World Conference on Transport Research - WCTR 2023 Montreal 17-21 July 2023

Pavement Performance Prediction using Machine Learning: Supervised Learning with Tree-Based Algorithms

[[1]](#footnote-1)

Abstract

This study employs supervised machine learning tree-based algorithms to predict the performance of flexible pavements, as the pavement quality index was used the International Roughness Index (IRI). Three algorithms, namely Decision Tree, Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) were applied to assess pavement quality. The data to develop the machine learning models were sourced from the Long-Term Pavement Performance InfoPave database, selecting 55 experimental sections of asphalt concrete on granular and asphalt concrete on bound bases. The study considered only pavements without maintenance or rehabilitation. Likewise, models were trained using the structural number, annual average daily truck traffic, precipitation, and temperature, with IRI as the target variable. The results demonstrated that the XGBoost model outperformed the others, achieving an R-squared value of 0.98, while the Random Forest model achieved an R-squared value of 0.95. These findings indicate the potential of tree-based algorithms in predicting pavement performance, offering valuable insights for future research.

© 2023 The Authors. Published by Elsevier B.V.  
This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by-nc-nd/4.0>)

Peer-review under responsibility of the scientific committee of the World Conference on Transport Research – WCTR 2023.

*Keywords:* Pavement; Performance; Machine Learning; Supervised Learning; Decision Tree; Random Forest; XGBoost; IRI.

Highlights

* Tree-based algorithms show promising results for predicting the performance of flexible pavements.
* The models with the best results were XGBoost and Random Forest.
* The traffic is the most important feature for predicting the IRI of flexible pavements.

# Nomenclature

AADTT Annual Average Daily Truck Traffic

AAT Annual Average Temperature

FHWA Federal Highway Administration

IRI International Roughness Index (IRI)

LTPP Long-Term Pavement Performance

ML Machine Learning

PMS Pavement Management System

RMSE Root-Mean-Square-Error

SN Structural Number

TAP Total Annual Precipitation

XGBoost eXtreme Gradient Boosting

1. Introduction

To maintain a road network in good service condition, a Pavement Management System (PMS) is essential. Likewise, one of its main components is the pavement performance model (Ferreira et al., 2009; Jorge and Ferreira, 2012). This study aims to understand the best machine learning techniques for predicting the quality of road paving, with specific objectives to identify the most suitable algorithms and features, as well as to determine the optimal amount of data required for training an effective model and the appropriate pre-processing and treatment of this data.

This study compares three supervised Machine Learning (ML) models applied to predict the performance of flexible pavements. Tree-based algorithms were employed. This choice was based on the capacity and interpretability demonstrated by these algorithms. Also, the pavement quality indicator was the International Roughness Index (IRI), a widely used pavement smoothness and irregularity monitoring parameter (Fwa, 2006).

The algorithms selected were Decision Tree (Quinlan, 1986), Random Forest (Breiman, 2001), and XGBoost (Chen and Guestrin, 2016). The model created aims to predict the IRI of flexible pavements until its first intervention. Therefore, only pavements without maintenance or rehabilitation were considered.

The workflow adopted is to study the problem, obtain data, prepare the data, apply the model, evaluate the results, improve the model, and conclusively put the model into production (O'Neil and Schutt, 2013; Géron, 2017; Thakur, 2020). The data used are from the Federal Highway Administration (FHWA) program called Long-Term Pavement Performance (LTPP) InfoPave (LTPP, 2020). As a result, data from monitoring sections of the asphalt concrete on granular base and asphalt concrete on bound base were analyzed. The data represents IRI, pavement structure, traffic, and climate conditions.

Our work contributes to the understanding that supervised learning tree-based algorithms can effectively calculate the IRI. Additionally, we identified the most important features for training the models and the best-performing algorithms.

Finally, this article is structured into five sections. Section two presents the background. Section three provides information about tools, models, data, and methods. Section four discusses the results, and section five summarizes the main findings.

1. Background

A pavement performance model predicts a road network's quality based on monitoring data. According to Meegoda et al. (2014), truck traffic volume, climatic data, and the pavement's structural condition are recommended variables for pavement performance prediction models. As well, IRI is the main parameter used to define pavement smoothness (Fwa, 2006). The roughness of a road is defined by the variation in the surface that induces vibration in vehicles (Sayers et al., 1986). Likewise, Gillespie (1992) states that the IRI "is a scale for roughness based on the response of a generic motor vehicle to the roughness of the road surface". Therefore, the IRI is an index representing the vertical response of a hypothetical quarter-car traveling at 80 km/h and is calculated based on the vehicle's responses to variations in the pavement surface, generating an average longitudinal profile of the road (Fwa, 2006).

Since the end of the 1980s, the LTPP program has aimed to provide information that supports the development of models to explain the performance of road pavements (FHWA, 2019). The LTPP program has an online viewing platform called InfoPave (LTPP, 2020), where data from 2581 monitoring sections are available. This database supports several studies and is the most important database with road pavement data.

This study uses supervised machine learning models to predict IRI based on features such as traffic, pavement structure, and climate data. In supervised learning problems, multiple targets are associated with the dataset's existing data, so the model learns by example (Thakur, 2020).

The tree-based models are built on supervised learning, i.e., it is a mathematical structure that, given a target value Y, a prediction is made given an input X. Thus, to use these models, it is necessary to have a dataset with training values (features) and target values (labels) (Géron, 2017). In a tree-based model, the aim is to get a classifier to simulate the structure of a tree. The nodes correspond to the features, and the leaves to the classifications or predictions (Quinlan, 1986). Figure 1 shows a generic example of the Decision Tree structure. Lastly, decision trees are popular because they provide good results and are easy to implement and interpret (Gottard et al., 2020; Lundberg et al., 2020).

Diagram

Description automatically generated

Figure 1. Generic tree example

1. Materials and Methods

The algorithms applied were Decision Tree (Quinlan, 1986), Random Forest (Breiman, 2001), and XGBoost (Chen and Guestrin, 2016). The workflow adopted to train the models followed the steps covered by Thakur (2020), Géron (2017), and O'Neil & Schutt (2013). Initially, it is necessary to understand the problem, from which the data was collected and then pre-processed, transformed, and prepared for the model. At that time, the data was split into training and testing datasets. The training dataset is utilized to train the model. Then, an independent test dataset is used to evaluate the performance. Depending on the metrics adopted, the model may advance to production, or it may be necessary to improve the algorithm. The workflow is shown in Figure 2.

Diagram

Description automatically generated

Figure 2. Typical machine learning workflow

* 1. Experiment environment

The repository with all data, models, and instructions is available at: [github.com/tamagusko/predict-iri-trees](https://github.com/tamagusko/predict-iri-trees).

The Python version used was 3.9.15. The computer used to develop models have the following specifications:

* CPU: AMD Ryzen 9 5950X
* Memory: 64 GB DDR4 3000 MHz RAM
* GPU: NVIDIA GeForce RTX 3090 24 GB
* Operating System: Arch Linux, Kernel 6.0.8-arch1-1
  1. Data

The data to train the ML model were collected from InfoPave (LTPP, 2020). Monitoring sections classified as asphalt concrete on granular base and asphalt concrete on bound base were selected. The training features used data representing the pavement structure (Structural Number), performance (IRI), traffic (Annual Average Daily Truck Traffic), and climate indicators (Total Annual Precipitation and Annual Average Temperature). Only sections without maintenance and rehabilitation were selected. This choice was due to the complexity that different pavement interventions could add to the model. After the filters, 55 sections distributed across the US, Canada, and Puerto Rico became available (Figure 3). The distribution of monitoring sections is not uniform in the American territory. Most of them are located near the coast and on the east coast. This can generate biases in the models developed. However, the LTPP database is the most significant existing and has been the source of several studies. In the future, our team plans to use data augmentation and synthetic data to overcome this challenge. Table 1 describes the features selected to develop the machine learning models.

Map

Description automatically generated

Figure 3. Sections of the LTPP program used in this study

Table 1. Descriptive statistics of features used to train the model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Feature | Mean | Min | 25% | 50% | 75% | Max |
| Year | 1996 | 1989 | 1992 | 1995 | 1999 | 2017 |
| IRI (m/km) | 1.34 | 0.61 | 0.92 | 1.20 | 1.60 | 4.10 |
| SN | 4.16 | 1.10 | 2.80 | 4.20 | 5.30 | 8.00 |
| AADTT | 452 | 4 | 142 | 262 | 526 | 3126 |
| Precipitation (mm) | 1010 | 92 | 715 | 1061 | 1334 | 2091 |
| Temperature (°C) | 13.80 | 2.70 | 7.90 | 15.30 | 17.50 | 25.90 |

The data under study are distributed between 1989 and 2017. Meanwhile, information is not available for each section in all monitoring years. Most of the data used are concentrated between 1992 and 1999. Also, the data are concentrated in sections with an IRI lower than 2.0 m/km. Consequently, the model is expected to have better results for lower IRI values.

The training features were Structural Number (SN), Annual Average Daily Truck Traffic (AADTT), Total Annual Precipitation (TAP), and Annual Average Temperature (AAT), while the target (value to be predicted) is the IRI.

The data used to develop the ML model is presented in Figure 4. The lower part of the triangle represents the scatter plot, and the upper part represents the correlation between variables. As expected, the highest correlation is related to the AADTT and the SN, i.e., the roads with higher traffic were designed with a more robust structure.

A picture containing diagram

Description automatically generated

Figure 4. Matrix with processed data for model development

* 1. Data processing

The data were organized in a relational database (rows and columns). For each row, the information presented corresponds to a section and year. There is no correlation between different years for the same section. Consequently, each row corresponds to an individual input for the machine learning model, with features and a target. A snippet with the first five rows is shown in Table 2.

Table 2. Data processed for the study

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Year | Precipitation (mm) | Temperature (°C) | AADTT | SN | IRI (m/km) |
| 773 | 1989 | 434 | 15.1 | 520 | 4.7 | 0.90 |
| 1181 | 1989 | 1404 | 4.4 | 195 | 2.8 | 0.87 |
| 1182 | 1989 | 1404 | 4.4 | 195 | 2.8 | 0.85 |
| 1183 | 1989 | 1404 | 4.4 | 195 | 2.8 | 0.86 |
| 1106 | 1989 | 1178 | 5.1 | 586 | 2.8 | 1.76 |
| 773 | 1989 | 434 | 15.1 | 520 | 4.7 | 0.90 |

Furthermore, rows with missing values were excluded. Thus, from 3,166 entries from the 55 monitoring sections, 2,203 remained with valid data for training and testing the model after processing. In this study, the outliers were not cleaned. It was observed that some of the values statistically identified as outliers were actually more degraded pavements, with IRI values greater than 3.5. Thus, techniques to rebalance the datasets should be employed in future work to improve the distribution. Therefore, it is expected that the models developed on these data will have better results for new or good pavements due to the characteristics of the training data.

Finally, a random split was performed with 75% of the data for training and 25% for testing. Accordingly, 1,652 sections were used for training and 551 for testing.

* 1. Parameter Optimizing

The optimization of the tree's depth was made (Figure 5). The objective is to find the optimal value for which a higher R-squared and the lowest Root-Mean-Square-Error (RMSE) are obtained. This process was essential to reduce the likelihood of overfitting and underfitting, which could negatively impact the model's performance.

Default values were utilized for the other training parameters to streamline the process and focus on the depth optimization.

Depth optimization is a critical aspect of model training, as it directly influences the model's ability to generalize and learn from the given data. Too shallow a tree might lead to underfitting, where the model cannot capture the underlying patterns and complexities in the data. Conversely, a too deep tree could result in overfitting, where the model is too specific to the training data and needs to generalize better to unseen data. By carefully tuning the tree depth, the study aimed to balance these two extremes, ensuring that the models could provide accurate and reliable predictions for pavement performance.

Chart, line chart

Description automatically generated

Figure 5. Tree depth tuning

1. Results

The execution time of these models on the workstation described in item 3.1 was only few seconds. However, this value is unimportant for a PMS, since the prediction of pavement quality is an activity that must be performed only few times a year. Thus, it can be said that the computational power required to process 55 monitoring sections is low. The first element evaluated is the importance of the features to the predictions (Table 3).

Table 3. Feature importance for models

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Decision Tree | Random Forest | XGBoost |
| AADTT | 34.4% | 34.0% | 32.3% |
| SN | 22.8% | 17.3% | 29.1% |
| Temperature | 17.0% | 18.9% | 15.8% |
| Precipitation | 11.9% | 17.4% | 11.8% |
| Year | 13.9% | 12.4% | 11.0% |

The importance of features had a similar order to the models developed. The most important feature was traffic (AADTT), corresponding to approximately one-third for all models. The second position was the pavement structure (SN) for the Decision Tree and XGBoost algorithms. However, in Random Forest, there was an inversion with temperature in second place. As the Random Forest algorithm gave less importance to structural capacity, greater weight was allocated to climatic factors, mainly precipitation. Thus, the XGBoost model, unlike the previous ones, took more into account the structural capacity. Based on Table 3, the Random Forest algorithm seems more robust because it better distributes the weights among the various features. Finally, the percentages show that the choice of features to train the models was appropriate.

Predictive ability and errors were tested using R-squared and RMSE (Root Mean Square Error) indicators to determine the best model. The ideal R-squared value is as close as possible to one, indicating that predicted values match actual field measurements. On the other hand, the RMSE measures the average magnitude of the prediction errors, with a lower value indicating a better model fit. Figure 6 illustrates the results and compares the models. The vertical axis represents predictive values, while the horizontal axis represents measured values. In an ideal model, points would be situated on the diagonal connecting the axes, representing accurate predictions with minimal errors.

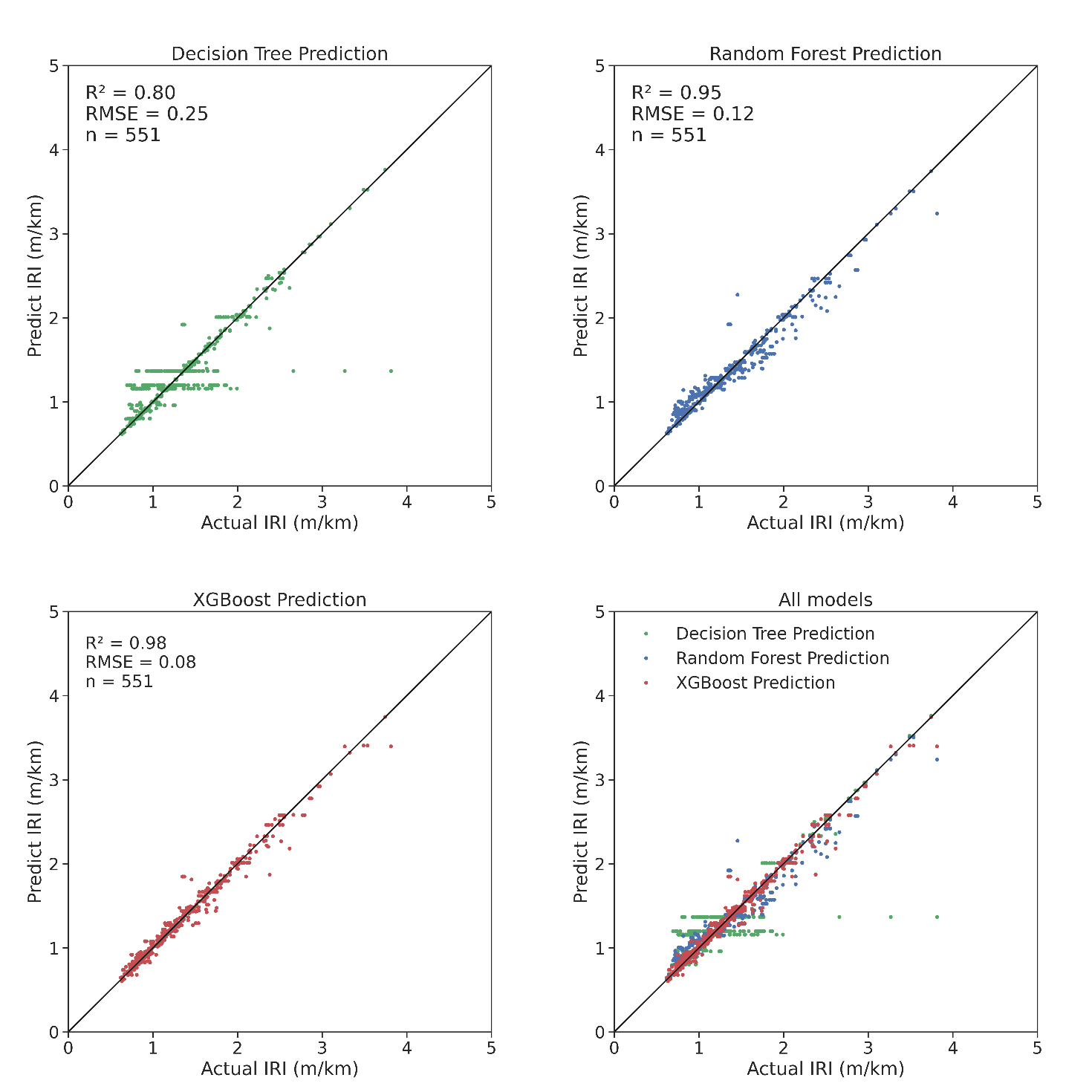


Figure 6. Results for the models evaluated on the test data

In the first model, the Decision Tree (green), some predicted values fall outside the diagonal, indicating the model's challenge in generalizing values for the test data. This disparity may be due to the approach taken or the data used, which might need to be more suitable for this model.

The Random Forest model (blue) produced favorable results, with an R-squared of 0.95 and an RMSE of 0.12 for the test sample. The model demonstrated robustness and exhibited satisfactory performance, particularly for well-maintained pavements. This model could be improved by incorporating more training data from degraded pavements.

The XGBoost model (red) outperformed the others, achieving an R-squared of 0.98 and an RMSE of 0.08 for the test sample. The model converged quickly to desirable values during parameter tuning. The model's inherent boosting techniques may have contributed to this performance. The distribution of errors is adequate across the test data, even for higher values.

These results are not definitive and carry the inherent limitation of relying on training data from pavements without maintenance and rehabilitation. Consequently, the machine learning models are exposed only to higher-quality pavements. While these models can predict the quality of road pavements before their first intervention, they still need to be improved before being integrated into a PMS.

1. Conclusions

The results of the ML models were promising, particularly for IRI values below 1.6 m/km. A potential explanation for this behavior is that these lower IRI values correspond to the early stages of the pavement life cycle, during which pavement deterioration is typically almost linear and predictable. Additionally, the training data contains a considerable number of IRI values below 2.0 m/km (see table 1), indicating that the dataset was unbalanced for good quality pavements. Consequently, the results were well-suited to the training data, taking into account the limitations imposed by focusing only on pavements without maintenance or rehabilitation.

As the model is at the moment, it can be used to predict the IRI with SN, AADTT, and climate information (precipitation and temperature) without interventions. Besides, the model can be adapted to other structural or traffic parameters. Nevertheless, the models must be retrained, and the model's performance may be different. Even so, these developed models are good candidates to be the basis for transfer learning. Thus, agencies with fewer data can benefit from these models to apply in their local context. The models require little computational power for processing. The most time-consuming process was parameter optimizing, where the models were trained several times until they found the best values. Therefore, there are no problems applying this ML model by any agency with input parameters that want to predict the IRI's evolution.

Finally, the best model developed is the XGBoost, obtaining an R-squared of 0.98 and RMSE of 0.08, closely followed by Random Forest, with an R-squared of 0.95 and RMSE of 0.12. In the future, this model will be expanded to consider pavement interventions. Furthermore, models for 5, 10, and 20 years will be developed. It is also intended to create an auxiliary module able to test the IRI result according to planned interventions. Another objective will be to test the robustness of the model for data from sources other than LTPP. Data and models are available at [github.com/tamagusko/predict-iri-trees.](https://github.com/tamagusko/predict-iri-trees)

Acknowledgments

The first author would like to thank the support by the Portuguese Foundation for Science and Technology (FCT), which awarded the PhD scholarship grant 2020.09565.BD. The authors also would like to thank the support of the Research Centre for Territory, Transports and Environment (CITTA) and also ACIV for the presentation of this paper in the World Conference on Transport Research.

References

Breiman, L., 2001. Random forests 28. https://doi.org/http://dx.doi.org/10.1023/A:1010933404324

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 785–794. https://doi.org/10.1145/2939672.2939785

Ferreira, A., Picado-Santos, L., Antunes, A., 2009. Pavement Performance Modelling: State of the Art. Comput. Tech. Civ. Struct. Eng. 58, 157–164. https://doi.org/10.4203/ccp.58.8.1

FHWA, 2019. Long-Term Pavement Performance Information Management System User Guide 2019 [WWW Document]. URL https://infopave.fhwa.dot.gov/ (accessed 9.25.20).

Fwa, T., 2006. The handbook of highway engineering.

Géron, A., 2017. Hands-On Machine Learing With Scikit-Learn & Tensor Flow, Hands-on Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media. https://doi.org/10.3389/fninf.2014.00014

Gillespie, T.D., 1992. Everything You Always Wanted to Know about the IRI, But Were Afraid to Ask! https://doi.org/10.1109/msp.2006.65

Gottard, A., Vannucci, G., Marchetti, G.M., 2020. A note on the interpretation of tree-based regression models. Biometrical J. 62, 1564–1573. https://doi.org/10.1002/bimj.201900195

Jorge, D., Ferreira, A., 2012. Road network pavement maintenance optimisation using the HDM-4 pavement performance prediction models. Int. J. Pavement Eng. 13, 39–51. https://doi.org/10.1080/10298436.2011.563851

LTPP, 2020. Long-Term Pavement Performance program's data portal [WWW Document]. URL https://infopave.fhwa.dot.gov (accessed 9.25.20).

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Meegoda, J.N., Gao, S., 2014. Roughness progression model for asphalt pavements using long-term pavement performance data. J. Transp. Eng. 140, 1–7. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000682

O'Neil, C., Schutt, R., 2013. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1, 81–106. https://doi.org/10.1007/bf00116251

Sayers, M.W., Gillespie, T.D., Queiroz, C.A.V., 1986. The International Road Roughness Experiment (IRRE): establishing correlation and a calibration standard for measurements.

Thakur, A., 2020. Approaching (Almost) Any Machine Learning Problem.

1. \* Corresponding author. Tel.: +351-239-797-101.

   *E-mail address:* adelino@dec.uc.pt [↑](#footnote-ref-1)