4G Signal RSSI Recommendation System for ISP Quality of Service Improvement

Tanatpon Duangta
Faculty of Informatics
Burapha University
Chonburi,Thailand
63910071@go.buu.ac.th

Watcharaphong Yookwan
Faculty of Informatics
Burapha University
Chonburi, Thailand
wyookwan@informatics.buu.ac.th

Anuparp Boonsongsrikul
Faculty of Engineering
Burapha University
Chonburi, Thailand
anuparp@eng.buu.ac.th

Krisana Chinnasarn
Faculty of Informatics
Burapha University
Chonburi, Thailand
krisana@buu.ac.th

***Abstract*— 4G Signal RSSI Recommendation System is one of the monitoring methods. The usage rate of local users improves the quality of traffic signals to cycle to receive increased traffic. This paper proposed a method for Prediction and the traffic of data rates used within the area at each location. The result of the proposed approach comparing the performance of models was: The RMSE Gradient Boost, Decision Tree, and Random Forest were 0.291, 0.316 and 0.346, respectively. The correlation will be 0.976, 0.971, and 0.966 for Gradient Boost Tree, Decision Tree, and Random Forest, respectively, and the accuracy of Gradient Boost, Decision Tree, and Random Forest were 97.8%, 97.4% and 97%, respectively. The results of ensemble learning methods, the RMSE, Correlation, and accuracy were: 0.312, 0.972 and 97.5%**

Keywords— Recommendation System, Prediction, RMSE, Gradient Boost Tree, Decision Tree, Random Forest, Quality of Service

I. INTRODUCTION

At certain times, users of mobile network services do not have access to the data as planned by the provider. Therefore, users will lose their usage benefits due to unsupported usage restrictions. If there are many users in such areas, the person who loses the advantage is, therefore, the service provider. Therefore, give users access to the information and get the most out of their intended use.

Service providers will need to adjust their plans and improve performance for mobile data usage, with demand doubling every year, and monthly traffic is expected to reach 77.5 Exabytes in 2022.[1], An essential aspect of its use is the Prediction for improving parameters to deal with Quality of Service (QoS), such as the rate of traffic into future communication networks. [2] [3].

The number of users in different areas and periods will cause problems for some users who don't have access to information and users who aren't get fully serviced.

Machine-Learning-Based Uplink Throughput Prediction from Physical Layer Measurements 2022 [4] has selected features: for modelling for Prediction, which includes RSRP SNR RSRQ. This research predicts three locations: Batman/Turkey, Melbourne/FL/USA, Houston/TX/USA. The Experiment Result achieved the Average RMSE [Mbps] Analysis of ML Algorithms for Melbourne, FL, Batman, Turkey, and Houston, TX were 4.846, 4.972 and 8.616, respectively.

Uplink Throughput Prediction in Cellular Mobile Networks 2020 [5] has selected key features for modelling, model performance evaluation, and Prediction, which includes RSRP SNR RSRQ. This research used a Gradient Boost Tree the Experiment result achieved the correlation between UL and RSRP, RSRQ and SNR were 0.85, 0.29, and 0.62, respectively.

4G LTE Network Throughput Modelling and Prediction 2020 [6] has selected features, including RSRP SNR RSRQ RSSI DL-bitrate UL-bitrate Latitude Longitude, for model performance evaluation and Prediction. This research used Gradient Boost Tree, Decision Tree Classifier and Random Forest for the classifier. The experiment results achieved, the RMSE Gradient Boost Tree, Decision Tree Classifier, and Random Forest were 0.291, 0.316 and 0.346, respectively.

A Recommendation enables Internet Service Providers (ISPs) to reduce power consumption from low-user towers and scale the signal to support users in high-density areas. The exact amount of signal strength emits the general power consumption of the antenna. Which sometimes has a small number of users, causing a waste of energy

The researchers then developed a process of predicting 4G Signal RSSI Recommendation System for ISPs. Using machine learning methods to help predict the speed of data transmission so that service providers can prepare, and support data based on the growing number of users

II. PROBLEM ANALYSIS

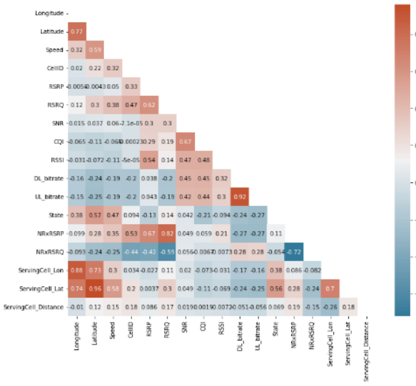

Figure 1 Correlation of Features

The features selected by Conventional Methods are insufficient, thus making the results of model performance evaluations faultier than the proposed method. The features used between Conventional Methods and the proposed method are compared. The problem is that in most conventional method sections, there are only main features: RSRQ, SNR, RSRP, RSSI, Latitude, Longitude, Date, and Time. That results from a very dislocated evaluation of the model's performance. Compared to the proposed method, it uses features CQI, CellID, DL_bitrate, Latitude, Longitude, NetworkMode, Operatername, RSRP, RSRQ, and Speed to help assess the performance of the model for better results. It leads to creating a Recommendation System to address the problem of supporting enough users in the area and allocating the signal intensity emissions in the densely populated place.

III. METHODOLOGY

This section describes the process of data import, data cleaning, and feature selection. To use machine learning modelling, model optimization processes, and finding the best model to ensemble learning for recommendations. System. Results were obtained from testing the performance of the model. The overview process diagram will show in figure 2.

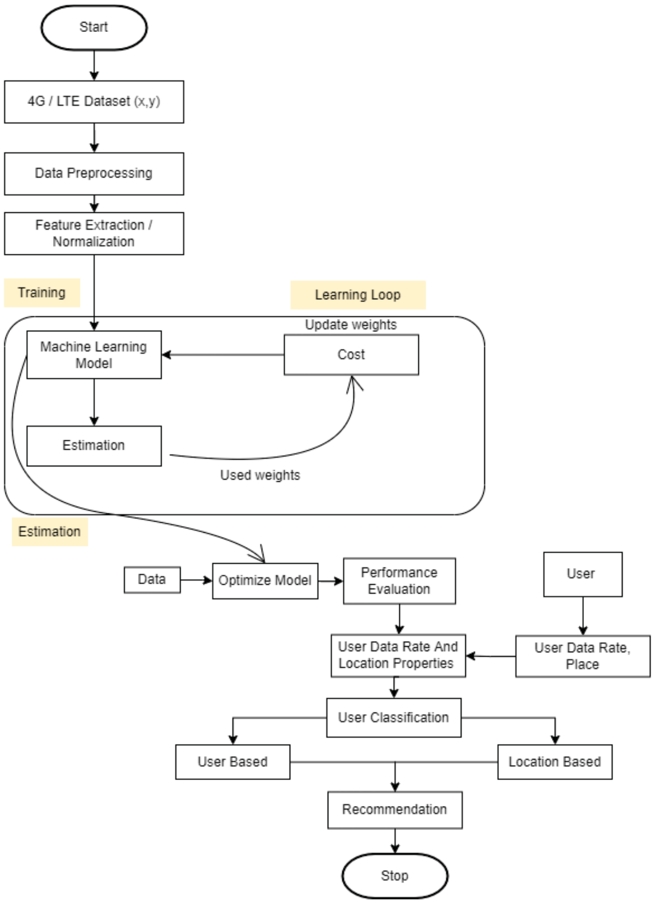

Figure 2. Overview Process Diagram

A. 4G / LTE Dataset (x,y)

The data is imported as a public 4G /LTE Dataset from the Kaggle website.

B. Data Preprocessing

The data preprocessing process starts by extracting data from the types of data that divided into sections and then performing data cleansing to complete the data with the interpolation method. Data Preprocessing Process with Extract Transform Load (ETL) method, The ETL (process extracts data from the data types into sections. To handle missing value data and transform it into the same format, it will load all data into the same data set in preparation for further evaluation of the model's performance.

1) Data Extraction

Data Extraction Process is part of extracting data as the input data is divided into dates. Therefore, the information obtained from multiple sources must be combined into a single data set.

2) Data Cleansing

Data Cleansing is a part of data management. Since the imported data contains a certain amount of error data, Data Cleansing is required.

3) Fill Missing Data

Fill Missing Data Process is filling missing data to make it continuous, in which the data acquisition is filled by a method of Linear Interpolation. Linear Interpolation is a method that uses a function to study the relationship with linear equations, which gives an estimated value of a process with acceptable accuracy within the equations (1):

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1} (\bar{x} - x_i) \quad (1)$$

where two points are known as (x_1, y_1) and (x_2, y_2)
$\bar{x}$ is mean of x

4) Data Transform

Data Transform Process is the process of transforming data from a data set. Text or characters such as Columns State datasets are converted to numbers 1 and 2 for use in machine learning models' processes.

C. Feature Extraction / Normalization

The feature extraction process was selected using a selection method. Correlation features of relevant data were used to determine data for use in Machine Learning Models.

1) Correlation

Correlation is studying the relationship between two or more variables to see how they are related. And are described in any direction can be calculated from the equation (2)

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (2)$$

where x_i is actual value and $\bar{x}$ = mean of x
y_i is actual value and $\bar{y}$ = mean of y

In The Figure 3 will show the information that interests for prediction is UL_bitrate data, so we have found the relationship between tables, UL_bitrate, and other data.

	Longitude	Latitude	Speed	CellID	RSRP	RSRQ	SNR	CQI	RSSI	DL_bitrate
UL-Bitrate	-0.15	-0.25	-0.19	-0.2	0.043	-0.19	0.42	0.44	0.3	0.92

Figure 3 table of data relationships between UL-bitrate

D. Machine Learning Model

In the selection of features to train in the modelling, Features are used to create a machine learning model to test the performance of the selected data against the model used to evaluate the performance of each machine learning model. The Figure 4 will show the relative error of each machine learning process helps to test and evaluate model performance.

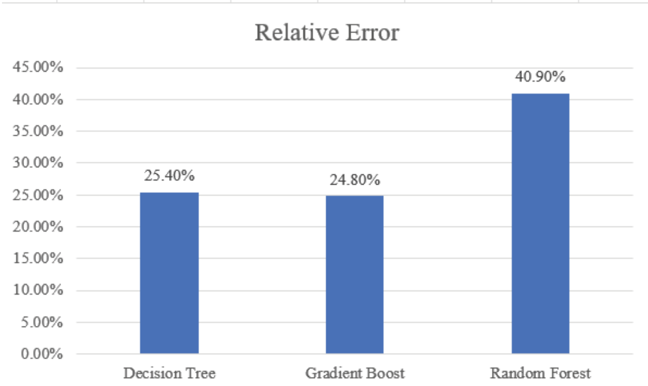

Figure 4 The Relative Error of Machine Learning Model

1) Decision Tree

Decision Tree is one of Machine Learning's tools that help you analyze events or situations to help you make quick decisions. It looks like a tree. Used to evaluate model performance with the equations (3) and (4):

$$Entropy\ (c) = \sum_{i=1}^{n} - P_i \ log_2(P_i) \quad (3)$$

$$Information\ Gain(A) = Entropy(c) - Entropy_A(c) \quad (4)$$

c is the class that interests us, and it's calculated by Entropy as a probability of i.

P_i is the probability of a class in each dimension.

A is Information of the features or dimensions that is interested\

It is a calculation of the probability of each Attribute. Compared to Class, find the Attribute with the highest probability to be the Root Node of the Decision Tree.

2) Random Forest

Random Forest is one of the prediction models. That consists of identical models used to train multiple similar models on the same data set. It uses a data training method that selects different pieces of data. Then decide on those models to vote on which class is the most chosen.

3) Gradient Boost Tree

The principle is that several decision trees like Random Forest will be created, the Figure 5 will show each of which will be built about the previous model, with the current tree being designed to increase the score or weight of the false prediction of the first tree rather than the correct prediction, which will make the model better, more accurately predictable.

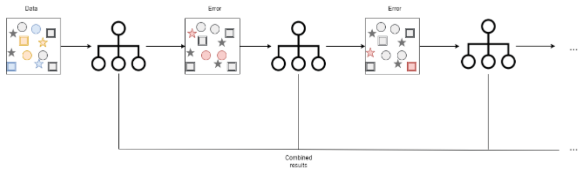

Figure 5 Gradient Boosting Decision Tree Structure

E. Optimize Model

Optimize Model process to test the data and determine the best performance the model can achieve with the 4G LTE Dataset. The Figure 6 That show the process of evaluating and ignoring model performance has few tolerances and gets the best results.

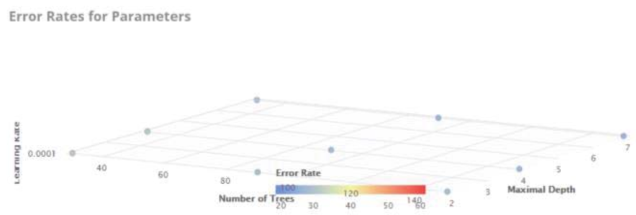

Figure 6 Error Rated Parameters Optimization Model

The TABLE 1. will show the optimization process of Decision Tree Model for finding the best results in prediction.

Table 1. Optimization Parameters

Number of Tree	Maximal Depth	Learning Rate	Error Rate
90	7	0.1	0.27422861
150	7	0.1	0.27463948
90	4	0.1	0.28077924
150	4	0.1	0.28078417
90	2	0.1	0.28531175
150	2	0.1	0.28824952
30	7	0.1	0.30174078
30	4	0.1	0.30525791
30	2	0.1	0.31198877

F. Performance Evaluation

Performance Evaluation is the process of evaluating the performance of the model used in testing. In this paper, the model's effectiveness was assessed using three different assessment methods as follows:

1) Root Mean Square Error (RMSE)

Root Mean Square is a Loss Function that takes Mean Square Error (MSE) values into the Square Root, it has properties that are like MSE values, but the difference is that the error unit has no squares So, so it's easier to read the value. This is because the units of the RMSE are the same as the values predicted by the model. which has in the equation (5).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{5}$$

Where $\hat{y}_i$ is predicted value
y_i is actual value

2) Confusion Matrix

Confusion Matrix is an essential table for measuring the performance or capability of machine learning. Finding the ratio of predicted data (Prediction Model) to actual data (Actual Value). In Table 2 will describe the confusion matrix in each dimension. Which has in the equation (6)

TABLE 2. Confusion Matrix

	Actually Positive (1)	Actually Negative(0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

$$Accuracy = \frac{TPs+TNs}{(TPs+TNs+FPs+FNs)} \tag{6}$$

where True Positive (TP) = what is predicted corresponds to what is actual in the case that the prediction is true and what is actual is true

True Negative (TN) = What is predicted is exactly what actual. If the prediction is False and what actual is False.

False Positive (FP) = What is predicted does not match what actual, is predicted to be true, but what actual is False.

False Negative (FN) = What is predicted does not match what actual. is a prediction that is False but what actual is true.

G. User Data Rate and Location Properties

User Data Rate and Location Properties uses a method of surveying user data based on the location of each point on the map.

H. User Classification

User Classification is a process for classifying user data by comparing the characteristics of the data to check user patterns and predicting the results using Ensemble Learning.

Ensemble Learning is the process of a machine learning model for combining similar or different machine learning classifiers to predict the output based on the highest majority. Figure 8 will show the process of using machine learning models resulting from predictions to create a Voting Algorithm for generating predictions. Recommendation System.

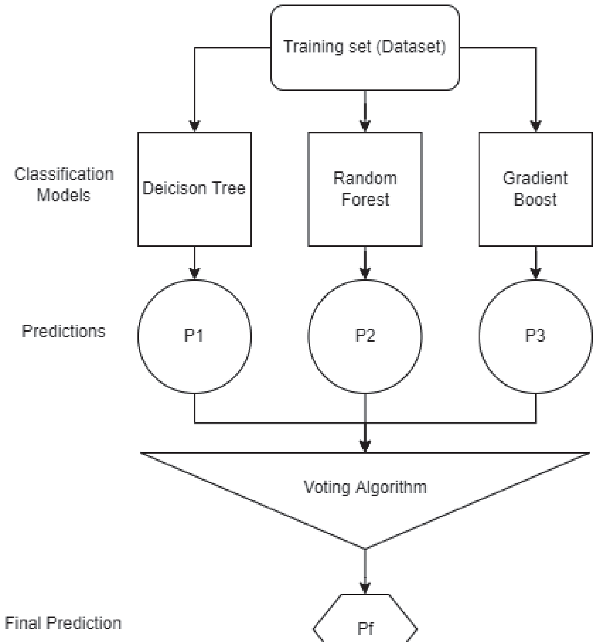

Figure 8 Voting Classifier Process

I. User Clustering

User Clustering is segmenting Received signal strength indicators (RSSI) which are designed to divide the data reception into 4 groups of RSSI using a statistic model and high internet speeds. 3 ranges with percentile.

A quartile is the statistic model for dividing data into four equal parts. Equations (7) (8) (9) show the result of the proportions in each quartile.

$$Lower\ Quartile\ (Q1) = \frac{(n+1)}{4} \quad (7)$$

$$Middle\ Quartile\ (Q1) = \frac{(n+1)}{2} \quad (8)$$

$$Upper\ Quartile\ (Q1) = (n+1) \times \frac{3}{4} \quad (9)$$

Where n is data points

Percentile is a method of dividing information into proportional parts that we define. This proposed will split the data into 33% and 66% to be used to share the data of User usage estimates.

J. User based

User based is to find the traffic of nearby users in an area that interests them. Then, they select the portion of the traffic and the transmission rate received from the antenna the user uses.

K. Location based

Location based finds the signal station in each tower from nearby users using Latitude and Longitude coordinates.

L. Recommendation System

The Recommendation System in this research proposed used the model obtained from the Voting Classifier to create a recommendation system.

1) Collaborative Filtering

Collaborative Filtering use a behavioral learning approach from other users that can be calculated from the equation (7)

$$\cos(\theta) = \frac{U \bullet L}{||U||||L||} = \frac{\sum_{i=1}^{n} U_i L_i}{\sqrt{\sum_{i=1}^{n} U_i^2}\sqrt{\sum_{i=1}^{n} L_i^2}} \quad (7)$$

where U_i are components of vector User Based
L_i are components of vector Location Based

IV. RESULTS AND DISCUSSION

1) RESULTS

In this purposed method. The data is organized and cleaned so that the data can be used. The modeling process evaluates the performance of the model for best results, with the results shown in Figure 9

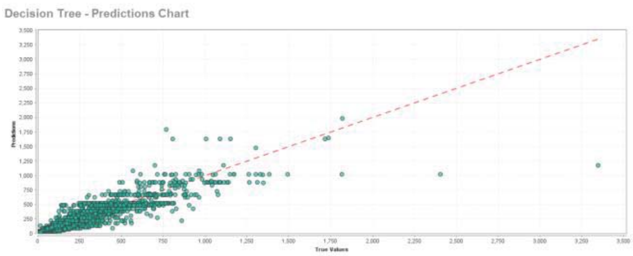

Figure 9(a) Prediction Decision Tree Model

in Figure 9(a), a Prediction Chart of the Decision Tree Model obtained from data entry Predicted results in correlation were 0.971 and RMSE were 0.316.

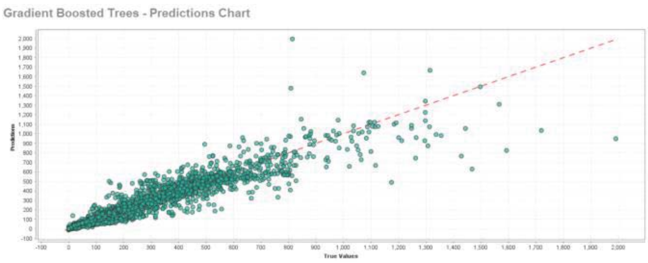

Figure 9(b) Prediction Gradient Boost Tree Model

in Figure 9(b), a Prediction Chart of the Gradient Boost Tree Model obtained from data entry Predicted results in correlation were 0.976 and RMSE were 0.291.

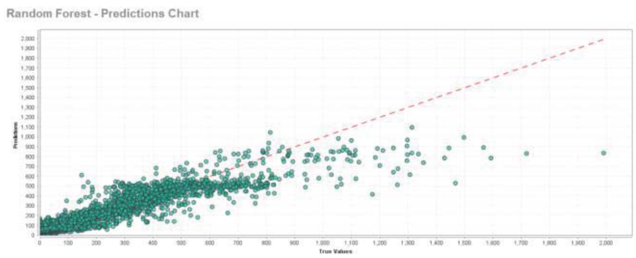

Figure 9(c) Prediction Random Forest Model

in Figure 9(b), a Prediction Chart of the Random Forest Model obtained from data entry Predicted results in correlation were 0.966 and RMSE were 0.346.

Table 3 will show the model's performance with all three models.The Gradient Boost Tree delivering the best results: RMSE 0.291, Correlation at 0.976

TABLE 3. Prediction Algorithm

Models	RMSE	Correlation	Accuracy
Gradient Boosting	0.291	0.976	0.978
Decision Tree	0.316	0.971	0.974
Random Forest	0.346	0.966	0.97

As a result of this, the prediction Decision Tree Model. The results from TABLE 4. That shown the prediction were achieved by RMSE 0.316, Correlation at 0.971 and Overall Accuracy Model were 97.49%. which in TABLE 5. will given the result of accuracy, precision, recall and F1 Score .

TABLE 4. Confusion Matrix Decision Tree

	Actual Class				
Predict Class	1506	0	354	0	0
	0	742	0	0	0
	103	0	6618	0	0
	0	0	0	1097	0
	0	0	0	0	7806

TABLE 5. Confusion Matrix Score of Decision Tree

Class	n (Truth)	n (Classified)	Accuracy (%)	Precision	Recall	F1-Score
1	1609	1860	97.49	0.81	0.94	0.87
2	742	742	100	1	1	1
3	6972	6721	97.49	0.98	0.95	0.97
4	1097	1097	100	1	1	1
5	7806	7806	100	1	1	1

As a result of this, the Prediction Gradient Boost Tree Model. The results from TABLE 6. That shown the prediction resulted in results by RMSE 0.291, Correlation at 0.976 and Overall Accuray Model 97.89%. which in TABLE 7. will given the result of accuracy, precision, recall and F1 Score .

TABLE 6. Confusion Matrix Gradient Boost Tree

	Actual Class				
Predict Class	1535	0	325	0	0
	0	742	0	0	0
	59	0	6662	0	0
	0	0	0	1097	0
	0	1	0	0	7806

TABLE 7. Confusion Matrix Score of Gradient Boost Tree

Class	n (Truth)	n (Classified)	Accuracy (%)	Precision	Recall	F1-Score
1	1594	1860	97.89	0.83	0.96	0.89
2	743	742	99.99	1	1	1
3	6987	6721	97.89	0.99	0.95	0.97
4	1097	1097	100	1	1	1
5	7806	7807	99.99	1	1	1

As a result of this, the Prediction Gradient Boost Tree Model. The results from TABLE 8. That shown the prediction resulted in results by RMSE 0.346, Correlation at 0.966 and Accuracy Model 97%. which in TABLE 9. will given the result of accuracy, precision, recall and F1 Score .

TABLE 8. Confusion Matrix Random Forest

	Actual Class				
Predict Class	1404	0	456	0	0
	0	742	0	0	0
	81	0	6640	0	0
	0	0	0	1097	0
	0	4	0	0	7803

TABLE 9. Confusion Matrix Score of Random Forest

Class	n (Truth)	n (Classified)	Accuracy (%)	Precision	Recall	F1-Score
1	1485	1860	97.05	0.75	0.95	0.84
2	746	742	99.98	1	0.99	1
3	7096	6721	97.05	0.99	0.94	0.96
4	1097	1097	100	1	1	1
5	7803	7807	99.98	1	1	1

As a result of three models for our prosposed to combining three machine learning classifier call ensemble learning methods results TABLE 10. That shown the prediction were produced by RMSE 0.312, Correlation at 0.972 and Accuracy Model 97.5%. which in TABLE 11. will given the result of accuracy, precision, recall and F1 Score

TABLE 10. Confusion Matrix ensemble learning

	Predict Class				
Actual Class	1478	0	382	0	0
	0	742	0	0	0
	62	0	6659	0	0
	0	0	0	1097	0

	0	1	0	0	7806

TABLE 11. Confusion Matrix Score of ensemble learning

Class	n (Truth)	n (Classified)	Accuracy (%)	Precision	Recall	F1-Score
1	1540	1860	97.56	0.79	0.96	0.87
2	743	742	99.99	1	1	1
3	7041	6721	97.56	0.99	0.95	0.97
4	1097	1097	100	1	1	1
5	7806	7807	99.99	0	1	1

In TABLE 12. on our test sample for every test user, our model generates the top 5 recommendations for making a recommendation.to monitor the usage of users in the area. It is also used to adjust the size of the signal distribution of nearby antennas to increase the efficiency in supporting users' use in the area. With high density.

TABLE 12. Comparison of Method

Method	Matching Average (%)	Confident
Proposed	**90.16**	**0.835**
Collaborative Filtering	86.6	0.816
Content-based Filtering	83.24	0.798

2) *DISCUSSION*

Discussing the results will be discussed in three main sections discussing modelling results. The first section discusses the application of ensemble learning methods to build a Recommender System.

The process of accessing the user's usage characteristics to adjust the size of the signal and the user's usage data pattern. And a modelling process for predicting signal strength to solve the problem of user inaccessibility due to heavy usage. And the last part is to improve the service quality of service providers by reducing or increasing the antenna's transmission power according to the density of users.

An ensemble learning method derived from signal strength. Prediction modelling creates a Recommender System using user data in each area to determine the local signal strength. Is it enough for users to verify the signal transmission in that area? How does It adjust the size of the antenna's transmission power?

The modelling process and the signal intensity prediction are designed to be used to create a recommendation system that will be useful. It is divided into two parts. Internetwork because service providers are not scaling up to support users in dense areas. The broadcast signal is, therefore, inaccessible—the quality of Service (QoS) to everyone. The system recommends that Internet Speed Providers (ISPs) adjust the transmission power of cell towers in certain areas. That uses fewer users to reduce power consumption and increase the transmission power of the antennas in regions. A high density makes it more convenient for users and reduces traffic congestion. The second part will be to create a system to advise Internet Speed Providers (ISPs) to know growth rates. In the area, they are currently serving to allow ISPs to increase the number of small cell towers. The power of the cell towers to promptly support the number of users in different areas

V. CONCLUSIONS AND FUTURE WORKS

This research proposed a method to predict the rate of traffic. Which Data is used as data from the Kaggle public website, splitting data into five sections for handling missing value data by interpolating and transforming the data into the same format.

After that, the data for evaluating the model's performance and converting the data for use in the model performance evaluation process is compared, and the performance of the models used in testing and evaluation is compared. It also improves model performance to minimize errors. In this proposed method, comparing the performance of models, the results were: The RMSE Gradient Boost, Decision Tree, and Random Forest were 0.291, 0.316 and 0.346, respectively. The correlation will be 0.976, 0.971, and 0.966 for Gradient Boost Tree, Decision Tree, and Random Forest, respectively, and the accuracy of Gradient Boost, Decision Tree, and Random Forest were 97.8%, 97.4% and 97%, respectively. The results of ensemble learning methods, the RMSE, Correlation, and accuracy were: 0.312, 0.972 and 97.5%

The recommendation is built to monitor users' traffic at each tower to adjust the size of the signal distribution to reduce the burden on a single tower. It also allows users to have uninterrupted access to the signal. The matching average of the proposed method, collaborative filtering, and content-based filter was 90.16, 86.6 and, 83.24. the confidence was: 0.835, 0.816 and, 0.798.

Furthermore, the implement of our recommendation system that classify has derived from the model and model it for managing additional antenna positioning for ISPs to optimize Quality of Service (QOS) and managing the power consumption to the antenna's

ACKNOWLEDGMENT

This work was financial supported by Research Grant of Burapha University and Faculty of Informatics Burapha University.

REFERENCES

[1] J. Thomas Barnett, "Cisco Visual Networking Index (VNI) Global and Americas/EMEAR Mobile Data Traffic Forecast, 2017–2022," Cisco, 2019.

[2] Y. Egi, E. Eyceyurt, I. Kostanic, C. E. Otero, "An Efficient Approach for Evaluating Performance in LTE Wireless Networks," Las Vegas, 2017

[3] 5GAA, "Making 5G proactive and predictive for the automotive industry," 5GAA, White Paper, Jan. 2020.

[4] Eyceyurt, Engin, Yunus Egi, and Josko Zec. "Machine-Learning-Based Uplink Throughput Prediction from Physical Layer Measurements." Electronics 11.8 (2022): 1227.

[5] Eyceyurt, Engin, and Josko Zec. "Uplink throughput prediction in cellular mobile networks." International Journal of Electronics and Communication Engineering 14.6 (2020): 149-153.

[6] Elsherbiny, Habiba, et al. "4g lte network throughput modelling and prediction." GLOBECOM 2020-2020 IEEE Global Communications Conference. IEEE, 2020.