Prediction of Execution Time for PCIe Performance with Regression Models

Ameer Zaaman Raja Salim[1]**, Assoc. Prof. Ir. Ts. Dr. Saidatul Norlyana Azemi**
[1]Advanced Computer Science, Centre of Excellence (CoE), Universiti Science Malaysia (USM), Penang, Malaysia
[2]Computing Centre of Excellence Advanced Computing (AdvComp), Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia.

Corresponding author: First A. Ameer Zaaman (e-mail: ameerpfs@gmail.com).

ABSTRACT In manufacturing company that build product based on their customer preferences and requirements due to their strategic build to order manufacturing model. Based on this approach, they have a high combination of products which causes PCIe (peripheral connector interphase) having inconsistency performance in compliance testing. The product that has a bad performance was unable to determine and prevented thus affecting the overall performance of the graphic functionality. Furthermore, data storage using solid-state drive (SSD) currently has limitations includes time consuming, manually process and data leakage. The PCIe data only saved with SSD that possible misplaced and easily occupied. All the chipset machines having a lot of raw data which easily occupied the storage. Due to the limitation of data storage for PCIe in local host and importance of PCIe data to the company, thus, this project aims to enable AV Cloud with Heidi SQL to upload and retrieve from CDS (cloud domain storage) on PCIe data for each machine connected via internal network. Moreover, speeding up execution time on PCIe is needed to improve the performance in compliance testing that predict execution time for PCie using machine learning approaches including five different regression models ((K-Nearest Neighbour, AdaBoost, Bagging, Linear Regression, Random Forest). The evaluation experiments show that the overall predictive models can predict the execution time for PCIe performance with accuracy more than 70%. The project will be improved in future using the predicted execution times to optimize the PCIe testing and recommended to work on different data science techniques for development process.

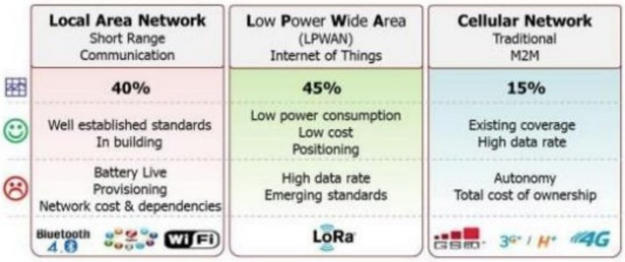
KEYWORDS *Regression, Predictor, Dependent variable, Machine learning, Electronics.*

I. INTRODUCTION

Semi-conductors manufacture motherboards chipsets, network interface controllers and integrated circuits, flash memory, graphics chips, embedded processors, and other devices related to communications and computing. Using cloud service, Intel organizations use shared and storage resources Rather than building, operating, and improving infrastructure on their own. Cloud is a model that enables the following features:

- Resources can be scaled up or down automatically, depending on the load.
- Resources are accessible over a network with proper security.
- Cloud service providers can enable a pay-as-you-go model, where customers are charged based on the type of resources and per user.

Many deployed networks are using mesh network architecture. In a mesh network, the individual end nodes forward the information nodes to increase the range and cell size, but it also adds complexity, reduces network capacity and battery lifetime by forwarding information from other nodes that are likely irrelevant for them **(Paternina et al., 2020).** LPWAN using star architecture for preserving battery lifetime when long-range connectivity can be achieved. In Lorawan, network nodes are not fixed with a single gateway **(Seller, 2021).** Perhaps, data transmitted by a node is typically received by multiple gateways. Each gateway will forward the received packet from the end-node to the cloud-based through network server either cellular, Ethernet, satellite, or Wi-Fi as shown in Figure 1. The intelligence **(Cycleo et al., 2012)** and complexity are pushed to the network server, which manages the network and filters the data redundant received packets, perform security checks, schedule acknowledgments.



	Local Area Network Short Range Communication	Low Power Wide Area (LPWAN) Internet of Things	Cellular Network Traditional M2M
▦	40%	45%	15%
☺	Well established standards In building	Low power consumption Low cost Positioning	Existing coverage High data rate
☹	Battery Live Provisioning Network cost & dependencies	High data rate Emerging standards	Autonomy Total cost of ownership
	Bluetooth 4.0 WIFI	LoRa	GSM 3G / H+ 4G

Figure 1. Lorawan Topology (**Seller, 2021**)

II. PCIe (Peripheral Component Interconnect Express)

PCIe slots is the modules in computer subsystem provide the speed in PC which more on graphic lane speed. To make the PC working as faster as the memory could drive, PCIe was introduced in 2004 (**Intel et al.,2004**). Intel created the PCI Express bus in 2004 to meet the rising need for bandwidth. PCI Express, which was originally developed to allow high-speed music and video streaming, is now utilized to increase the data rate from measuring devices to PC memory by up to 30 times over the regular PCI bus.

Execution time in PCIe will be important measurement as need to complete all the compliancy test within the period (**Intel et al.,2004**). Almost every semi-conductor company keen focus on Signal Integrity test as the main parameters in PCIe modules. But, to make PCIe slots working as much the execution of testing needed to focus as well to enhance the PCIe performance.

III. DATA SCIENCE TECHNIQUES (REGRESSION)

Regression analysis is one of the most frequently used data analyses (Galton, 1989) Sir Francis Galton is credited with coining the phrase "regression." Galton was Charles Darwin's cousin who became interested in science, specifically biology. When evaluating a continuous dependent variable from a set of independent factors, deployment on regression analysis is needed (**Galton et al.,1821**). Logistic regression should be used if the dependent variable is dichotomous approaches. If the split between the two levels of the dependent variable is close to 50-50, logistic and linear regression will provide comparable findings. Regression analysis is a type of predictive modeling approach that examines the connection between a dependent (goal) variable and an independent variable (s) (predictor) (**Galton et al.,1821**). This method is used for forecasting, time series modeling, and determining the cause-and-effect relationship between variables.

A. AV CLOUD

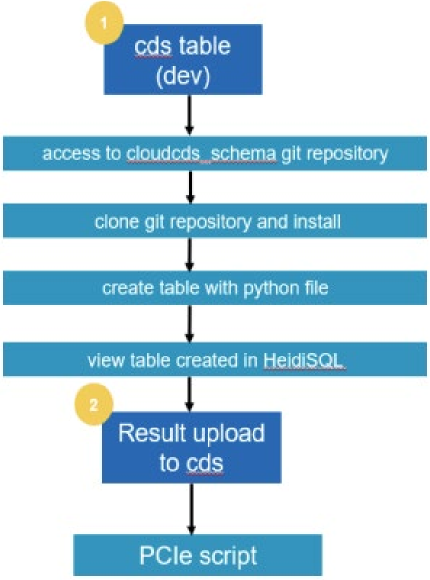

1
cds table (dev)
access to cloudcds_schema git repository
clone git repository and install
create table with python file
view table created in HeidiSQL
2
Result upload to cds
PCIe script

Figure 2. Flowchart of the AVCloud

AV Cloud database solution presents to the user the possibility to store/retrieve data in the database. The solution is based on the Intel Cloud dbaas.intel.com database service. AV Cloud presents the following options, SQL-based relational database (which is based on TSQL language, Microsoft SQL Server 2016) MongoDB documents database, based on mongo database 3.4/4.0. Both options have development and production servers, the first can be used for the development of tools and database tables, the second should be used for real data that are working with. For SQL server, these options differ by user access rights. Development servers allow users to create and kill tables. In production, server tables can be created and killed by AET group only. Based on the Figure 2, the diagram explained the process of the data being uploaded into cloud storage domain (CDS) from the host repository. The data uploaded can be sight in Heidi SQL. This implementation required programming in python to create table in cloud to match the data that user want to view in Heide SQL.

IV. DATA SCIENCE LIFECYCLE

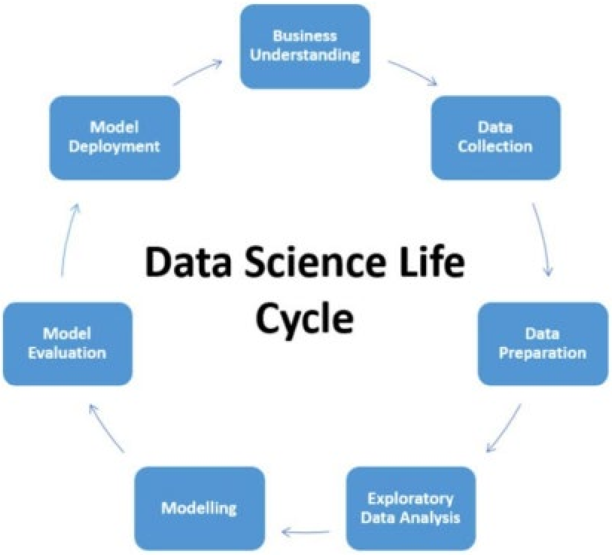


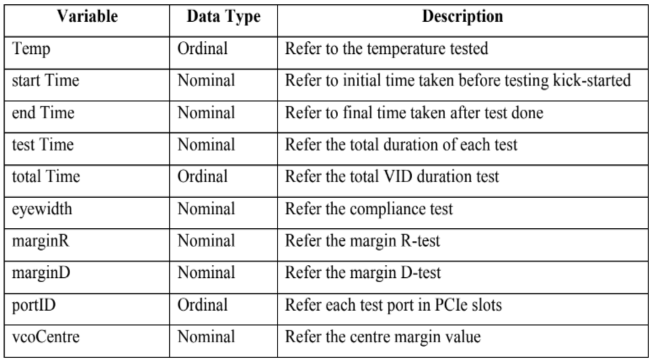
Figure 3. Flow of Data Science Life Cycle (**Daimler-Benz et al.,1994**).

These core data science cycles explained the important aspects that need to consider enhancing the need of the data science analytics in figure 3 to be executed correctly following the business aspects to deliver a comprehensive performance (**Daimler Benz et al.,1994**). Stage 1 is the Business understanding defined to identify data science problems in the company (execution time problem) Stage 2 explained on Data understanding needed to be more on the structure of data. Stage 3 covered the Data preparation include the pre-processing activities that encompass data cleaning, data reduction, and data encoded. Stage 4 explained the Modeling tested with various regression models through evaluation and model deployment were applied.

A. DATA UNDERSTANDING

In this step, the data query from the cloud databases using Heidi SQL to process the data. It is because the file formats like excel, CSV received due using Python, not only that they have specific packages that can read data from these data sources. From this, the excel file was directly downloaded from the database into the local machine. To do this, the script was implemented to pull data from the table on the cloud. In this script, implemented a VID filter means we can download a specific sorted data without double the effort pulled the unwanted data from the cloud. As within this project scope, the data pulled from the cloud only with PCIe test compliance to enhance the performance of the PCIe test through validate the execution time follow the Industrial specification.

From the table 1, each variable explained detailed one-by one to understand the data type before proceeding to data preparation. There are 2 types of data types in this data set which is ordinal and nominal. Most of the data effects the execution time in PCIe compliance testing while it is important to explore the description of each variable clearly portrayed in this table 1.



Variable	Data Type	Description
Temp	Ordinal	Refer to the temperature tested
start Time	Nominal	Refer to initial time taken before testing kick-started
end Time	Nominal	Refer to final time taken after test done
test Time	Nominal	Refer the total duration of each test
total Time	Ordinal	Refer the total VID duration test
eyewidth	Nominal	Refer the compliance test
marginR	Nominal	Refer the margin R-test
marginD	Nominal	Refer the margin D-test
portID	Ordinal	Refer each test port in PCIe slots
vcoCentre	Nominal	Refer the centre margin value

Table 1. Datatype and description for each variable in the dataset.

B. REGRESSION EVALUATION METRICS

There are four models explore namely 1) mean squared log error, 2) mean squared error, 3) mean squared error, 4) R2 score. It contained details of mean squared log error, mean squared error, r2 score, and mean absolute error. The loss function of mean squared log error (MSLE) only cares about the relative difference between the real and predicted values. Usage of this error calculation when the target is normally distributed (**Lehmann et al.,1998**). Mean squared error (MSE) measuring the average of the squares of the errors between the estimated values and the actual values derived from the square of quantity of an estimator **(Lehmann et al.,1998).** MSE has the same units of measurement as the acquirer of the quantity being estimated taking the square root of the quantity being calculated.

R-squared (R2) coefficient of determination, denoted r2 is the proportion of the variance in the dependent variable that is predictable from the independent variables (**Steel et al.,1960**). It is a statistic employed in the context of statistical models, the primary aim of which is either the prediction of future events or the testing of hypotheses based on other relevant information. Based on the fraction of total variance described by the model, it gives a measure of how well-observed outcomes are duplicated by the model **(Steel et al.,1960**).

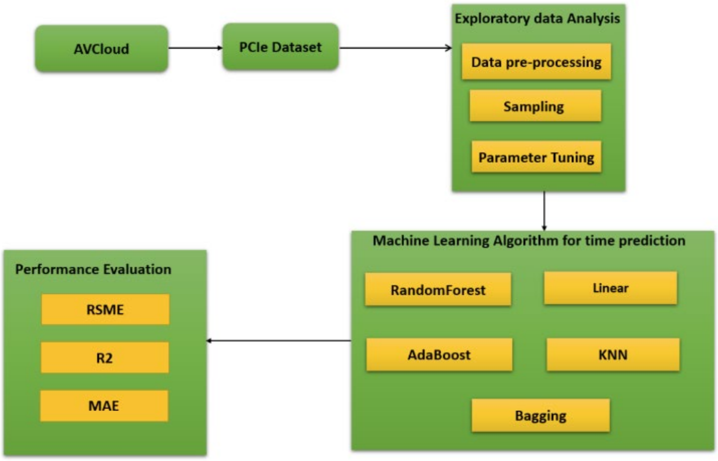


Figure 4. Proposed framework and solution.

After the finalized parameters, the dataset undergoes a framework of data science techniques to find the best model as shown in the above figure 4. Data exploratory analysis using data pre-processing, sampling and parameter tuning to avoid over fitting and data redundancy. This helps to reduce the error before design the modeling concerning each machine learning algorithm. Several algorithms were used for this project. Such as, Random Forest, AdaBoost, Bagging, Linear, and KNeighbors. Each algorithm was evaluated using rsme, r2, mae scores to see the performance on the model.

V. PARAMETERS AND ALGORITHMS TUNING

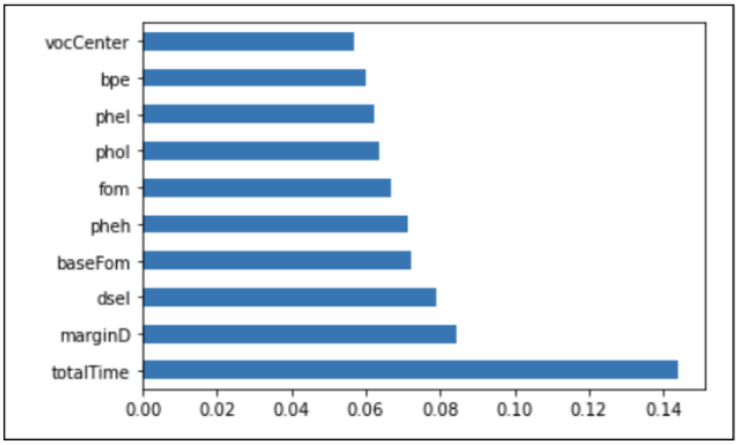


Figure 5. Importance of selected features for development of models

The selected features and parameters prior to design the regression model. Features is selected based on discussion positive correlation between execution time includes vocCentre, bpe, phel, phol, gorm, pheh, baseForm, dsel, marginD and totalTime. Figure 5 shows the features importance that has been selected to design and develop the predictive model using five different regression models. Besides that, the parameters and algorithm tunings are shown in Table 2.

Machine learning algorithms	Parameters	Parameter settings
Bagging	max_features	0.5
	max_samples	0.5
KNN	n_neighbors	2
Linear	fit_intercept	True
	normalize	False
	copy_X	True
	n_jobs	None
	positive	False
AdaBoost	n_estimators	100
	random_state	0
Random Forest	max_depth	2
	random_state	0

Table 2. Tabulation on algorithm and parameter settings for each model.

From the regression perspective, this will correlate which model has the highest accurate values corresponding to this project. There are five different regression model that has been focused which are linear regression, Adaboost, KNN, Bagging and Random Forest.

A. LINEAR REGRESSION

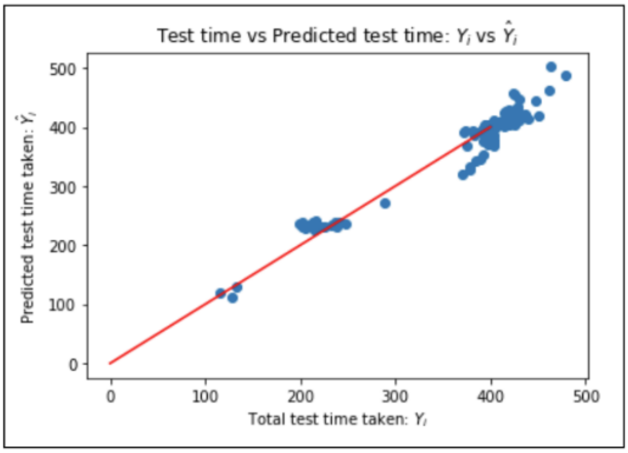


Figure 6. Graph on actual time and predicted time using linear regression.

Based on the ten features selected in Figure 6, linear regression is developed using Linear Regression package in Python Programming Language based on the parameter settings in Table 2. The result is evaluated based on the performance metrics. Figure 7 shows the actual execution time for PCIe and predicted execution time for PCIe using linear regression model.

Figure 6 plotted to see the positively linear relation between the total test time and predicted test time taken for PCIe test to complete. From the Figure 7, the data using total time feature having steady positive correlation but need to see on the accuracy score. This r2_score tell us the model have the high accuracy values when closest to 1. Which clearly see, using linear regression the values of the r2_score is 0.969. This model is compared with other regression model to see the variance on r2_score to predict the best model for this project.

B. K-NEAREST NEIGHBOURS (KNN)

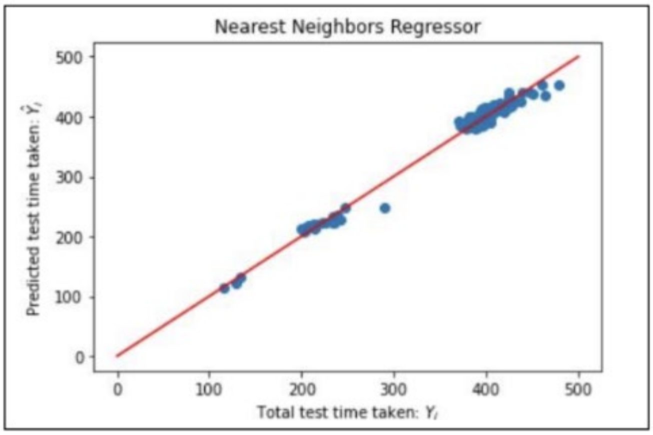


Figure 7. Graph on actual total time and predicted test time taken for KNN Regressor.

KNN algorithm is one of the simplest learning algorithms which can be used in classification problem as well as regression problem. Figure 7 shows KNN is non-parametric algorithm with linear plotted graph shows the r2_score for KNN. The purpose is to use the data points from time perspective into several classes to predict the classification of a new sample point.

C. ADABOOST REGRESSOR

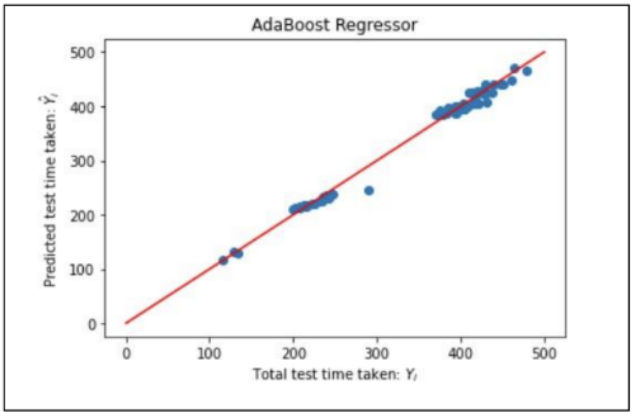


Figure 8. Graph on actual time and predicted time for AdaBoost Regressor.

AdaBoost Regressor is applied to the PCIe data. This model was tested using the same feature as discussed in the linear regression model. The result capture as a graph to tell the r2_score values. From this, the result of the r2_score capture on the following section. Figure 8 show the plotted graph of actual total time and predicted time for PCIe performance and the r2_score result respectively.

D. RANDOM FOREST

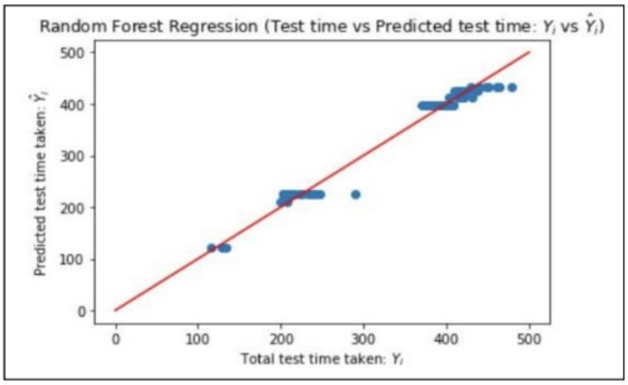


Figure 9. Graph on actual total time and predicted test time taken in Random Forest.

Random Forest is an ensemble learning algorithm. It can be used to improve the accuracy of the model. This random forest regression is a supervised learning algorithm user to tell the method by constructing several training time data and output the mean of the classes data of the overall data. Random forest and bagging regression are used to the predict the execution time for PCIe performance. Figure 9 showing graph plotted using random forest regression on actual total time and predicted test time taken.

E. BAGGING

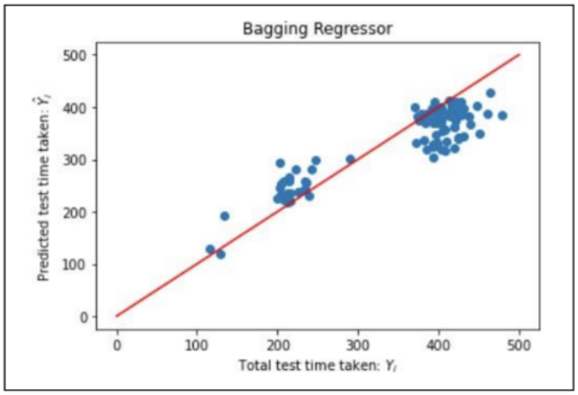


Figure 10. Graph on actual total time and predicted test time taken in Bagging.

Bagging (bootstrap aggregation) is an ensemble machine learning algorithm were used in classification regression problem for decrease the variance in prediction by generating additional training from the dataset using combination of reptations multi-set from original data. It is used to avoid overfitting. Figure 10 shows the graph on actual execution time and predicted execution time for PCIe dataset. Result of r2_score is also calculated.

VI. DISCUSSION

Measure\ *Regression models*	***Linear Regression***	***K-Nearest Neighbour Regressor***	***Random Forest Regressor***	***AdaBoost Regressor***	***Bagging Regressor***
R-squared (R2)	0.969	0.994	0.999	0.995	0.999
Root Mean Squared Error (RMSE)	0.203	0.028	0.022	0.029	0.023
Mean Absolute Error (MAE)	9.91	9.89	13.76	8.96	25.23

Table 3. Result tabulation for all the models.

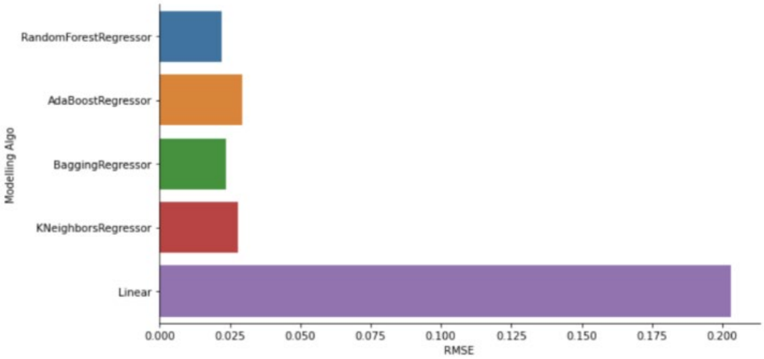


Figure 11. Graph plotted on RMSE in different regression models

Figure 11 shows the error values between the regression models one by one. The blue color indicates the values on Random Forest. Orange indicates the AdaBoost regression. Green indicates Bagging Regression. K-Nearest Neighbors in red column and purple indicates the Linear regression. Based on the above diagram, clearly proven the best model having lowest error for this dataset is Random Forest who lies on 0. 0222.This model show the best model achieved with this dataset is Random Forest to predict the total time testing in PCIe data with lowest error and have the highest accuracy rate in r square

(R2) value 0.999 for this dataset. Thus, overall perfect model for this dataset from the discussion is using Random Forest regression model.

VII. CONCLUSION AND FUTURE WORKS

Data Science analytics tools reduce the time for processing the raw data by the engineers. With this tool, the development even increases the quality of data to maximize the error and duplicate data. Nevertheless, it helps on finding the best model achieved with PCIe data. According to the project perspective, the project will be improved in future using the predicted execution times to optimize the PCIe testing. In addition, the future work will be in the modelling phase using other models in regression. Examples, Arima and Logistic regression to prove the time-series models to validate the accuracy. Even, Decision Trees and Bayes will be added to the benefits through this dataset. In term of cloud implementation, there will need to add another section on analytic structure directly from the cloud without download the CSV files which need more time to implement it.

REFERENCES

[1]. Paternina, C., Arnedo, R., Dominguez-Jimenez, J. A., & Campillo, J. (2020). LoRAWAN Network Coverage Testing Design using Open-source Low-Cost Hardware. 2020 IEEE ANDESCON, 1–6. https://doi.org/10.1109/ANDESCON50619.2020.9272128

[2]. Seller, O. (2021). LoRaWAN Link Layer.Journal of ICT Standardization. https://doi.org/10.13052/jicts2245-800X.911

[3]. Puput Dani Prasetyo Adi, Yuyu Wahyu,The error rate analyzes and parameter measurement on LoRa communication for health monitoring, Microprocessors and Microsystems, Volume 98,2023,104820,ISSN,0141-9331, https://doi.org/10.1016/j.micpro.2023.104820.

[4]. Intel. (2004). Intel, Intel 64 and IA-32 architecture software developer’s manual, volume 3B: system programming guide.

[5]. Galton, F. (1989). Kinship and Correlation. Statistical Science, 4(2). https://doi.org/10.1214/ss/1177012581

[6]. Mariscal,G.,Marban,O.,Fernandez,C. (2010). "A Survey of Data Mining and knowledge discovery process Models and methodologies". *The Knowledge Engineering Review*. 25 (2): 137–166. doi:10.1017/S0269888910000032. S2CID 31359633

[7]. Lehmann, E. L. (1998). Nonparametrics: Statistical Methods Based on Ranks, revised first edition. Prentice Hall, Upper Saddle River, New Jersey. [Previous edition by Holden-Day (1975).

[8]. Steel, R. G. D.; Torrie, J. H. (1960). Principles and Procedures of Statistics with Special Reference to the Biological Sciences. McGraw Hill.

[9]. Hamilton, J. D. (1994). Time series analysis. Princeton University Press.

[10]. Akella Amarendra Babu. (2019). Analysing various Regression Models for Data Processing. Volume-8.

[11]. Anava, O., & Levy, K. Y. (2017). k*-Nearest Neighbors: From Global to Local. ArXiv:1701.07266 [Cs, Stat]. http://arxiv.org/abs/1701.07266

[12]. Breiman, L. (1996). Bagging Predictors. Machine Learning,24(2),123–140. https://doi.org/10.1023/A:1018054314350

[13]. Collins, M., Schapire, R. E., & Singer, Y. (2002). Logistic Regression, AdaBoost and Bregman Distances. Machine Learning, 48(1/3), 253–285. https://doi.org/10.1023/A:1013912006537

[14]. Cristianini, N., & Shawe-Taylor, J. (2013). An introduction to support vector machines: And other kernel-based learning methods. Cambridge University Press.

[15]. Guido van Rossum. (2003). What’s New in Python?

[16]. Harvey, G. (2007). Microsoft Office Excel 2007 for dummies. Wiley.

[17]. Lykiardopoulou, E. M., Zucca, A., Scivier, S. A., & Amin, M. H. (2020). Improving nonstoquastic quantum annealing with spin-reversal transformations. ArXiv:2010.00065[Quant-Ph]. http://arxiv.org/abs/2010.00065

[18]. Mr. K. Mohamed Amanullah, & Mrs. V. Ramya. (2017). Regression Analysis with Cloud computin Technology in the field of Agriculture. International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified, Vol6.https://ijarcce.com/upload/2017/july 17/IJARCCE%2050.pdf.

BIOGRAPHIES

Ameer Zaaman Raja Salim. He born in Penang, Malaysia. First degree in Communication Engineering major in Electronics and Electrical. His working experience in Intel Technology focusing in Analog Design for the silicon on chip (SOC) for the graphic processors.

Started his career in 2018 as Electrical Engineer in Intel focusing in USB2 and GPIO for analog SoC design. After 5 years, manage to lead small team supporting electrical analog validation for Intel's newest business, Intel Arc with 10% of bug-free. Provide technical training in electrical, power delivery, SoC Firmware/Bios and developed automation for global use case.

Awarded as Division Department Award for resolved Electrical issue globally and successfully driving global team to unlock the gating issue as well panelist for Malaysia Young Talent Advisor in 2019. Merit Award for the degree final year project as one of the best project in 2014. One of the Debug Forum members for Malaysia Design Centre for analog graphic products 2021.