UDC 621
DOI: 10.15587/1729-4061.2022.259242

DEVELOPMENT OF PREDICTIVE MODELING AND DEEP LEARNING CLASSIFICATION OF TAXI TRIP TOLLS

Suhad Al-Shoukry
Corresponding author
Lecturer
Department of Computer Systems Techniques
AL-Najaf Technical Institute
AL-Furat Al-Awsat Technical University
Babylon-najaf str, Kufa, AL-Najaf, Iraq, 54003
E-mail: inj.suhadaalzhra2010@atu.edu.iq

Bushra Jaber M. Jawad
Assistant Lecturer
Department of Accounting
College of Administration and Economics
University of Kerbala
Kerbala, Iraq, 56001

Zalili Musa
Senior Lecturer, Doctor of Communication Engineering
Department of Computing
Universiti Malaysia Pahang
Pekan, Pahang, Malaysia, 26600

Ahmad H. Sabry
Doctor of Control and Automation Engineering
Department of Sustainable Energy
Universiti Tenaga Nasional
Jl. Ikram-Uniten, Kajang, Selangor, Malaysia, 43000

Several studies discussed the predictive modeling of deep learning in different applications such as classifying tissue features from microstructural data, Crude Oil Prices, mechanical constitutive behavior of materials, microbiome data, and mineral prospectively. Commercial navigation includes a wealth of trip-related data, including distance, expected journey time, and tolls that may be encountered along the way. Using a classification algorithm, it is possible to extract drop-off and pickup locations from taxi trip data and estimate if the tour would incur tolls. In this work, let's use the classification learner to create classification models, compare their performance, and export the findings for additional study. The workflow for the classification learner is the same as for the regression learner. The purpose is to make predictions based on fresh data in order to see how well the model performs with new data. To train the model, it's critical to separate the data set. The combined training and validation data is next pre-processed, which involves tasks such as cleaning and developing new features skills. Once the data has been prepared, it's time to begin the supervised machine learning process and test a number of ways to identify the best model, such as the type of model that should be used, the important features, and the best parameters of the model to find the best fit for the considered data. The results of analyzing different predictive multiclass classification models with taxi trip tolls show that it is possible to use a machine learning-based model when we like to avoid road tolls depending on historical data on taxi trip tolls. The outcome of this study can help to expect road tolls from the drop-off and pickup locations of a taxi data

Keywords: machine learning, deep learning, multiscale classifications, Taxi trips tolls, Prediction

Received date 06.05.2022
Accepted date 06.06.2022
Published date 30.06.2022

How to Cite: *Al-Shoukry, S., M. Jawad, B. J., Musa, Z., Sabry, A. H. (2022). Development of predictive modeling and deep learning classification of taxi trip tolls. Eastern-European Journal of Enterprise Technologies, 3 (3 (117)), 6–12. doi: https://doi.org/10.15587/1729-4061.2022.259242*

1. Introduction

Deep learning is utilized in a wide range of applications, from identifying disease risk factors to developing superior automotive safety systems. The purpose of supervised deep learning is to create a predictive model from data that includes a collection of attributes as well as the known response for each observation as shown in Fig. 1.

Several studies discussed the predictive modeling of deep learning in different application such as classifying tissue features from microstructural data [1], Crude Oil Price [2], mechanical constitutive behavior of materials [3], microbiome data [4], Brain Tumor [5], mineral prospectively[6], analyze cardiac electrophysiology data [7], forecast the potential of patients from the electronic health measurements [8], medical records (EHR) [9, 10], tensile strength prediction in fused deposition modeling [11], and bridge vortex-induced vibrations from field monitoring [12]. Metal deformation has been studied using traditional simulations based on the crystal plasticity finite element approach. The deep learning-based methods were proposed also for predicting macroscopic attributes based on microstructure features with low human bias. The model can anticipate property against a given structure in the dual phase, isotropic elastic-plastic regime [13]. Machine learning techniques are increasingly being used to identify patterns and insights from the growing stream of geospatial data, however they may not be appropriate when system behavior is influenced by geographical or temporal context [14]. For these approaches to be useful, they must have extremely high accuracy and low false-negative rates. One example of a Google map showing the estimated travel time, trip distance, and tolls that might meet on routes is shown in Fig. 2.

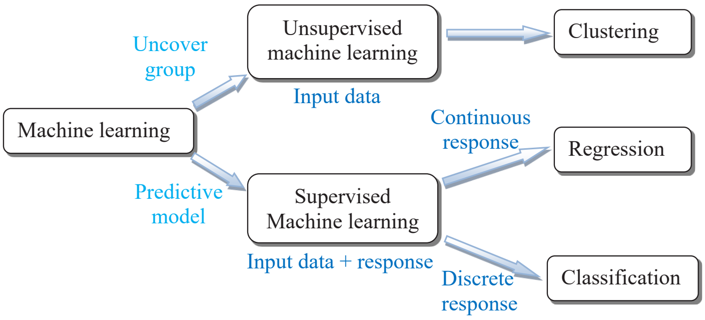

Fig. 1. Types of machine learning

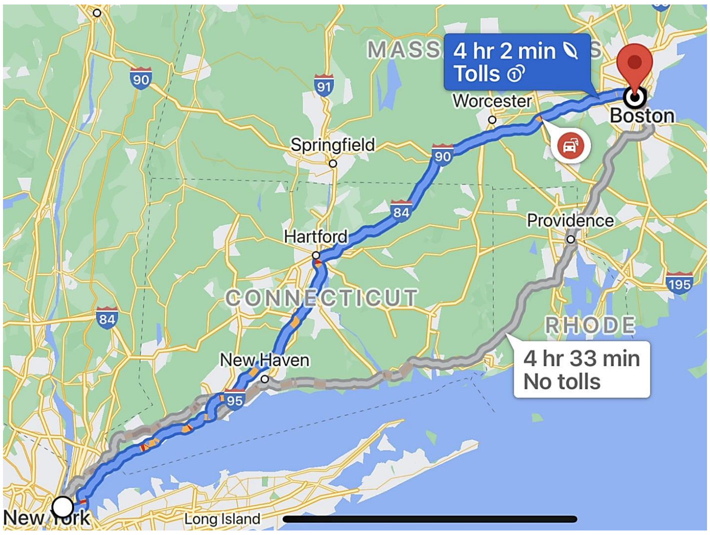

Fig. 2. One example of Google map showing the estimated travel time, trip distance, and tolls that might meet on routes [15]

Deep learning may have a revolutionary impact on how let's simulate the constitutive features of soft biological tissues, according to this finding [1]. Therefore, the importance of such a study can help to expect if the journey will include a toll from drop-off and the pickup locations of taxi data using binary or multi-classification models.

2. Literature review and problem statement

The study [1] proposed a model for predicting mechanical properties of vascular tissue features from microstructural data by developing a hybrid modeling framework that blends advanced theoretical notions with deep learning. This hybrid modeling system is only trained with data from 27 tissue samples. Although this study obtains a median coefficient of determination of 0.97, it was limited to tissue samples with mechanical properties in the range usually observed. The paper [2] presented modules of outlier detection, recurrence analysis, data preprocessing, predictive modeling, and feature selection based on deep learning with the goal of obtaining probabilistic and deterministic predictions to model the nonlinear dynamics in crude oil price. However, the presented model can make only accurate probabilistic and deterministic predictions with a narrow application range and feasibility. Predictive data-driven constitutive modeling by deep learning was proposed by [3] for mechanical constitutive behavior of materials where no stress-strain data are available, but this study fails to compare the model with other prior studies. In microbiome data application, the study [4] built a prediction model for clinical outcomes based on microbiome data that can predict both binary and continuous outcomes. The presented model has been applied in both binary and regression classification but with a complex and training time-consuming network structure. The transfer learning-based predictive model of paper [5] was applied to detect the growth of malignant tissue by looking at a patient's brain magnetic resonance imaging (MRI) data using three pre-trained models. Although the results were compared to one another, the study didn't show and compare with more models. This issue is explored by [6] when presented data-driven predictive models including a series of machine learning approaches were discussed but with limited input datasets of only 118 known occurrences derived from long-term exploration of this brownfield area. According to the modeling findings, the CNN model obtains the best classification performance with a 92.38 percent accuracy, followed by the RF model (87.62 percent). However, this model can be satisfied in other applications. The paper [8] offered a unique unsupervised deep feature learning technique for generating a general-purpose patient representation from electronic health records data. Although this study used data of around 700,000 patients and improved clinical predictions, the offered system was not accurate enough for other applications.

According to the studies mentioned above, data-driven-based machine learning is a promising technology for various classification, especially in predictive modeling. Therefore, this motivates to development of a predictive model to estimate taxi trip tolls accurately with an appropriate classification algorithm, which is not addressed as an application in the prior studies.

3. The aim and objectives of the study

The main aim of the study is to develop predictive modeling and deep learning classification of taxi trip tolls.

The following aim have been set to achieve the objectives:

– to analyze different predictive multiclass classification models with taxi trip tolls;

– to obtain the best accurate predictive binary classification model for the same data.

4. Materials and methods

The data that is considered in this work is the taxi data, which is commercial navigation data that offer several information concerning the trip distance, tolls that may meet on routes, and estimated travel time. The considered workflow to evaluate regression and classification models is shown in Fig. 3.

The initial step is to explore and import the data whether the data is enough before beginning to develop a deep learning model. The purpose is to make predictions based on fresh data in order to see how well the model performs with new data. It is essential to divide the dataset to train the model. This dataset contains training and validation data, the remaining data are referred to as test data, in which the new observations (has never been seen) are used to be simulated in the final model. Let's locate the assessment dataset to the side until the final model is obtained.

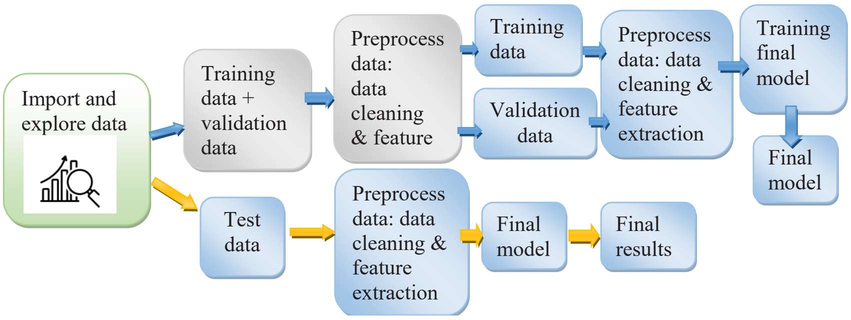
Import and explore data
Training data + validation data
Preprocess data: data cleaning & feature
Training data
Validation data
Preprocess data: data cleaning & feature extraction
Training final model
Final model
Test data
Preprocess data: data cleaning & feature extraction
Final model
Final results

Fig. 3. Supervised deep learning workflow

Next, pre-processing the validation data and combined training, which contains generating new features and cleaning skills. Once the data is prepared, the procedure of supervised deep learning is started and trying to find the best model by trying a range of methods including the important features, the kind of model need to use, the method to find the parameters of the optimal model to obtain the best fit for the considered data. The final step is to use the test data to obtain the model to predict results.

Commercial navigation includes a wealth of trip-related data, including distance, expected journey time, and tolls that may be encountered along the way. Using a binary classification algorithm, it is possible to extract drop-off and pickup locations from taxi trip data and estimate if the tour would incur tolls. MATLAB-based classification learner is employed to follow for achieving the objectives.

According to the above, let's first import and prepare the data into the proposed model, the data in this study is the January taxi data [16].

Let's consider only three columns from the dataset that has a (76518*3) dimension. The response variable is the third column represented by the total tool paid.

5. Results of predictive modeling with deep learning of taxi trip tolls

5. 1. Multiclass classification models

The scatter plot showing the distribution of the data pickup and drop off locations and grouped according to the toll paid variable is shown in Fig. 4.

The blue dots point out that zero tolls were paid. Initially, a classification type of Fine Tree model was used to train the data, and it has over (74 %) accuracy. In order to provide more clearance on the data, a confusion matrix is shown in Fig. 5.

In order to show the minimum classification error over 30 iterations, a plot of this error along with the number of iterations is shown in Fig. 6.

Because of the imbalance in class sizes, only approximately ten of the actual tolls are correctly predicted by the model. Out of the over 76518 journeys, fewer than 7000 had a toll. Because tolls are used on just around 36 % of trips, a model that always forecasts no toll would have 74 % accuracy. It would simply be ineffective. Models like logistic regression and SVM have a hard time dealing with datasets that have an uneven distribution of classes as described in Fig. 7.

Several attempted models have been conducted that potentially obtain more accuracy than the Fine Tree classification model. It is found that some models have to scatter plots with significantly fewer visible x's as listed in Table 1.

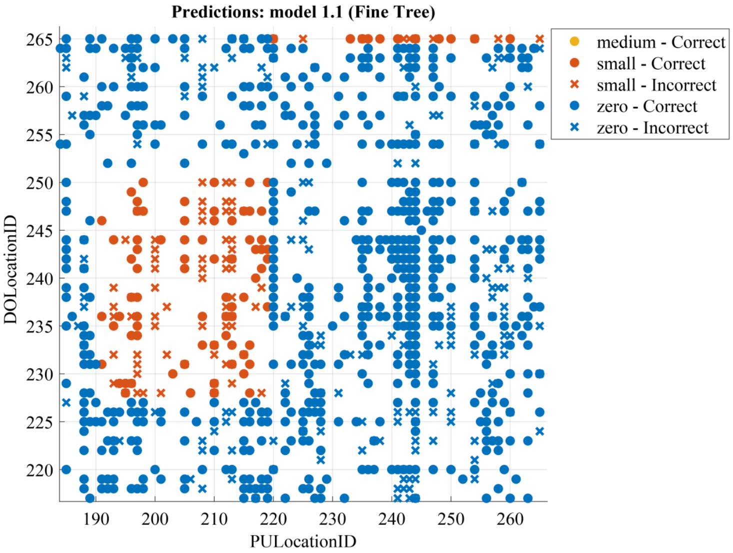

Fig. 4. The scatter plot shows the distribution of the data pick-up and drop-off locations along with the toll paid variable

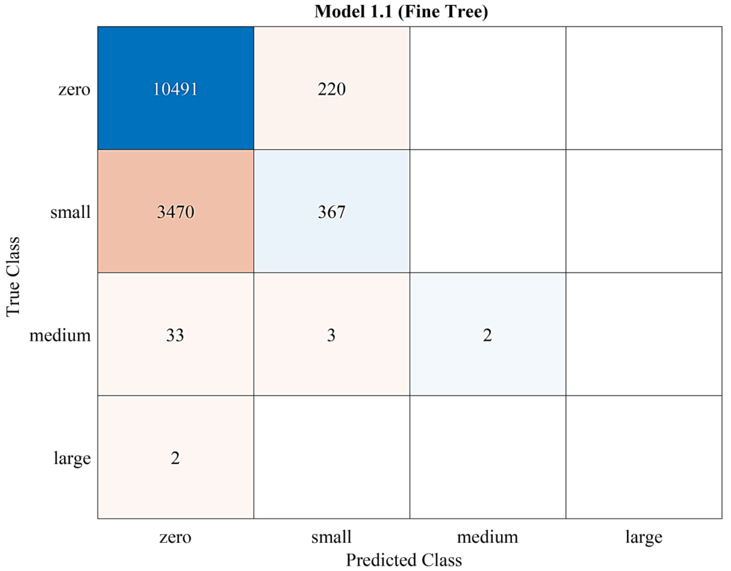

Fig. 5. The confusion matrix shows four classes when the Fine Tree classification model was used

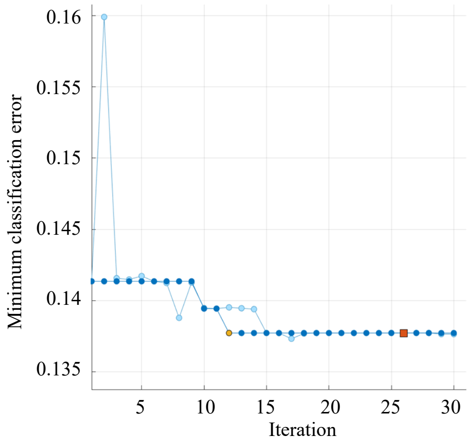

Fig. 6. The minimum classification error over 30 iterations for Tree classification models

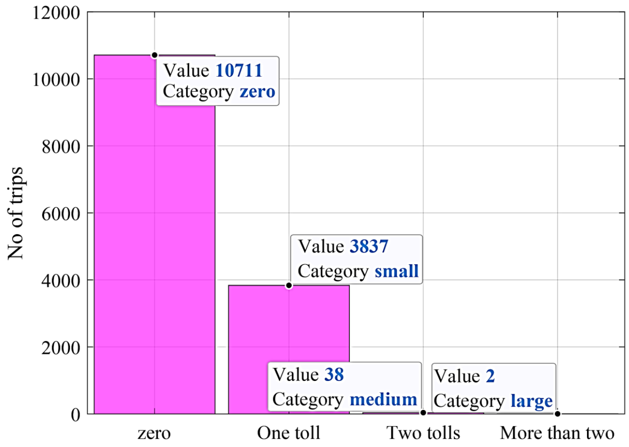

Fig. 7. Distribution of the taxi trips data

Table 1

Classification models with their key parameters for evaluation

Model Type:	Accuracy (Validation) (%)	Total cost (Validation)	Prediction speed (obs/sec)	Training time (sec)	Maximum number of splits
Fine Tree	74.4	3728	~460000	7.7861	100
Medium Tree	73.6	3849	~230000	1.7407	20
Coarse Tree	73.4	3877	~1300000	0.86175	4
Linear Discriminant	73.4	3877	~550000	1.5393	NA
Gaussian Naive Bayes	73.4	3877	~570000	1.3596	NA
Kernel Naive Bayes	73.4	3877	~1100	42.186	NA
Cubic SVM	26.2	10770	~37000	531.8	NA
Fine Gaussian SVM	73.4	3877	~4100	23.58	NA
Medium Gaussian SVM	73.4	3877	~6700	13.648	NA
Coarse Gaussian SVM	73.4	3877	~7200	12.265	NA
Fine KNN	94.1	854	~300000	1.3836	1
Medium KNN	75.5	3569	~73000	1.7005	10
Coarse KNN	73.4	3877	~41000	1.8231	100
Cosine KNN	73.9	3814	~12000	4.4402	10
Cubic KNN	75.5	3572	~79000	1.5247	10
Weighted KNN	94.4	824	~190000	0.90813	10
Boosted Trees	73.5	3862	~65000	3.2712	20
Bagged Trees	94.0	873	~36000	4.3785	NA
Subspace Discriminant	73.4	3877	~60000	2.2534	NA
Subspace KNN	72.1	4066	~16000	4.746	NA
RUSBoosted Trees	40.3	8704	~92000	2.0928	20
Narrow Neural Network	73.4	NA	~830000	16.34	NA
Medium Neural Network	73.4	NA	~490000	18.433	NA
Wide Neural Network	73.4	NA	~620000	43.32	NA
Bilayered Neural Network	73.4	NA	~1100000	18.947	NA
Trilayered Neural Network	73.4	NA	~740000	21.269	NA

The models with accuracy more than 90 % look promising. It is found that the KNN models could be a good selection for such forecast.

5. 2. Binary classification models

In order to inspect the feasibility of using classification models to split between taxi trips with and without toll payment, a binary classification models are used. Therefore the data has been categorized logically and can be shown in Fig. 8.

Here let's show only the best scatter plot in terms of accuracy showing the distribution of the data pickup and drop off locations and grouped according to the toll paid variable. This is shown in Fig. 9.

The blue dots refer to zero tolls that were paid during the tour. A classification type of Weighted KNN model was used and consumed about 1.2835 sec to train the data with (94.4 %) accuracy. The group scatters plot employs markers to show where the model accurately anticipated that a toll will be paid. Dots represent right predictions, whereas (x's) represent wrong predictions. The total cost (Validation) was 813 with a prediction speed of ~130000 Observation/sec). In order to provide more clearance on the data, a confusion matrix is shown in Fig. 10.

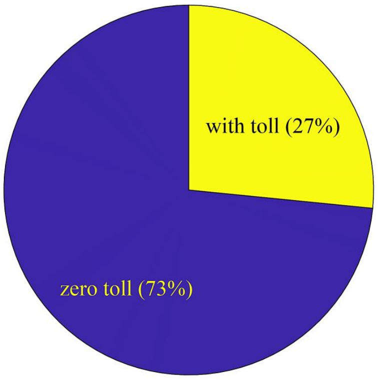

Fig. 8. The binary distribution of the taxi trips data according to the toll paid

The minimum classification error over 30 iterations for KNN binary classification model is shown in Fig. 11.

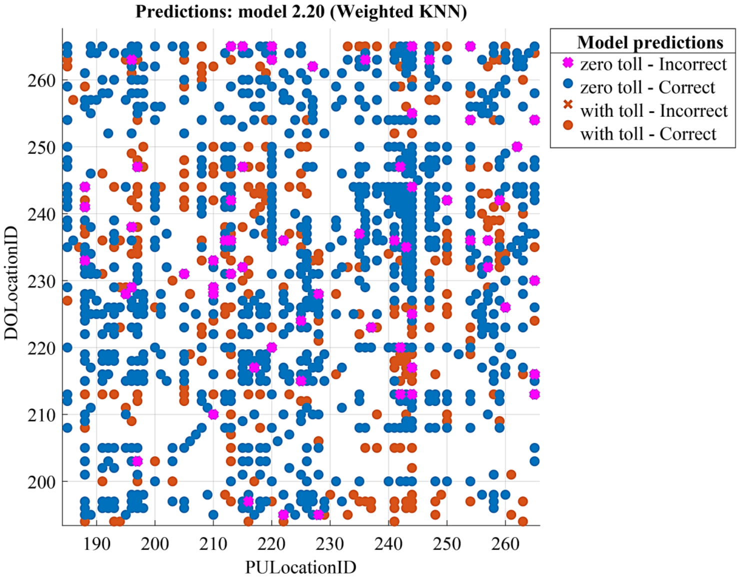

Fig. 9. The scatter plot showing the distribution of the data pickup and drop off locations and grouped according to the toll paid a variable for binary classification

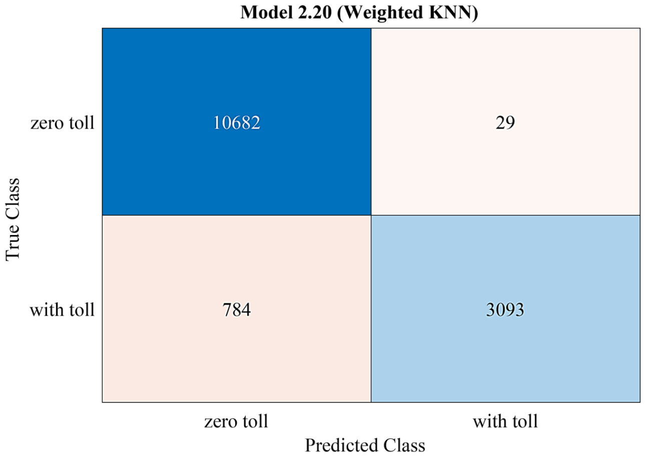

Fig. 10. The confusion matrix showing four classes when Fine Tree classification model was used for binary classification

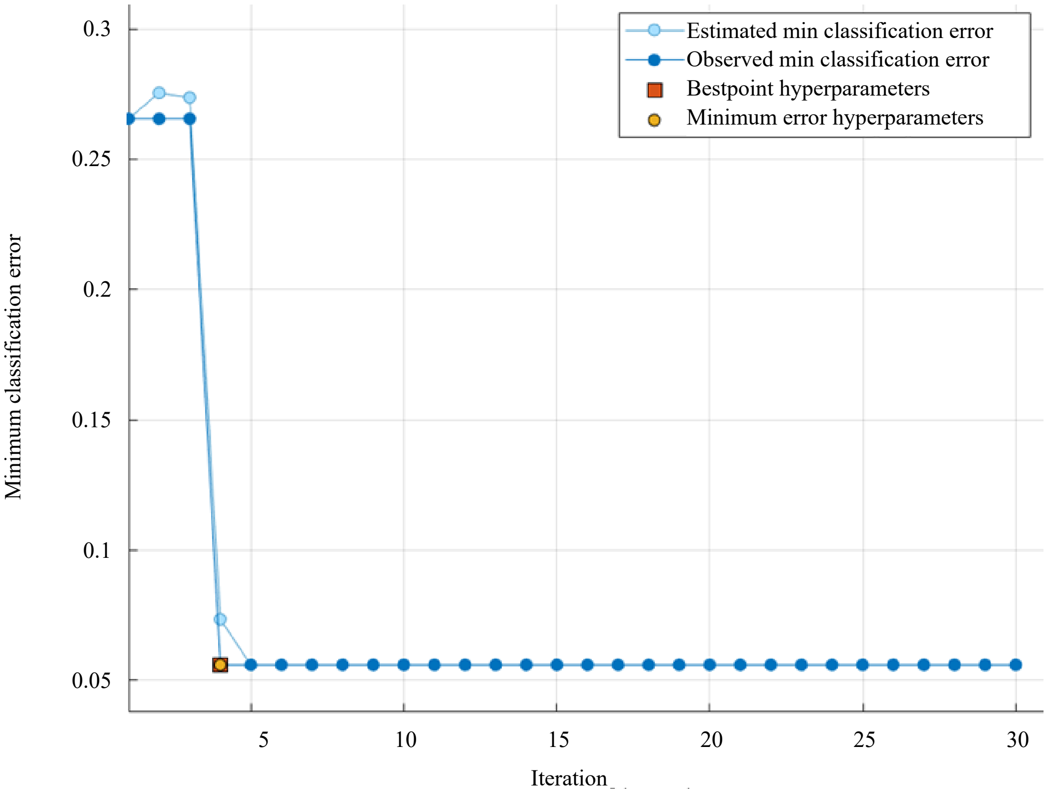

Fig. 11. The minimum classification error over 30 iterations for Optimizable KNN model

Fig. 11 shows about 0.052 of minimum classification error has been obtained using Weighted KNN binary classification model with the considered Green taxi trips data.

6. Discussion of the results of the developed predictive model of taxi trip tolls

Referring to Fig. 4, the total cost (Validation) was 3728 with a prediction speed of (~460000 Observation/sec). The training time with this model was (7.7861 sec) with a maximum number of splits of 100. More particularly blue (x's) specify wherever there was no toll paid but the model estimated a toll, and orange (x's) indicate where a toll is paid but the model didn't predict it, which is also shown in Fig. 9.

Fig. 5 provided the confusion matrix to show a more good indication of the modeled data distribution. This graph displays how often a model gets each class correct and which class is chosen when a data point is misclassified. The rows represent the data set's value of (tool paid). The top row of the matrix depicts travels without a toll, while the below rows depict trips with a toll ranging from a little to a big paid sum. The model predictions are shown by the columns. The earliest column shows travels for which the model projected there would be no tool, while the last column shows trips for which the model anticipated there would be a tool. The trips where the model properly predicted the reaction, either true negatives where there was no toll and the model projected no toll or directed to true positives where there was a tool and the model predicted a tool, are the primary diagonal elements. The off-diagonal entries indicate where the model predicted incorrectly. The same concept is applied in Fig. 10.

For the confusion matrix of the Weighted KNN binary classification model (Fig. 10), it is possible to see that 10682 observations have been classified as no toll (zero tolls) correctly. 784 (with paid toll) classes have been wrongly classified as (zero tolls) and 29 (zero tolls) classes have been wrongly classified as (with paid toll). 3093 (with paid toll) classes have been correctly classified as (with paid toll).

In this case, a false positive could be a good surprise for individuals who expected a toll but did not have to pay one, but a false negative would mean a surprising toll, which would be bothersome if there were too many of them.

Limitations of such a model as getting an accuracy of less than 100 % is being when representing for example a disease data, a false positive could be frightening and result in additional testing, while a false negative could result in a serious condition staying undiscovered, delaying treatment options. When it comes to judging when false positives and false negatives are acceptable, your domain knowledge will help you. It's possible that minimizing one or the other is crucial. We'll need more performance metrics to figure out how good we are.

7. Conclusions

1. The results of analyzing different predictive multi-class classification models with taxi trip tolls show that it is possible to use a machine learning-based model when we like to avoid road tolls depending on historical data on taxi trip tolls. The tree model was used to train the data, and it has over (74 %) accuracy. The models with an accuracy of more than 90 % look promising. It is found that the KNN models could be a good selection for such a forecast.

2. The Weighted KNN model is the best accurate predictive model in both binary and multiclass classification for the same data.

A classification type of Weighted KNN model was used and consumed about 1.2835 sec to train the data with (94.4 %) accuracy.

Acknowledgments

I would like to acknowledge and give my warmest thanks to the Department of Computer system Techniques, AL-Najaf Technical Institute, AL-Furat Al-Awsat Technical University, AL-Najaf, Iraq.

References

1. Holzapfel, G. A., Linka, K., Sherifova, S., Cyron, C. J. (2021). Predictive constitutive modelling of arteries by deep learning. Journal of The Royal Society Interface, 18 (182), 20210411. doi: https://doi.org/10.1098/rsif.2021.0411
2. Niu, T., Wang, J., Lu, H., Yang, W., Du, P. (2021). A Learning System Integrating Temporal Convolution and Deep Learning for Predictive Modeling of Crude Oil Price. IEEE Transactions on Industrial Informatics, 17 (7), 4602–4612. doi: https://doi.org/10.1109/tii.2020.3016594
3. Linka, K., Hillgärtner, M., Abdolazizi, K. P., Aydin, R. C., Itskov, M., Cyron, C. J. (2021). Constitutive artificial neural networks: A fast and general approach to predictive data-driven constitutive modeling by deep learning. Journal of Computational Physics, 429, 110010. doi: https://doi.org/10.1016/j.jcp.2020.110010
4. Wang, Y., Bhattacharya, T., Jiang, Y., Qin, X., Wang, Y., Liu, Y. et. al. (2020). A novel deep learning method for predictive modeling of microbiome data. Briefings in Bioinformatics, 22 (3). doi: https://doi.org/10.1093/bib/bbaa073
5. Saxena, P., Maheshwari, A., Maheshwari, S. (2020). Predictive Modeling of Brain Tumor: A Deep Learning Approach. Innovations in Computational Intelligence and Computer Vision, 275–285. doi: https://doi.org/10.1007/978-981-15-6067-5_30
6. Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z., Hu, Z. (2020). Data-Driven Predictive Modelling of Mineral Prospectivity Using Machine Learning and Deep Learning Methods: A Case Study from Southern Jiangxi Province, China. Minerals, 10 (2), 102. doi: https://doi.org/10.3390/min10020102
7. Cantwell, C. D., Mohamied, Y., Tzortzis, K. N., Garasto, S., Houston, C., Chowdhury, R. A. et. al. (2019). Rethinking multiscale cardiac electrophysiology with machine learning and predictive modelling. Computers in Biology and Medicine, 104, 339–351. doi: https://doi.org/10.1016/j.compbiomed.2018.10.015
8. Miotto, R., Li, L., Kidd, B. A., Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports, 6 (1). doi: https://doi.org/10.1038/srep26094
9. Sun, M., Tang, F., Yi, J., Wang, F., Zhou, J. (2018). Identify Susceptible Locations in Medical Records via Adversarial Attacks on Deep Predictive Models. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. doi: https://doi.org/10.1145/3219819.3219909
10. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M. et. al. (2018). Scalable and accurate deep learning with electronic health records. Npj Digital Medicine, 1 (1). doi: https://doi.org/10.1038/s41746-018-0029-1
11. Zhang, J., Wang, P., Gao, R. X. (2019). Deep learning-based tensile strength prediction in fused deposition modeling. Computers in Industry, 107, 11–21. doi: https://doi.org/10.1016/j.compind.2019.01.011
12. Li, S., Laima, S., Li, H. (2021). Physics-guided deep learning framework for predictive modeling of bridge vortex-induced vibrations from field monitoring. Physics of Fluids, 33 (3), 037113. doi: https://doi.org/10.1063/5.0032402
13. Beniwal, A., Dadhich, R., Alankar, A. (2019). Deep learning based predictive modeling for structure-property linkages. Materialia, 8, 100435. doi: https://doi.org/10.1016/j.mtla.2019.100435
14. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566 (7743), 195–204. doi: https://doi.org/10.1038/s41586-019-0912-1
15. How to use less gas when driving with Google Maps. Popular Science. Available at: https://www.popsci.com/diy/fuel-efficient-route-google-maps/
16. TLC Trip Record Data. Available at: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page