which were used as inputs for the models. The implementation stages of the artificial neural network for fuel consumption modeling are as described:

- Creating a database with data collected using PEMS devices during on-road testing of modern heavy-duty vehicles;
- Eliminating test trips that are less than 5 min duration as the trips may not capture information sufficient for the model to generalize well;
- Selecting the parameters that affect the fuel consumption based on parameters collected and domain knowledge;
- Performing correlation analysis on the input parameters selected to eliminate multi-collinear variables;
- Developing the neural networks and identifying the network with best-performing hyperparameters. The hyperparameters include the number of hidden layers, number of hidden neurons per layer, learning rate, and optimization function;
- Calculating the correlation coefficient on the reduced database using the best-performing model selected;
- Perform the generalization analysis of the model by calculating the performance measures such as MAE, RMSE, and R^2;
- Evaluating the performance of the model by comparing the predicted values with the actual values collected during on-road testing.

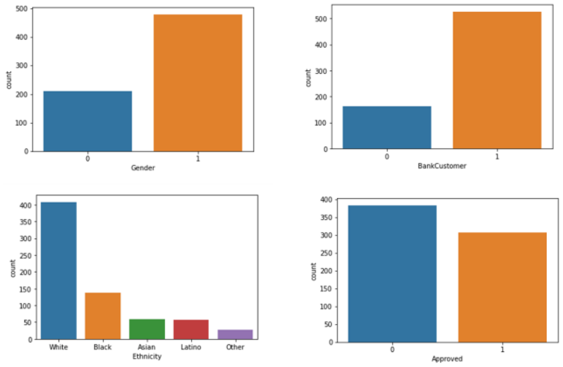

Data Collection using PEMS device
Data Collection
Remove trips with duration less than 5 min
Replace NULL values with the last known value
Data Pre-processing
Remove outliers due to sensor failure
Determine the correlation between dependent features and independent feature
Train the Neural Network
Model Training
Analyze the performance of Neural Network
Performance Analysis

Data Collection and Pre-Processing:

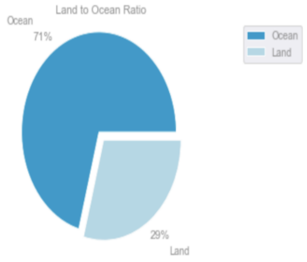
Data collection methods such as onboard emission measurement , laboratory measurement, and tunnel study have been used in past. An on-road data collection method using PEMS is increasingly being used, which makes it possible to collect real-world fuel consumption and emission data and has proved to be reliable. The data used in the current study was collected using a PEMS device mounted on the vehicle during on-road testing at a frequency of 1Hz. PEMS software outputs for the sensor ports were used to process second by second data into a comma-separated values (CSV) file for each trip. Over 100 parameters such as fuel rate (L/h), engine speed (rpm), speed (km/h), gas temperature, CO_2, NOx, GPS altitude, GPS longitude, GPS latitude, etc. were collected for each trip based on data logger settings. Data were collected from two modern heavy-duty trucks with the same make/model of diesel engine manufactured in Detroit in 2016 were used in this study. The trucks were Cascadia models manufactured by Freightliner with DD13 engines and used as goods movement trucks.



MULTIPLE LINEAR REGRESSION:

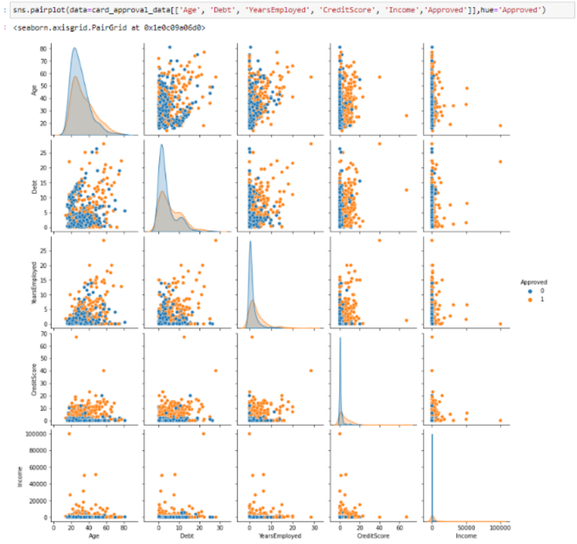


RANDOM FOREST:



Land to Ocean area Pie Chart

PERFORMANCE MEASURES:



```
sns.pairplot(data=card_approval_data[['Age', 'Debt', 'YearsEmployed', 'CreditScore', 'Income','Approved']],hue='Approved')
```

```
<seaborn.axisgrid.PairGrid at 0x1e0c09a06d0>
```

RELATED WORK:

Ability to model and predict the fuel consumption is vital in enhancing fuel economy of vehicles and preventing fraudulent activities in fleet management. Fuel consumption of a vehicle depends on several internal factors such as distance, load, vehicle characteristics, and driver behavior, as well as external factors such as road conditions, traffic, and weather. However, not all these factors may be measured or available for the fuel consumption analysis

1. Sandareka Wickramanayakein April 2016 considered a case where only a subset of the aforementioned factors is available as a multivariate time series from a long distance, public bus. Hence, the challenge is to model and/or predict the fuel consumption only with the available data,

while still indirectly capturing as much as influences from other internal and external factors.In this paper, they compared the predictive ability of three ML techniques in predicting the fuel consumption of the bus, given all available parameters as a time series. Based on the analysis, it is concluded that the random forest technique produces a more accurate prediction compared to both the gradient boosting and neural networks.

2. InAddition,Alexander Schoen in his paper(A Machine Learning Model for Average Fuel Consumption in Heavy Vehicles) described a data summarization approach based on distance rather than the traditional time period when developing individualized machine learning models for fuel consumption. This approach is used in conjunction with seven predictors derived from vehicle speed and road grade to produce a highly predictive neural network model for average fuel consumption in heavy vehicles. The proposed model can easily be developed and deployed for each individual vehicle in a fleet in order to optimize fuel consumption over the entire fleet.
3. In order to enhance the accuracy of fuel consumption Mohamed A. HAMED proposed a model.The proposed model is based on the Support Vector Machine algorithm. The Fuel Consumption estimation is given as a function of Mass Air Flow, Vehicle Speed, Revolutions Per Minute, and Throttle Position Sensor features. The proposed model is applied and tested on a vehicle's On-Board Diagnostics Dataset. The observations were conducted on 18 features. Results achieved a higher accuracy with an R-Squared metric value of 0.97 than other related work using the same Support Vector Machine regression algorithm.
4. Gonçalo Pereira proposed a fuel consumption estimation model that was developed using Machine Learning (ML) algorithms supported by data, which were gathered through several sensors, in a specially designed data logger with wireless communication and opportunistic synchronization, in a real context experiment. The results demonstrated the viability of the method, providing important insight into the advantages associated with the combination of sensorization and the machine learning models in a real-world construction setting.
5. Gonçalo Pereira proposed a fuel consumption estimation model that was developed using Machine Learning (ML) algorithms supported by data, which were gathered through several sensors, in a specially designed datalogger with wireless communication and opportunistic synchronization, in a real context experiment. The results demonstrated the viability of the method, providing important insight into the advantages associated with the combination of sensorization and the machine learning models in a real-world construction setting.
6. In June 2002, the author Yawen Li says that “the fleet vehicle industry is the main source of fuel combustion and environmental pollution”. Therefore, in this paper, we propose a multi-view deep neural network (MVDNN) to analyse the key factors affecting the fuel consumption of automobiles. The experiments show that the introduction of human input improves the prediction accuracy and the root mean square error (RMSE) achieves 0.993. In addition, this paper also finds that for drivers, driving habits, driving frequency, and safety awareness are the most important factors affecting the fuel consumption of vehicles by combining Lasso regression with MVDNN. Finally, by comparing the prediction accuracy of different experiments, relevant policy suggestions are made.
7. Ying Yao on her journal of advanced fleet transportation improved fuel consumption monitoring databases based on mobile

phone data. The fuel consumption prediction models are built using back propagation (BP) neural network, support vector regression (SVR), and random forests. The results show that the average speed, average speed except for idle (ASEI), average acceleration, average deceleration, acceleration time percentage, deceleration time percentage, and cruising time percentage are important indicators for fuel consumption evaluation. All three models could predict fuel consumption accurately, with an absolute relative error less than 10%.

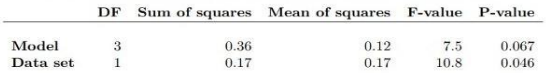
8. In August 2022, The author of the journal “High accuracy fuel consumption based on distance correlation analysis”Ding Rui ,ensured that the accuracy of the model, an integrated structure of the steady-state base module and the transient correction module is determined as the overall structure of the model. Based on the steady-state fuel consumption data, the steady-state base module is established. Then, based on the easily obtained vehicle and engine state parameters, principal component analysis and cluster analysis are used to reasonably classify different driving conditions of the vehicles. Following that, the distance correlation analysis is applied to find the combination of state

RESULTS:

Dividing and normalizing:

the data Before building the different models the data is divided into training, validation and test data sets. The training set is used to train the different models and the validation test is used to verify and evaluate the models during iterative training in order to select the best meta parameters. Once the meta parameters have been chosen and a finished model selected the final model is tested on the test data to evaluate its performance.

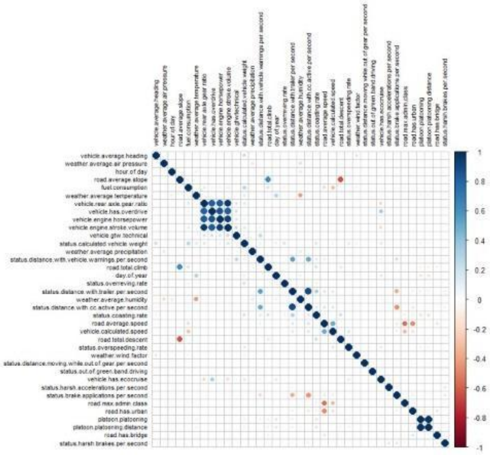
. The validation and test data sets are also split based on date partitioning so that the two sets do not contain observations from the same dates. Partitioning based on date ensures that there is no strong correlation or dependencies between the data in the different sets. If a random sampling method had been applied there would be a risk that observations coming from the same vehicle and the same time period would appear in both training and test sets, giving a strong correlation between the data set



	DF	Sum of squares	Mean of squares	F-value	P-value
Model	3	0.36	0.12	7.5	0.067
Data set	1	0.17	0.17	10.8	0.046

DATA ANALYSIS:

The data is analyzed to find correlations between different features. A correlation plot is presented in Figure 5.1. Strong correlations are found between the features describing the engine characteristics. One interesting question is if the engine features could be reduced to one or two descriptive features using principal component analysis (pca)) or some other method for dimensionality reduction. This question is beyond the scope of this study but could be worth investigating in future research .



CONCLUSION:

In conclusion, the study demonstrates the modeling of fuel consumption in modern heavy-duty vehicles with an artificial neural network using very few technical parameters. An attempt was made to develop a model using very few parameters collected under different conditions. Data from modern heavy-duty trucks with the same make and model, driven by different persons on various routes under different external conditions, were used for training the artificial neural network. The model relies on very few parameters that could be obtained quickly and easily from a vehicle during a trip, unlike other parameters such as road grade, latitude, longitude, traffic information, etc. Moreover, the three parameters used were able to capture a minimum of 78% of the variance in the fuel rate compared to other studies where many parameters are used. Adding more input parameters would improve the performance of ANN , but collecting such data might require additional equipment setup. The performance measures MAE, RMSE, and R^2 indicate that accurate prediction can be obtained with the model. The data modeling can help to identify the trend in instantaneous fuel consumption and to calculate the total fuel consumed by the vehicle for each trip, which can further help in diagnosing vehicle performance in the case of abnormalities. Models that are accurate, fast, and able to predict in real-time will enable the optimization of fuel consumption. The model can be fine-tuned easily to model more complex data from other vehicles with different makes and models that do not have the amount on-road data needed to train a network. This work can be extended to include other factors such as time, traffic information, road information, GPS data, etc. that affect fuel consumption, and to estimate vehicle exhaust emissions.

REFERENCES

- Lee B., Quinones L. and Sanchez J. 2011 Development of greenhouse gas emissions model for 2014-2017 heavy-and medium-duty vehicle compliance SAE *Technical Paper, Tech. Rep.*
  Google Scholar
- [2]
- Fontaras G., Luz R., Anagnostopoulus K., Savvidis D., Hausberger S. and Rexeis M. 2014 Monitoring co2 emissions from hdv in europe-an experimental proof of concept of the proposed methodological approach *20th International Transport and Air Pollution Conference*
  Google Scholar
- [3]
- Wickramanayake S. and Bandara H. D. 2016 Fuel consumption prediction of fleet vehicles using machine learning: A comparative study *Moratuwa Engineering Research Conference* (MERCon), 2016. IEEE 90-95
  Google Scholar
- [4]
- Wang L., Duran A., Gonder J. and Kelly K. 2015 Modeling heavy/medium duty fuel consumption based on drive cycle properties SAE *Technical Paper, Tech. Rep.*
  Google Scholar
- [5]
- Perrotta F., Parry T. and Neves L. C. 2015 Application of machine learning for fuel consumption modelling of trucks *Big Data (Big Data)*, 2017 *IEEE International Conference on.* IEEE, 2017 3810-3815 Paper, Tech. Rep.
  Google Scholar
- [6]
- Dhanalaxmi B. Machine learning and its emergence in the modern world and its contribution to artificial intelligence 2020 *International Conference for Emerging Technology*, INCET 2020 9154058
  CrossrefGoogle Scholar
- [7]
- Nikhil D., Dr. Srinivasa Reddy K. and Dhanalaxmi B. 2020 Image processing

based cancer detection techniques using modern technology - A survey *Proceedings of the 5th International Conference on Communication and Electronics Systems,* ICCES 2020 1279-1282
Google Scholar

- [8]
- Nikhil D., Dhanalaxmi B. and Dr. Srinivasa Reddy K. The evolution of cloud computing and its contribution with big data analytics *Lecture Notes on Data Engineering and Communications Technologies* 46 332-341
CrossrefGoogle Scholar
- [9]
- Dhanalaxmi B., Apparao Naidu G. and Anuradha K. A Rule Based Prediction Method for Defect Detection in Software System *Journal of Theoretical and Applied Information Technology* 95 3403-3412 31st July 2017
Google Scholar
- [10]
- Dhanalaxmi B., Apparao Naidu G. and Anuradha K. A Survey on Design and Analysis of Robust IOT Architectute *International Conference on Innovative Mechanisms for Industry Applications* 375-378
CrossrefGoogle Scholar
- [11]
- Dhanalaxmi B., Apparao Naidu G. and Anuradha K. 2015 Adaptive PSO based Association Rule Mining Technique for Software Defect Classification using ANN *International Conference on Information and Communication Technologies, Procedia Computer Science* 46 432-442
Google Scholar
- [12]
- Dhanalaxmi B., Apparao Naidu G. and Anuradha K. 2017 Defect Classification using Relational Association Rule Mining based on Fuzzy Classifier along with Modified Articial Bee Colony Algorithm *Indian Journal of Applied Engineering Research* 12 2879-2886 June
Google Scholar
- [13]
- Dhanalaxmi B., Apparao Naidu G. and Anuradha K. 2016 A Fault Prediction Approach based on the Probabilistic Model for Improvising Software Inspection *Indian Journal of Science and Technology* 9 December
Google Scholar
- [14]
- Dhanalaxmi G., Naidu Apparao and Anuradha K. 2015 A Review on Software Fault Detection and Prevention Mechanism in Software Development Activities *Journal of Computer Engineering* 17 25-30 Nov – Dec.
Google Scholar
- [15]
- Dhanalaxmi B., Apparao Naidu G. and Anuradha K. 2016 Practical Guidelines to Improve Defect Prediction Model – A Review *International Journal of Engineering Science Invention* 5 57-61 September
Google Scholar