UNIVERSITY OF WATERLOO
Faculty of Mathematics

Rut Depth Modelling Using Random Forest

University of Waterloo - Centre for Pavement and Transport Technology
,

Sagun Malwatkar
Spring 2022 - B.Math. Mathematics / Financial Analysis and Risk Management
ID 20812957
August 18, 2022

Contents

List of Figures

1.0 Introduction

The Centre for Pavement and Transport Technology (CPATT) consists of specialists from industry and transportation authorities and academia, forming a partnership between universities, the public, and the private sector. CPATT's work deals with studying the preservation and replacement of Canada's public infrastructure, including the structural design, construction, and maintenance technology, materials, and geotechnical engineering, field evaluation methods, equipment and data processing, intelligent transportation systems, and safety, as well as risk and reliability methods. (About the Centre for Pavement and Transportation Technology, 2012) This report will mainly discuss the data processing performed to predict a specific aspect of pavement deterioration, i.e., rutting. Rutting is described as longitudinal deformation that occurs in flexible pavement. It decreases the life span of the road and can result in safety issues. The three primary factors that result in rutting are the asphalt layer, structural layer, and weak subgrade issues. (Naiel, 2010) The random forest machine learning

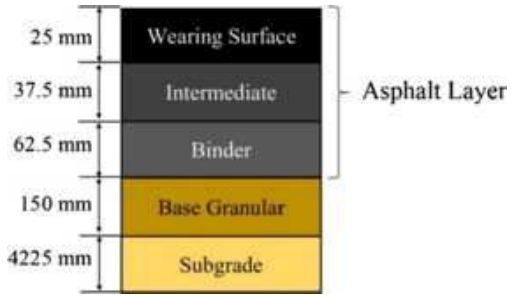


Figure 1: Pavement Layers

model was used to predict rutting. Data was extracted from the LTPP database. The LTPP database is a supply of extensive information about pavements including factors that may affect pavement performance. (LTPP FAQs — FHWA, n.d.)

2.0 Data Preprocessing

2.1 Initial Selection of Variables

This research aims to develop a Random Forest model to predict the rutting of flexible pavements. Several factors affect pavement performance that requires an understanding of pavement design and functional and structural properties. The independent variables were initially selected based on a literature review of previous studies using the LTPP data and the availability and limitation of the LTPP data and engineering knowledge. (Naiel, 2010) Pavement rut depth was selected to be the dependent (target) variable for this model and this data was extracted from the MON_T_PROF_INDEX_SECTION in the LTPP database. (LTPP InfoPave - Home, n.d.) The main variables selected dealt with traffic loads, temperature, precipitation and climate index, subbase thickness, subgrade material type, asphalt content, voids in the mineral aggregate, and air voids in the mix since it is difficult to address all the factors that contribute towards rutting.

The two types of traffic data used were historical traffic data and monitored traffic data. Since historical traffic data consisted of traffic data from the original construction date to 1990, monitored traffic data was additionally used as it provided annual estimates after 1990 that were either computed from raw data or provided by the concerned highway agency. These were the fields ANL_KESAL_LTPP_LN_YR in the tables TRF_HIST_EST_ESAL and TRF_MON_EST_ESAL respectively. (LTPP InfoPave - Home, n.d.; Naiel, 2010)

The environmental fields used were total annual precipitation, freeze index, maximum annual temperature average and days, and average number of days above 32 °C. These fields were taken from the tables CLM_VWS_PRECIP_ANNUAL for precipitation information and CLM_VWS_PRECIP_ANNUAL for temperature information. (LTPP InfoPave - Home, n.d.)

A better model can be made with more data points and hence, the three fields (air voids in the mix, asphalt content, voids in the mineral aggregate) were removed with subgrade

thickness, material type of subgrade, and characterization of subgrade stiffness as the only pavement layer properties to be included. Subgrade stiffness expresses the ability of the subgrade material to stand the traffic loads. The feature used was RES_MOD_AVG, i.e., resilient modulus which is a characterization of subgrade material stiffness. from the table TST_UG07_SS07_WKSHT_SUM. The features MR_MATL_TYPE (subgrade material type) and REPR_THICKNESS (layer thickness) were extracted from the table TST_UG07_SS07_A and TST_L05B respectively. (LTPP InfoPave - Home, n.d.)

CONSTRUCTION_NO indicates the number of rehabilitation and maintenance performed in the test section. The test section with CONSTRUCTION_NO 1 means that this section is not rehabilitated or maintained. Therefore, when the test section is maintained or rehabilitated the CONSTRUCTION_NO will increase by 1. (LTPP FAQs — FHWA, n.d.; Naiel, 2010) This feature was used along with SHRP_ID and STATE_CODE as keys to join the multiple tables.

2.2 Data Acquisition and Processing

The raw data was collected from the Long-Term Pavement Performance (LTPP) database which consists of information about pavement test sections in North America, including performance, traffic, weather, maintenance, and materials used. (LTPP FAQs — FHWA, n.d.; Naiel, 2010) First, the required data elements and the tables that included them were identified, and then a master spreadsheet was created with SHRP_ID, STATE_CODE, and CONSTRUCTION_NO as keys.

The tables mentioned in the previous section were downloaded using the Table Extraction tool on the LTPP website. Using MySQL and MySQL Workbench or alternatively, Python and Jupyter Notebook, all the tables were combined using an inner join (SQL function or Pandas merge function depending on the mode used). Then using Pandas and Numpy the data was further preprocessed. First, only the needed variables that were discussed in the

previous section and the variable keys that identified each data entry were kept by simply filtering out the unnecessary fields and reassigning the dataframe.

Exploratory univariate analysis was conducted to obtain descriptive data for all the variables required for the model. A statistical summary was obtained for each variable which helped determine any outliers or non-typical data points which were then excluded. (Naiel, 2010) The Pandas describe function was used to obtain this statistical summary which gives the count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile and maximum. The unique values for each field were also obtained to check for NaN values or any other inconsistencies. The empty data points were either filled with 0, the average of the feature, or the feature was completely excluded depending on the percent of missing values. The three factors - asphalt content, voids in the mineral aggregate, and air voids in the mix were removed from the final list of variables due to the lack of variation in the observations and due to significantly fewer observations. These fields had the highest percentage of missing values (95.4%) and were excluded from the data. Having these features in the data resulted in only 991 data points after inner joins on the tables which is not enough to create a good machine learning model. The five-number statistical summary (minimum value, first quartile, median, third quartile, maximum value) did not reveal any outliers in the data. Upon analyzing the correlation between the variables using the corr method and seaborn heatmap, the variable list was further decreased so that the independent variables have low collinearity between them. (Marcelino et al., 2021)

The final dataset had 7875 data points with 10 input features and 1 label. The data is then separated into two separate dataframes for the features named 'X' and the label (dependent or target variable) named 'y' and the Pandas dataframes are converted into arrays taking the data one step closer to being fed into the machine learning algorithm. The data is then sampled which assists in the generalization of the model. This refers to randomly splitting the data into train and test datasets, where the train set is fed into the model to train it and the

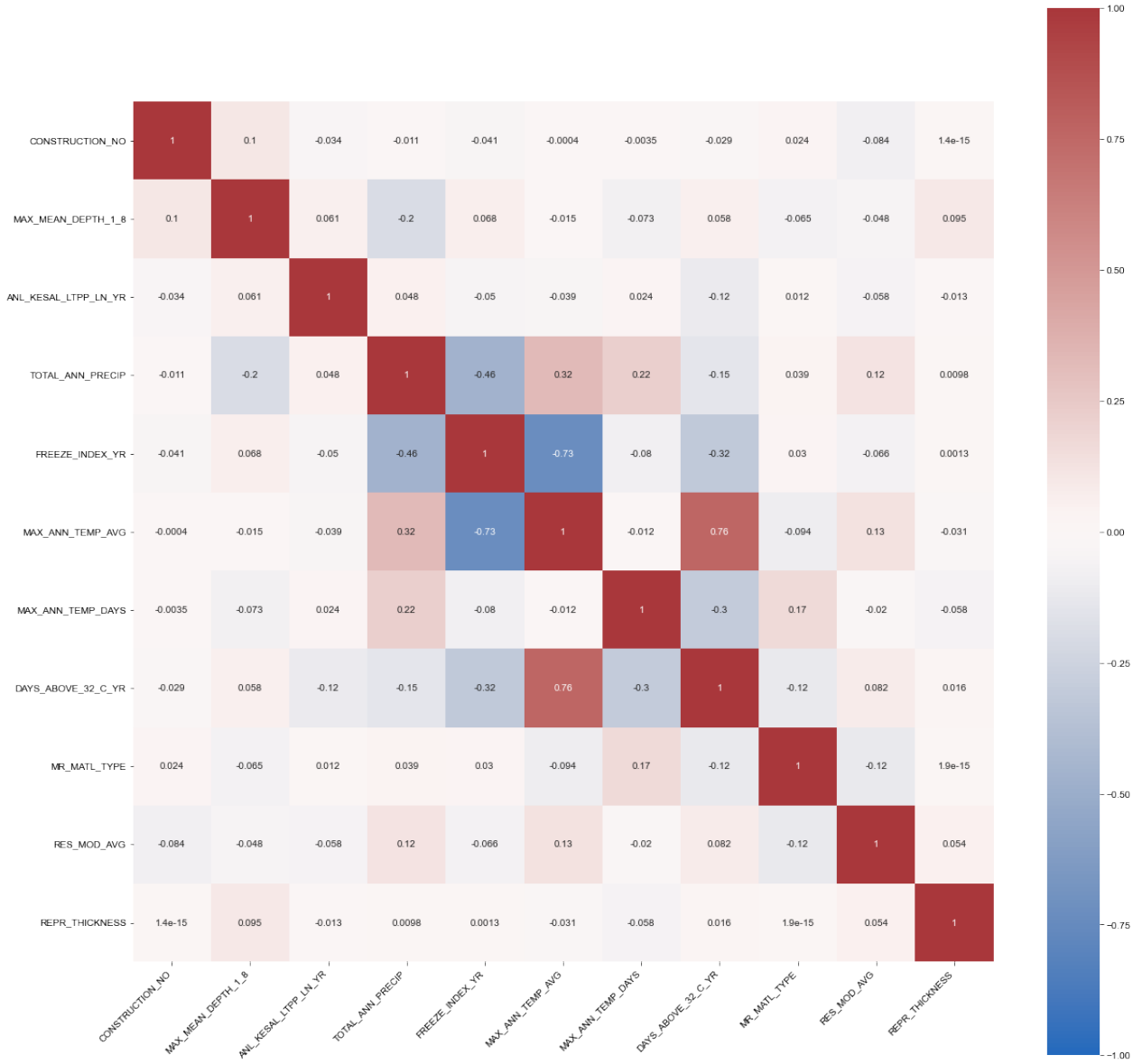


Figure 2: Correlation Matrix of all Variables

test set is used to assess the model.(Marcelino et al., 2021) For this study, the train set was 80% of the final dataset and the test set consisted of the remaining 20% of the data. This was done using the scikit-learn train_test_split function with the random state parameter equal to 42 for reproducibility of the data split. The RandomForestRegressor model is then fit to the data and after using the predict function for the train and test datasets, the mean squared error and R^2 statistic is calculated using the corresponding functions from the scikit-learn metrics library.

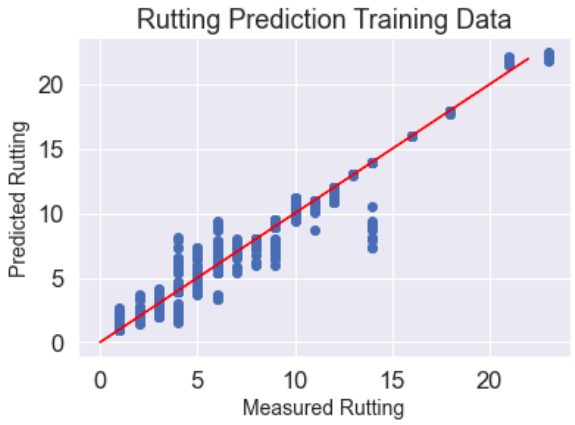
3.0 Model Fitting & Results

The Random Forest model was used to predict rutting in this study. This algorithm was used to better visualize the decision-making process for predicting rutting because the application of this algorithm may improve the accuracy of the predictive machine learning (ML) model. Random Forests are an ensemble method that reduces the variance of an ML model by combining different models and thus limiting overfitting issues.(Madeh Piryonesi El-Diraby, 2021) Scikit-learn was used to implement the model. Most machine learning algorithms have a group of parameters that are responsible for many aspects of the behavior of the algorithm and describe complex properties of the model. These parameters are called 'hyperparameters'. Hyperparameters are fixed before the learning process begins.(Marcelino et al., 2021) The Random Forest model has the following hyperparameters - number of trees in the forest (n_estimators), maximum depth of the tree (max_depth), number of features to consider when looking for the best split (max_features), the minimum number of samples required to be at a leaf node (min_samples_leaf) and minimum number of samples required to split an internal node (min_samples_split). (Pedregosa et al, 2011.) The first model had default parameters defined by scikit-learn which were later revised based on the results of hyperparameter optimization using grid search. The values of the parameters are given in the table below.

	Before GridSearch	After GridSearch
n_estimators	100	5
max_depth	None	22
max_features	1	8
min_samples_leaf	1	4
min_samples_split	2	8

The R^2 statistic measure can be described as a goodness-of-fit measure. The R^2 measure of the model for the training set was computed to be 98.476% and 95.821% for the test

```
Training data mean squared error: 0.19642711734854995
Training data R^2: 0.9847604577597364
Total Datasize: 7875
Training Datasize: 4427
0.9848136899648093
```




```
Testing data mean squared error: 0.48973400431165576
Testing data R^2: 0.9582110804682499
Total Datasize: 7875
Testing Datasize: 1107
```

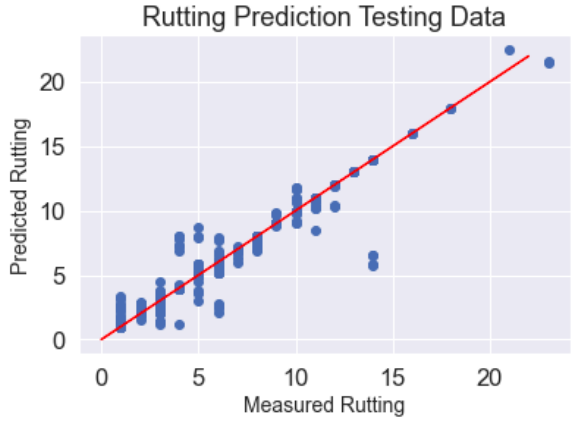


Figure 3: Scatterplots of Measured vs. Predicted Rutting

set. There is evidence of overfitting since the model predicts rutting for the training set significantly better than it predicts rutting for the test set. So, K-fold cross-validation was used to better evaluate the model by the method of resampling the data repeatedly, i.e., k times. In this study, 10-fold cross-validation was used to test the goodness-of-fit of the model, i.e., resampling was done 10 times. To improve the performance of the Random Forest model, hyperparameter optimization using GridSearch was considered. Modifying and fine-tuning the hyperparameters of a model may improve the model and its ability to predict accurately

for test data. The GridSearch method examines all possible combinations of hyperparameters to optimize the model. (Marcelino et al., 2021) By implementing the scikit-learn GridSearch

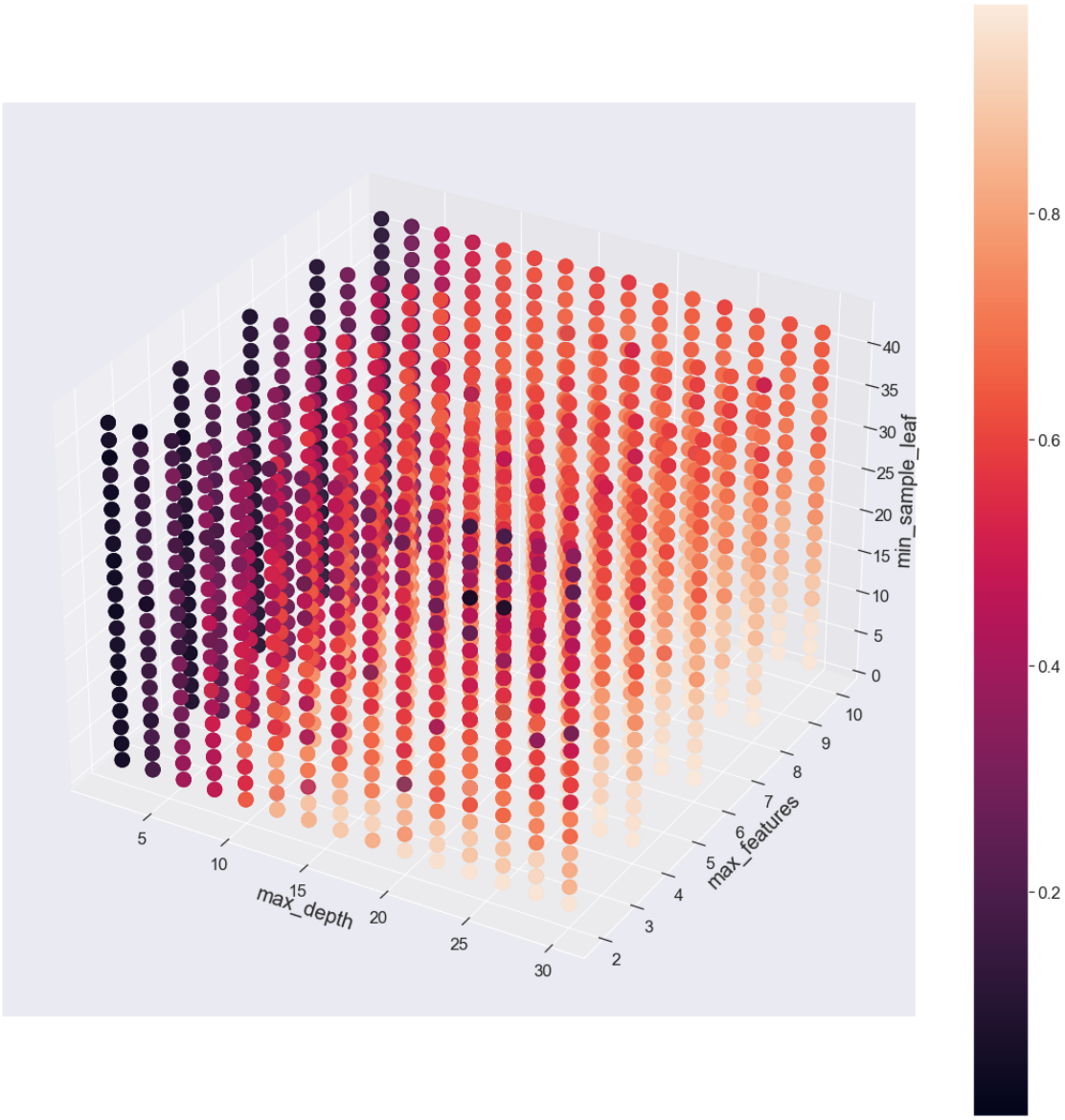


Figure 4: GridSearch Results Visualisation

algorithm (hyperparameter optimization), adapting the model to the best parameters found, and then implementing 10-fold cross-validation, the value of the R^2 statistic measure for the model was observed to be 96.5%. This is a better evaluation of the model and it also dealt with the bias of the training data.

In general, the results suggest that the proposed approach and model can predict rutting based on data from LTPP. Feature importance was obtained from using the feature_importances_ property of the model to make sure that the rutting prediction model made sense from an engineering outlook. This showed that the variables with the most impact were annual

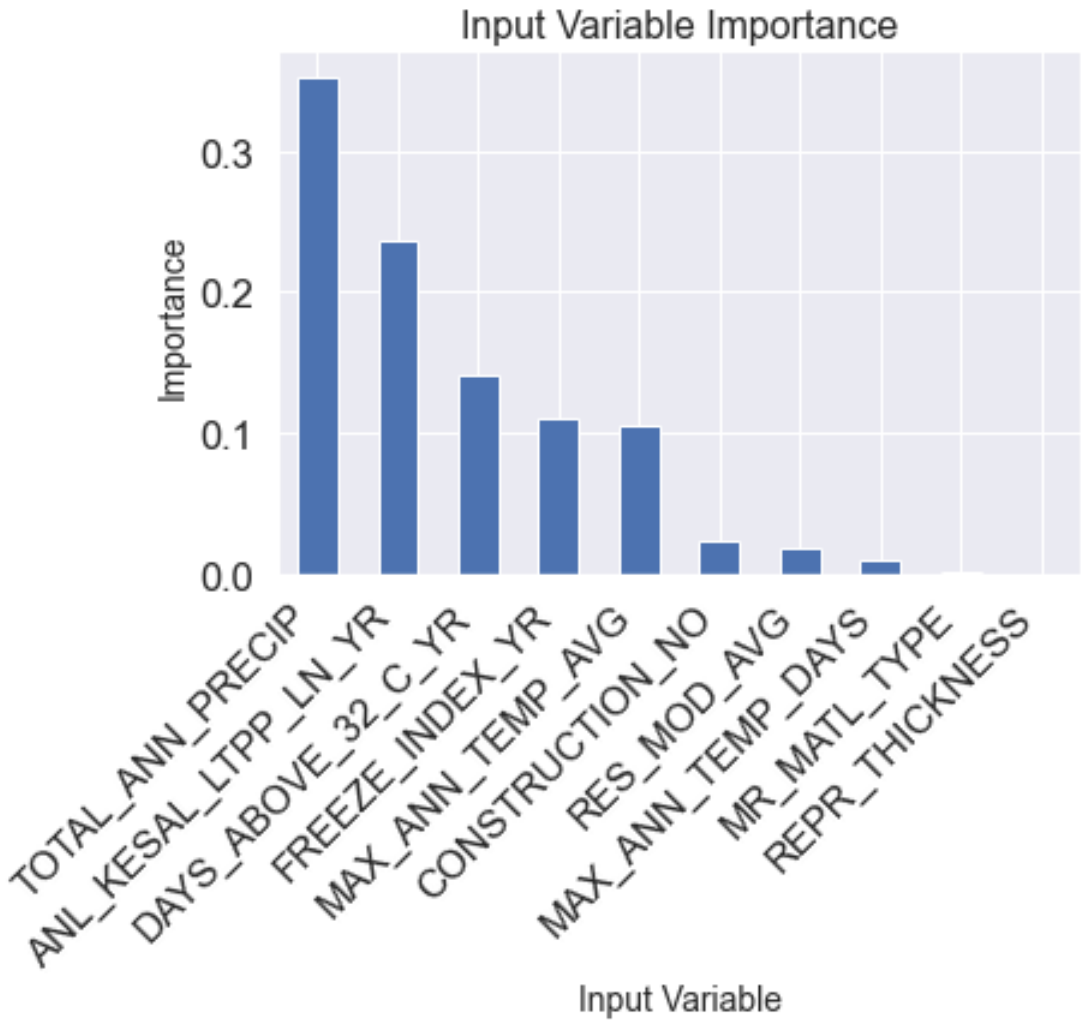


Figure 5: Feature Importance

precipitation, traffic loads, and temperature features which correctly follows the theoretical background of external factors affecting the rutting of flexible pavements.

The scatter plot results indicate that the model has an acceptable predictive capacity because the scatterplots show an R^2 value of 0.98476 and 0.95821 for the train and test datasets.

4.0 Conclusions

Prediction of pavement performance is an essential tool that supports the transport system. Machine learning models that have greater predictive ability can predict pavement performance more accurately, thereby playing a role in maintenance decisions. The aim of this study was to build a good Random Forest model that predicts rutting in flexible pavement. Thus, it is shown that through modelling and machine learning algorithms, accurate predictions can be made. The model built exhibited reasonable predictive ability with an acceptable accuracy of 96.5%.

The results of this study will be of interest to road authorities performing road maintenance work. By improving prediction accuracy, machine learning algorithms can optimize maintenance and rehabilitation work and reduce costs. The ideas behind this research can be applied to different machine learning models and different pavement performance indicators, depending on the needs and goals of road authorities.

Further research could explore other machine learning algorithms and on the prediction of various performance measures. (Marcelino et al., 2021)

References

About the Centre for Pavement and Transportation Technology. (2012, June 29). Centre for Pavement and Transportation Technology. https://uwaterloo.ca/centre-pavement-transportation-technology/about-centre-pavement-and-transportation-technology

LTPP FAQs — FHWA. (n.d.). Retrieved August 12, 2022, from https://highways.dot.gov/research/long-term-infrastructure-performance/ltpp/frequently-asked-questions

LTPP InfoPave - Home. (n.d.). Retrieved August 12, 2022, from https://infopave.fhwa.dot.gov/

Madeh Piryonesi, S., El-Diraby, T. E. (2021). Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling. Journal of Infrastructure Systems, 27(2), 04021005. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000602

Marcelino, P., de Lurdes Antunes, M., Fortunato, E., Gomes, M. C. (2021). Machine learning approach for pavement performance prediction. International Journal of Pavement Engineering, 22(3), 341–354. https://doi.org/10.1080/10298436.2019.1609673

Naiel, Asmaiel Kodan, "Flexible Pavement Rut Depth Modeling For Different Climate Zones" (2010). Wayne State University Dissertations. Paper 179.

Pedregosa et al (2011). Scikit-learn: Machine Learning in Python — Scikit-Learn 1.1.2 Documentation, JMLR 12, 2825-2830. Retrieved August 12, 2022, from https://scikit-learn.org/stable/index.html