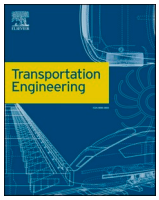
Transportation Engineering 18 (2024) 100282

Contents lists available at ScienceDirect

Transportation Engineering

journal homepage: www.sciencedirect.com/journal/transportation-engineering



Full Length Article

Predicting Marshall stability and flow parameters in asphalt pavements using explainable machine-learning models

Ibrahim Asi [a], Yusra I. Alhadidi [b,c], Taqwa I. Alhadidi [a,*]

[a] Civil Engineering Department, Al-Ahliyya Amman University, 19328, Amman, Jordan
[b] Ph.D. Candidate, Transportation Engineering Researcher, Illinois Center for Transportation 1611 Titan Drive, Rantoul, IL 61866
[c] Research Fellow AlBalqa'a Applied University, Salt, Jordan

ARTICLE INFO

Keywords:
Bitumen content
HMA
Marshall flow
Marshall stability
Prediction model
SHAP

ABSTRACT

The traditional method for determining the Marshall stability (MS) and Marshall flow (MF) of asphalt pavements is laborious, time consuming, and costly. This study aims to predict these parameters using explainable machine-learning techniques. A comprehensive database comprising 721 hot mix asphalt (HMA) data points was established, including variables such as aggregate percentage, asphalt content, and specific gravity. Models were constructed using the PyCaret Python library, and their performance was assessed using metrics such as the mean absolute error (MAE) and coefficient of determination (R^2). The CatBoost regression model outperformed the other models, achieving R^2 values of 0.835 and 0.845 for MS and MF, respectively. Additionally, Shapley values were used to quantify the variable effects on the predictions. This approach enables the efficient preselection of design variables, reducing the need for extensive laboratory testing and promoting sustainable construction practices.

1. Introduction

Pavement performance and longevity are significantly influenced by mechanical and volumetric mix characteristics, which are determined by the mix design. These properties are fundamentally related to field defects, such as aging, rutting, and reflective cracking, indicating the potential of pavements for permanent deformation [43]. Currently, the Marshall, Hveem, and Superpave mix designs are the most popular methodologies employed in many countries [19]. These methods differ in their compaction styles, specimen sizes, and mechanistic testing procedures. While the Superpave mix design has gained popularity in recent years, the Marshall method remains a preferred choice in Jordan due to its historical application and ease of use [21]. However, these methods are labor intensive, rigorous, and expensive. For instance, a typical Superpave mixture design can take approximately 7.5 working days, while the Marshall mix design can take approximately four working days [37]. Both methods required the preparation of multiple replicates with different percentages to determine the design mix formula (DMF). Given the extensive experimental work required for defining the DMF of asphalt mixes, prediction-based techniques are highly advantageous because they offer significant time savings, particularly when material sources or testing boundaries remain unchanged.

With the increasing need for efficiency in pavement design, numerical simulations, such as the discrete element method (DEM) and finite element method (FEM), and soft computing techniques, such as machine learning, have emerged as valuable tools for predicting mix performance [28]. For example, DEM has been employed to predict Marshall parameters for virtual specimens, estimating an asphalt binder content between 4.3% and 7.5%, with a maximum binder content of 5.3% [27]. Similarly, the DEM has been used to predict the volumetric properties of HMA, such as voids in mineral aggregates (VMA) [12,20,45]. Machine learning is a powerful data analysis method that automates the construction of analytical models and offers significant potential for predicting the Marshall mix parameters. Among these approaches, machine learning has proven to be a powerful method for automating data analysis and constructing predictive models, especially for estimating Marshall mix parameters [4,5,14], whereas others have applied machine-learning techniques to estimate various material properties relevant to pavement design, such as surface roughness and cutting force [28,39].

Studies have also employed machine-learning algorithms to predict the engineering and physical characteristics of asphalt mixtures [29,30] and soft computing techniques, such as artificial neural network, genetic

* Corresponding author.
E-mail address: t.alhadidi@ammanu.edu.jo (T.I. Alhadidi).

https://doi.org/10.1016/j.treng.2024.100282
Received 2 May 2024; Received in revised form 28 September 2024; Accepted 3 October 2024
Available online 5 October 2024

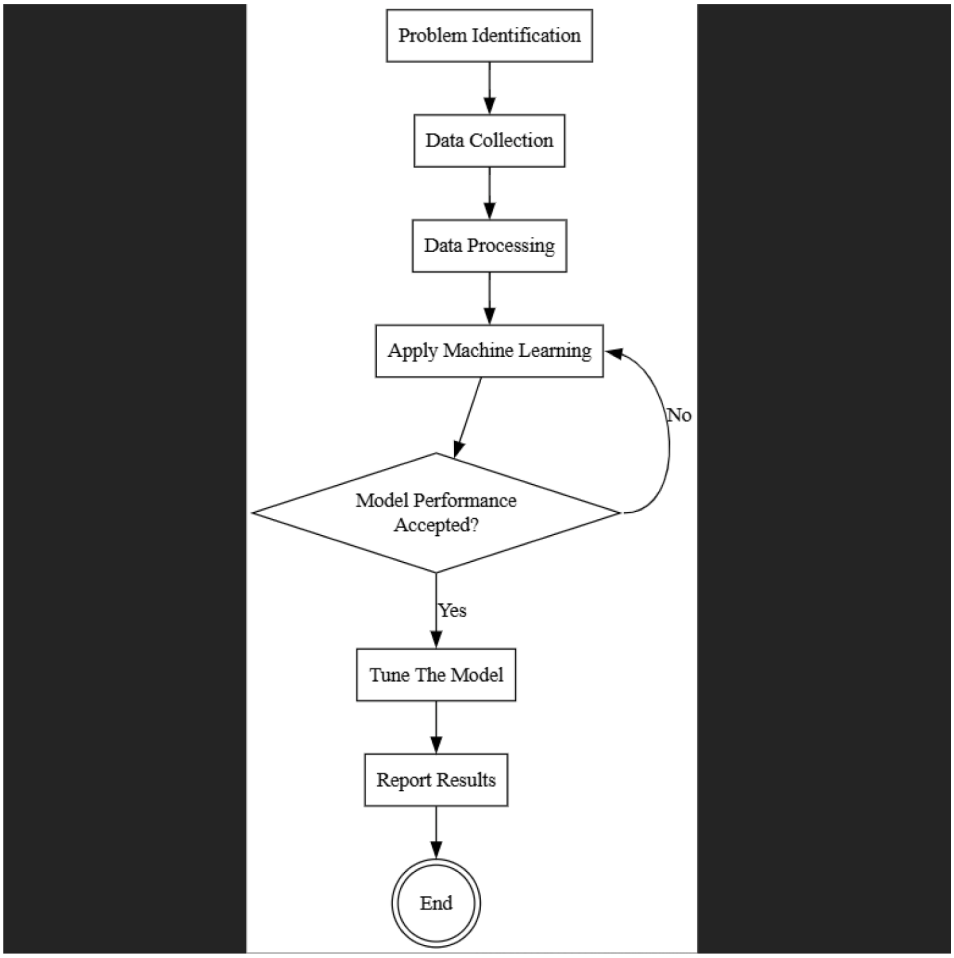


Fig. 1. Research Methodology.

algorithms, and fuzzy logic [31]. For example, researchers have used specific design inputs from Marshall mix designs, such as Va, density, and VMA, to predict the MS, MF, and Marshall quotient (MQ) [2]. Explainable machine-learning techniques provide semantically interpretable tools that can create knowledge resources and help understand and interpret complex models [6,17,34]. These techniques are valuable in pavement engineering, and understanding the influence of various mix parameters on performance outcomes is crucial. Physics-informed machine-learning models, which integrate physical laws and principles into the learning process, can express results more clearly with fewer rules [22]. Knowledge graphs and ontologies are key technologies for building explainable machine-learning models, facilitating both human and machine understanding of data [25].

In pavement engineering, several studies have applied machine learning to predict various properties of asphalt mixtures, including surface roughness and cutting force [28,39]. Additionally, soft computing techniques, such as artificial neural networks, genetic algorithms, and fuzzy logic, have been widely used to predict the engineering and physical characteristics of asphalt mixtures [29–31]. For example, researchers have utilized design inputs from Marshall mix designs, including air void content (Va), density, and VMA, to predict Marshall stability (MS), Marshall flow (MF), and Marshall quotient (MQ) [2].

However, while traditional machine-learning methods offer accurate predictions, they often lack transparency in their decision-making processes. This is where explainable machine-learning techniques, such as SHapley Additive exPlanations (SHAP), come into play. These methods provide semantically interpretable models, enabling users to understand the contribution of each input variable to the output prediction [6,17,34]. In pavement engineering, explainable models are particularly valuable because they help engineers interpret how various mix parameters affect performance outcomes. Furthermore, physics-informed machine-learning models, which integrate physical laws and principles into the learning process, can provide clearer and more interpretable results [22].

To bridge the gap between complex machine-learning models and practical applications, this study aims to leverage explainable machine-learning techniques embedded in the PyCaret Python library to predict Marshall performance parameters. By utilizing a comprehensive database comprising 721 HMA data points from 230 asphalt mixes, variables such as aggregate percentage, asphalt content, and specific gravity were incorporated into the models. Several machine-learning algorithms, including CatBoost, LightGBM, and ExtraTrees, were employed, and their performance was evaluated using metrics such as mean absolute error (MAE) and coefficient of determination (R^2). Among these, the CatBoost regression model demonstrated superior performance, achieving R^2 values of 0.835 for MS and 0.845 for MF.

This study's findings provide a robust framework for the efficient preselection of design variables, promoting sustainable construction practices and advancing the state-of-the-art in pavement engineering.

2. Methodology

This section presents the research methodology framework and discusses the data description and model building.

Re**search framework**: The study framework is illustrated in Fig. 1.

The investigation commenced with the collection of data from 230 construction and maintenance projects in Jordan that utilized bitumen 60/70. This resulted in a comprehensive dataset comprising 721 data points. From these mixes, several key variables were extracted, including MS, MF, percentage of aggregates in the total mix (Ps), percentage of asphalt content (Pb), bulk specific gravity of the compacted mixture (Gmb), maximum specific gravity of the mix (Gmm), specific gravity of the aggregate (Gsb), effective specific gravity of the aggregate (Gse), percentage of VMA, percentage of air voids (Va), percentage of coarse aggregate to fine aggregate ratio (CA/FA), and filler-to-bitumen ratio (f/Pb).

Data preprocessing: To minimize the complexity of the developed models and mitigate data anomalies, preprocessing and data division were implemented. This process encompassed the identification of outliers and verification that the data were free from abnormal values. Outliers were detected and addressed using the interquartile range (IQR) method to prevent adverse effects on model performance. A summary table was generated to evaluate the distribution of variables and identify any anomalies. It is noteworthy that the dataset utilized in this study was derived from real mixes, and consequently, no missing values were imputed because the data were complete. This approach ensures that the predictions generated by the machine-learning models reflect real-world conditions.

Multicollinearity was assessed using the Spearman rank correlation matrix, and variables exhibiting high correlations ($R^2 > 0.85$) were eliminated to prevent multicollinearity issues from compromising model performance. This procedure ensured that the remaining variables incorporated in the models were independent and contributed substantively to the prediction tasks.

Model building: The **PyCaret** library in Python was used to construct two machine-learning models: one for predicting **Marshall stability (MS)** and the other for predicting **Marshall flow (MF)**. The PyCaret framework provides a wide range of machine-learning algorithms, allowing for the comparison of various models. In this study, several machine-learning models were evaluated, including **CatBoost, LightGBM, XGBoost**, and **ExtraTrees**.

The rationale for selecting these models is based on several factors:

- **CatBoost** was chosen for its ability to handle categorical features without explicit encoding and its strong performance with relatively small datasets.
- **LightGBM** and **XGBoost** are both gradient-boosting algorithms known for their efficiency in large datasets and their high accuracy. These models are also highly regarded for their ability to handle missing values (if any) and perform robustly in various regression tasks.
- **ExtraTrees** was selected due to its capacity to reduce overfitting by using multiple decision trees and its ability to average predictions, thereby improving overall predictive accuracy.

Other models, such as random forest and support vector machines (SVM), were also tested but did not perform as well in terms of both predictive accuracy and computational efficiency compared to the models mentioned above. As such, CatBoost, LightGBM, XGBoost, and ExtraTrees were chosen as the final models for further evaluation.

Each model's performance was assessed using several evaluation metrics, including the mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and coefficient of determination (R^2). To ensure the robustness of the results, 10-fold cross-validation was used to validate the models.

Model evaluation: The predictive performance of the models was evaluated using a K-fold cross-validation approach with 10 folds. The models were ranked according to their performance, and the top-performing models were chosen for further analysis. To ensure that the models did not suffer from multicollinearity, residual plots were used and the correlation between the predicted and actual data was plotted to verify the accuracy of the predictions.

Explainability: SHAP (SHapley Additive exPlanations) values were used to quantify the impact of each variable on the predicted MS and MF. SHAP values provide insights into the contribution of each feature to the prediction, offering a deeper understanding of the decision-making process of the model. TreeSHAP was used to visualize feature attribution, showing how each feature value either increased or decreased the prediction accuracy.

2.1. Model structure and modeling technique

The model parameter selection is an initial and significant step in the development of appropriate models. Marshall properties depend on several parameters, including Ps, Pb, Gmb, Gmm, Gsb, Gse, Va, VMA, CA/FA, and f/Pb. In the previous step, we removed the multicollinearity problem because the Marshall parameters were dependent on Pb, Gmb, Gmm, Gsb, Gse, VMA, CA/FA, and f/Pb.

2.1.1. Model development using PyCaret

Modeling was performed using the PyCaret package in Python. PyCaret compared different regression models based on the parameters listed above [3]. PyCaret is a Python-based library that consolidates various machine-learning frameworks and libraries for both regression and classification tasks. The library includes more than 20 models [3]. It simplifies the machine-learning workflow by providing functionalities for model training, hyperparameter optimization, model selection, and deployment. The library offers a user-friendly interface for efficiently experimenting with different machine-learning algorithms and techniques [50]. It has also been utilized in optimizing hyperparameters for decision-tree models predicting gold nanorod sizes from spectra, showcasing its effectiveness in enhancing model performance through automated parameter tuning [36].

PyCaret has more than 20 machine-learning models; in our analysis, the best five models were the CatBoost regressor, ExtraTrees regressors, light gradient boosting, extreme gradient boosting, gradient-boosting regressors, and random forest regressors. An explanation for these models is presented below.

2.1.1.1. CatBoost regressor. CatBoost builds on decision trees by optimizing a loss function using gradient-boosting methods. It applies a unique ordered boosting technique in which the new tree is built based on the old prediction in addition to a new learning rate being added to the prediction function [10,24].

2.1.1.2. ExtraTrees regressor. The ExtraTrees regressor builds multiple decision trees and uses averaging to improve the predictive accuracy and control the overfitting from the different trees. It differs from random forests in how it splits nodes. In essence, each tree $h(x,\Theta k)$, where Θk are the randomly chosen parameters for tree k, is built on a random subset. The final prediction is an average of all individual tree predictions. The general equation for the ExtraTrees regressor is described in Eq. 1 [24].

$$y_i = \frac{1}{N}\sum_{N=1}^{N} h(x, \Theta k), \tag{1}$$

where y_i is the predicted value for y.

2.1.1.3. Light gradient-boosting machine (LightGBM). LightGBM uses gradient-based, one-sided sampling and exclusive features that bundle to process large datasets efficiently. It grows trees "leaf-wise." In fact, LightGBM adds predictions to the new iteration. The general equation for LightGMB is presented in Eq. 2 [36].

$$new\_prediction = old\_prediction + \eta \times f_t(x) \tag{2}$$

where η is the new learning rate and $f_t(x)$ is the new learning rate.

2.1.1.4. Extreme gradient boosting (XGBoost). XGBoost improves on standard gradient boosting through systems optimization and enhancements such as handling missing values and pruning trees. The general equation of the XGBoost was shown in Eq. 2.

2.1.1.5. Gradient-boosting regressors. Gradient-boosting constructs additive models in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. Each new tree makes up for the shortcomings of the existing combined model. Similar to the results of other boosting techniques, the new predicted value is built based on the previous predicted value after adjusting the predicted value to minimize the loss function.

2.1.1.6. Random forest regressors. Random forest regressors combine multiple trees to reduce the risk of overfitting and to increase the model's predictive accuracy. Each decision tree in the forest considers a random subset of features when forming questions and has access only to a random set of the training data points. The general equation of random forest is shown in Eq. 3.

$$y_i = \frac{1}{N}\sum_{N=1}^{N} tree_i(x) \tag{3}$$

The expressions reflect the basic operational mechanics of these ensemble methods, where the model successively refines predictions or classifications through an ensemble of simpler, weaker models, typically decision trees.

2.2. Hyperparameter optimization

To enhance the predictive accuracy and efficiency of our machine-learning models, an extensive hyperparameter optimization was conducted using PyCaret's robust suite of automation tools. This critical phase involved the following:

1. **Selection of hyperparameters**: A comprehensive range of hyperparameters was considered for tuning, including learning rates, depth of trees, and regularization coefficients. This selection was tailored to each model to explore the best combinations to improve performance.
2. **Optimization technique**: We employed grid search and random search methods facilitated by PyCaret, allowing us to systematically evaluate various parameter combinations. These techniques provided a structured approach to navigating the hyperparameter space efficiently.
3. **Evaluation metrics**: The optimization process was guided by key performance to reduce the RMSE. This reduction helped in identifying the configurations that yielded the most predictive and robust models.
4. **Results and selection**: The best-performing parameters were selected based on their performance metrics during cross-validation stages. This approach ensured that our models were not only accurate but also generalizable to new, unseen data.

2.3. Evaluation criteria and performance measures

The prediction performances of the models were assessed using six indicators and visual plots. The performance indicators included MAE, MSE, RMSE, (R^2), RMSLE, and MAPE. These indicators were estimated using Equations 4 through 9.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{4}$$

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

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{6}$$

$$R^2 = 1 - \frac{RSS}{TSS} \tag{7}$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i + 1) - \log(y_i + 1))^2} \tag{8}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{9}$$

In Equations 4 through 9, y_i is the actual value, $\hat{y}_i$ is the predicted value, n is the number of training or testing datasets, *RSS* is the residual sum of squares, and *TSS* is the total sum of squares. The maximum R^2 value was as high as 1.00. It can also be negative when the prediction performance is poor. Therefore, the closer this value is to 1.00, the higher the prediction performance of the model. The visual plots included residual plots to ensure that the model did not exhibit multicollinearity, by having a random residual of approximately zero.

2.4. K-fold cross-validation

Cross-validation (CV) was used to assess the generalization performance of the prediction model. It is more stable and comprehensive than the basic training–test split method. One of the most common CV methods is the k-fold CV. The number k is decided by the user and is commonly chosen as 5 or 10 [33]. The CV embedded in PyCaret was used in this study. In PyCaret, CV uses 10 folds to validate the model.

2.5. Explainable machine learning

Explainable machine learning (XML) aims to generate models that provide precise predictions along with transparency and interpretability in decision-making processes [42]. The primary objective of XML is to boost user trust by elucidating the reasoning behind machine-learning predictions [8]. This transparency is of paramount importance in fields such as healthcare, finance, manufacturing, and transportation, where decisions have significant real-world impacts [35]. Central to XML are explainable artificial intelligence (XAI) techniques, which enhance model interpretability and transparency [49]. XAI methods enable researchers to delve into the inner workings of complex models and comprehend the influence of individual features on predictions [47]. Explaining feature importance in predictions is a primary objective of SHAP (SHapley Additive exPlanations), a well-established method in the field [40]. The use of XAI techniques such as SHAP significantly improves the interpretability and trustworthiness of models [18]. The Morris method, also known as the one-at-a-time method, is a global sensitivity analysis technique in which only one input is adjusted per run. This approach is relatively fast, requiring fewer model executions compared to other sensitivity analysis algorithms. However, it has limitations in distinguishing between non-linearities and interactions among inputs. The Morris method is primarily used for screening purposes to identify which inputs are significant enough to warrant more detailed analysis [32]. A lower convergence index indicates that the sensitivity indices are stabilizing, suggesting that a sufficient number of trajectories have been used. A higher convergence index implies that more trajectories are needed for the sensitivity analysis to converge.

3. Results

In this section, explanatory data analysis, variable selection, and

Table 1
Study parameter statistics summary.

Variable (unit)	Mean	Std	Min	25%	50%	75%	Max	Coef of variation	Skewness	Kurtosis
Pb (%)	4.82	0.95	3.00	4.00	5.00	5.50	7.00	0.20	0.13	–0.72
Gmb (g/cm^3)	2.40	0.15	2.18	2.34	2.41	2.46	4.40	0.06	9.37	121.69
Gmm (g/cm^3)	2.51	0.08	2.30	2.45	2.52	2.57	2.71	0.03	–0.18	–0.69
Va (%)	4.64	2.29	0.30	2.70	4.30	6.40	11.10	0.49	0.38	–0.78
VMA (%)	14.51	0.81	13.00	14.00	14.40	15.00	19.00	0.06	0.77	1.50
Stability (kg)	1652.7	266.7	660.0	1452.0	1625.0	1826.0	2482.0	0.16	0.28	0.38
Flow (mm)	3.00	7.96	1.44	2.24	2.62	3.05	4.5	2.66	25.33	658.48
Gse (g/cm^3)	3.10	4.37	2.53	2.65	2.75	2.83	55.40	1.41	11.87	139.07
Gsb (g/cm^3)	2.69	0.08	2.48	2.65	2.72	2.74	2.81	0.03	0. 12	15.3
f/Pb (ratio)	0.94	0.32	0.15	0.73	0.93	1.14	1.71	0.34	–0.07	–0.26
CA/FA (ratio)	1.13	0.27	0.43	1.04	1.13	1.24	2.68	0.24	0.71	7.38

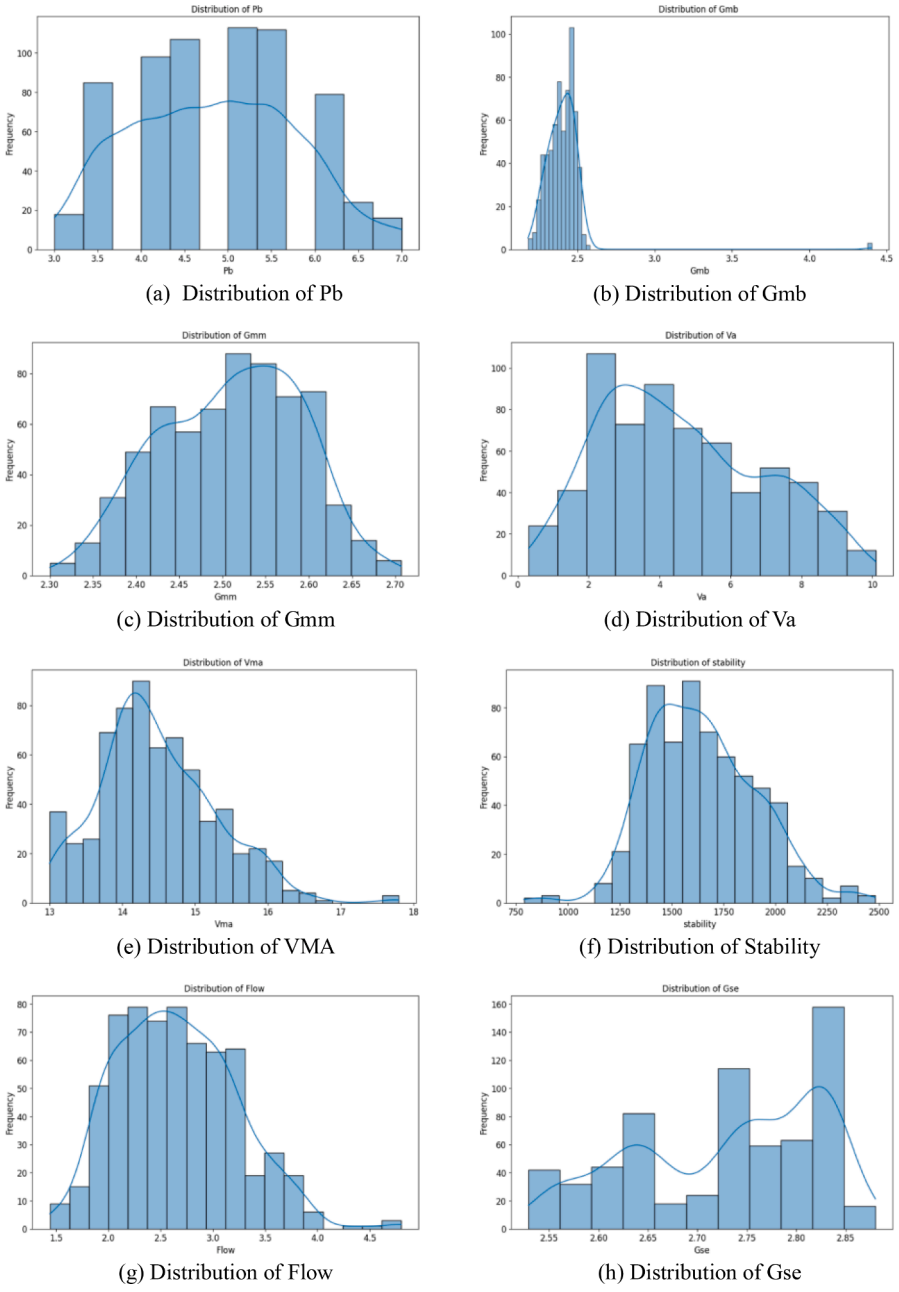
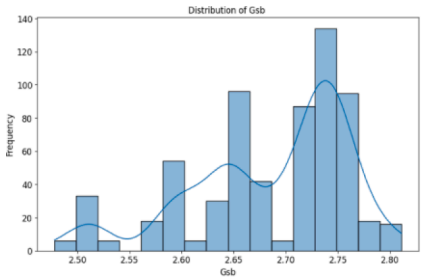
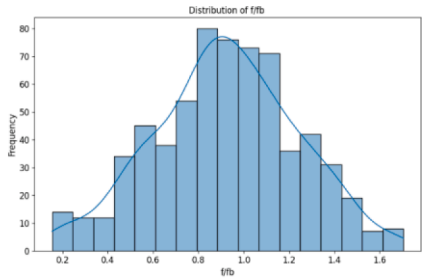


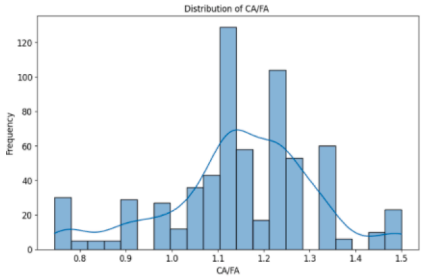
Fig. 2. Explanatory Variable Distribution.



(i) Distribution of Gsb



(j) Distribution of f/fb



(k) Distribution of CA/FA

Fig. 2. (*continued*).

models results, model performance, and explainable machine-learning results are presented.

3.1. Explanatory description analysis and variable selection

The first task in model building is to obtain a statistical summary of the variables and output correlation matrix. Two models were generated to predict each of the Marshall parameters (MS and MF). Table 1 shows a statistical summary of the variables and allows for an understanding of variable variation.

Table 1 lists the study variables, mean, median, standard deviation, coefficient of variance (dispersion), minimum and maximum (data extremes), skewness, and kurtosis (shapes of distribution), which make the interpretation of the datasets relatively straightforward. The numbers in Table 1 provide an understanding of the common material indices that influence the MS and MF. MS was found to range from 660 to 2482, and the MF range was found to range from 1.44 to 4.5. The distribution of all variables is shown in Fig. 2.

Fig. 2 displays the distributions of many crucial factors that are crucial for the Marshall mix design. Each variable is shown using histograms and kernel density estimates. The variable Pb, which represents the percentage of bitumen content, showed a uniform distribution roughly at a peak at value of 5%. The bulk specific gravity of the mix (Gmb) exhibited a distribution that is skewed to the right, indicating that most values were close to the lower values. This characteristic has the potential to affect the compaction and durability of the mix [13]. In contrast, the maximum specific gravity of the mix (Gmm) had a bimodal distribution, which signifies the presence of different aggregate or mix types that can impact the voids filled with asphalt and air voids (Va)—hence, altering the HMA performance [9]. The presence of many modes in the distribution of Va (air voids in the mix) indicated that the air void content varied, which in turn affected the mix's susceptibility to moisture damage and its durability [26]. The VMA exhibited a distribution that was skewed to the left, which is normally advantageous because it allows sufficient room for binder. However, if the VMA is too high, it may result in bleeding [38]. The presence of a multimodal distribution in Gsb and the effective specific gravity of aggregate (Gse) suggested that there is variability in the characteristics of the aggregate. These characteristics include physical ones such as porosity, permeable voids, shape, and texture, as well as chemical characteristics including the mineralogy and the composition of the aggregate particles, which are highly dependent on aggregate type and source. This variability can have an impact on the stability and compaction of the mixture, as noted by Fadhil et al. [11] and Yzenas [48]. The CA/FA ratio had a distribution with two distinct peaks, indicating the presence of different aggregate forms that enhanced stability but may necessitate the use of more binder for workability [16]. The F/b ratio was distributed normally, which is crucial for attaining the necessary stability and longevity of the mixture mix [7,15]. After the distribution of variables was plotted, variable selection was performed using the correlation matrix.

The second step involved the calculation of the Spearman rank coefficient matrix, which describes the strength of the relationship between variables. In the current study, models were generated for both MS and MF. Previous research demonstrated that integrating too many inputs with a low correlation with the desired output degrades model performance as complexity increases [1,44]. Multicollinearity, caused by the dependency of the input parameters, is a prevalent challenge in applications that employ machine-learning algorithms [41]. This can weaken the relationships between variables and reduce the strength of the models under development. To avoid this problem, it is recommended that the R-squared value between the two input parameters be smaller than 0.85 [23,46].

The correlation matrices for MS and MF are shown in Fig. 3(a) and 3 (b), respectively.

3.2. Modeling results

The Marshall performance was modeled by testing all PyCaret models. The PyCaret results were sorted using coefficients of determination. The five best models and their indicators are listed in Tables 2 and 3 for the testing dataset to predict MS and MF, respectively.

Tables 2 and 3 show that the CatBoost regressor testing results overcame the indicators of the different models to predict MS and MF. Intuitively, the best model is the one that had the lowest MAPE, MSE, RMSE, and RMSLE and the highest R^2. Model multicollinearity was checked using the residual plots for training and testing the dataset. Residual plots for both MS and MF are shown in Fig. 4(a) and 4(b) for MS and MF, respectively.

The following text illustrates the residuals for the CatBoost regressor when predicting MS and MF. The residual plots offer an evaluation of the

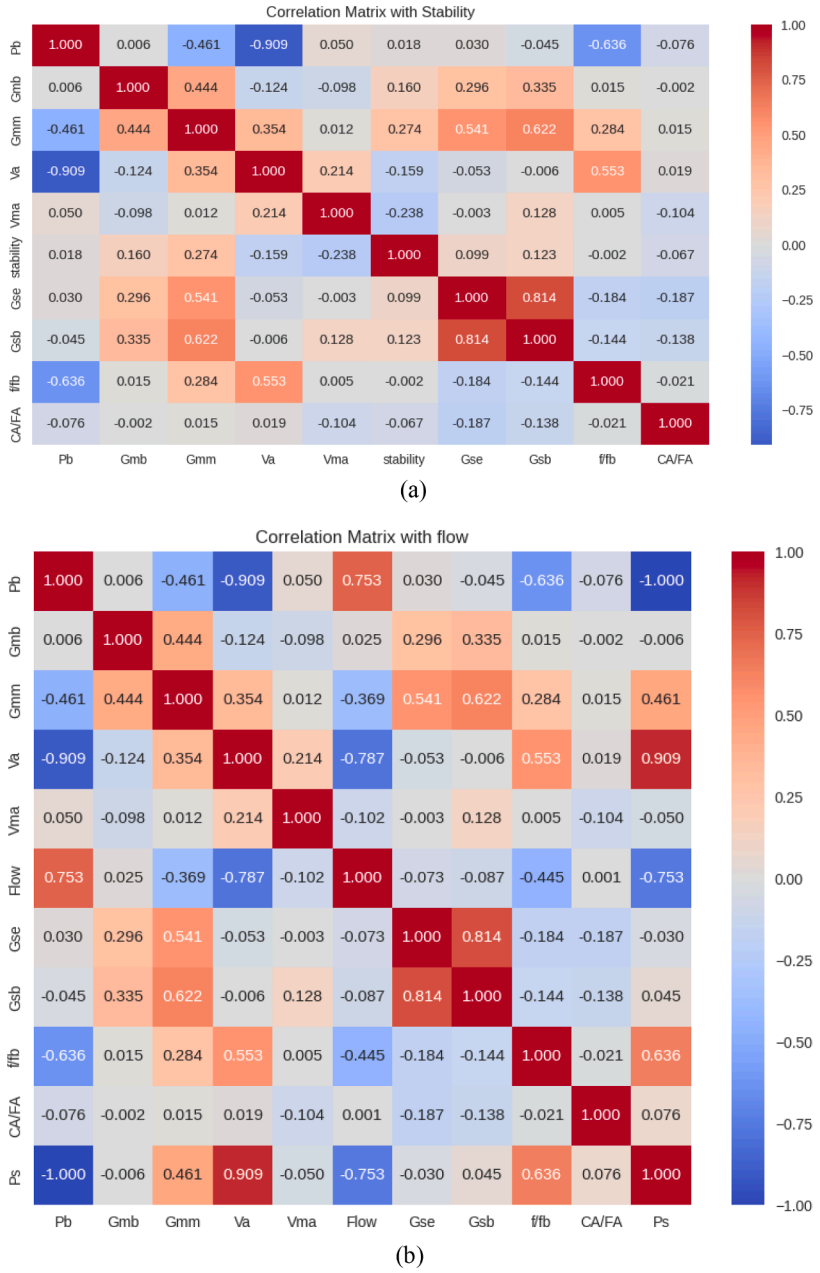


Fig. 3. Correlation Matrix: (a) Stability Correlation Matrix, (b) Flow Correlation Matrix.

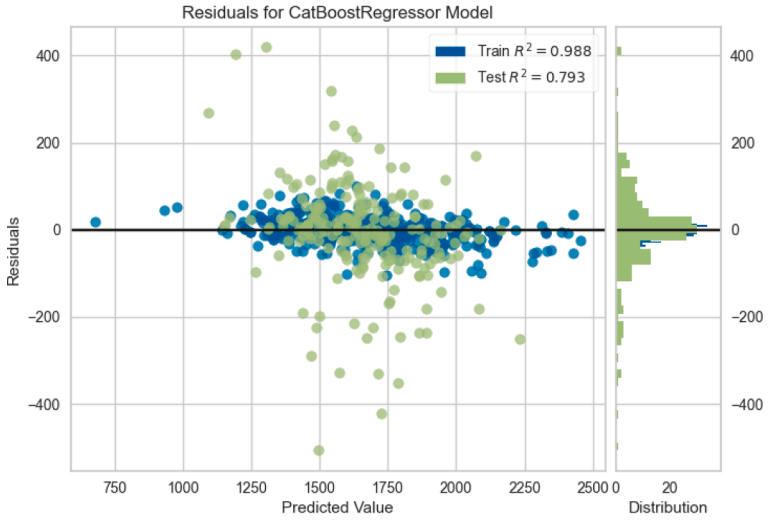
Table 2
Machine-learning model performance to predict Marshall stability for testing dataset.

	Model	MAE	MSE	RMSE	R^2	RMSLE	MAPE
CatBoost	CatBoost regressor	91.8499	17,113.3635	128.6616	0.793	0.0818	0.0576
XGBoost	Extreme gradient boosting	95.0633	20,336.5004	138.7943	0.7253	0.0881	0.0596
LightGBM	Light gradient-boosting machine	106.9091	22,394.9875	147.4065	0.6914	0.0931	0.0668
ET	ExtraTrees regressor	106.8980	23,770.2273	152.7934	0.6662	0.0958	0.0662
GBR	Gradient-boosting regressor	120.8133	26,273.4639	159.8434	0.6362	0.1001	0.0759

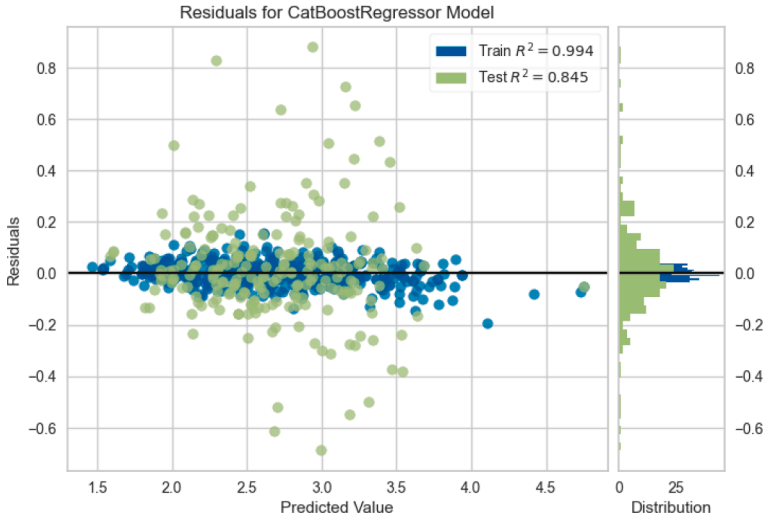
model's performance by examining the errors in its predictions for both the training and test datasets. The scatter plot depicts the residuals in relation to the predicted values, where the blue and green dots represent the training and test data residuals, respectively. The horizontal lines at zero on the plots indicate perfect predictions. The model exhibits high R^2 values, with 0.988 and 0.994 for the training data in MS and MF, respectively. However, the R^2 values for the testing data were 0.794 and 0.845 for MS and MF, respectively. Both subplots indicate that the residuals are symmetrically distributed around zero, suggesting that the models' predictions are unbiased and that the errors are normally distributed. The consistent spread of residuals across predicted values suggests that the models' errors have a constant variance, an advantageous trait. Although some outliers with larger residuals are present, the overall pattern indicates that the model performed well, maintaining a

Table 3
Machine-learning model performance to predict Marshall flow.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
CatBoost	CatBoost regressor	0.1393	0.0394	0.1947	0.8450	0.0525	0.0537
LightGBM	Light gradient-boosting machine	0.1570	0.0480	0.2162	0.8411	0.0576	0.0598
GBR	Gradient-boosting regressor	0.1721	0.0529	0.2273	0.8264	0.0611	0.0660
XGBoost	Extreme gradient boosting	0.1508	0.0525	0.2253	0.8237	0.0606	0.0576
ET	ExtraTrees regressor	0.1670	0.0597	0.2406	0.7997	0.0649	0.0642



(a)



(b)

Fig. 4. Model Residuals for Training and Testing Data: (a) Predicted Stability, (b) Flow Prediction.

homoscedastic error distribution and providing reliable predictions. Both MS and MF k-fold distributions are shown in Fig. 5.

Fig. 5 shows the MS and MF model CV results, in which the model accuracy increased with an increase in the training dataset. Fig. 5(a) shows that the learning curve for the training score remained consistently high (near 1.0) across all training instances, indicating an excellent fit to the training data but suggesting potential overfitting. The CV score for this model steadily increased from around 0.4 to about 0.8 as more training instances were added, indicating improved performance on unseen data. In contrast, the learning curve for predicting MF in Fig. 5 (b) shows that the training score started at around 0.9 and gradually increased to just above 0.95, suggesting a slightly higher bias but better overall generalization. The CV score for the MF prediction model also steadily increased from around 0.7 to about 0.85 with more training data. Both models benefited from additional training data, with CV scores consistently improving, underscoring the importance of sufficient training instances for enhancing model performance. After validating the model, we estimated the model testing error against the predicted value. The prediction error results are shown in Fig. 6.

Fig. 6 shows the correlation between the predicted and actual values. The coefficient of determination was computed and found to be 0.83 for stability prediction and 0.845 for flow prediction, indicating a highly correlated value. Another measure was used to model the importance of variables using Morris sensitivity analysis.

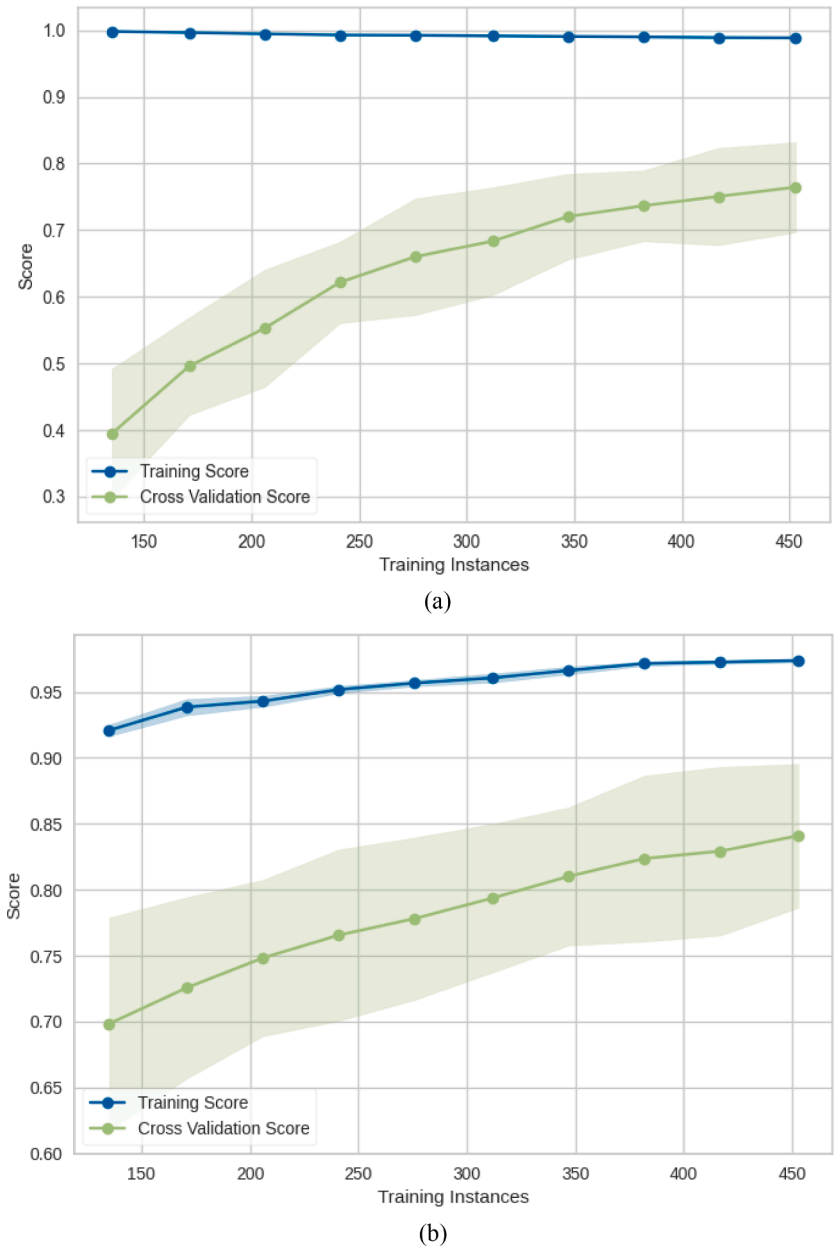


Fig. 5. 10-Fold Cross-Validation Results: (a) Stability Model, (b) Flow Model.

3.3. *Explainable machine-learning results*

The important variables are shown in Fig. 7. The convergence index indicates that the model is stable. Moreover, the results indicate that the VA and the CA/FA ratio had the strongest impact on MS prediction.

Fig. 7 reveals the relative importance of various input variables on the model outcomes, providing numerical sensitivity indices. As shown in Fig. 7(a), Va stood out as the most influential variable with a sensitivity index of approximately 0.50. Ps and Pb followed closely with sensitivity indices of around 0.30 and 0.28, respectively. CA/FA and Gmb exhibited moderate sensitivity, with indices of approximately 0.22 and 0.18. Other variables (f/fb, Gsb, VMA, Gse, and Gmm) had lower sensitivity indices, ranging from 0.10 to 0.15, indicating a lesser impact on the MS compared with Ps and Pb.

As shown in Fig. 7(b), Va remained the most sensitive variable, with a significantly higher sensitivity index of approximately 120. CA/FA followed, with an index of around 100, indicating a substantial impact on the model. Gmb, f/fb, and Gmm also showed considerable sensitivity, with indices of approximately 80, 75, and 70, respectively. The remaining variables —VMA, Gse, Pb, Ps, and Gsb—exhibited lower sensitivity indices, ranging from 40 to 60. The convergence indices of 0.095 and 0.097 confirmed the stability and reliability of these sensitivity measurements, underscoring the critical role of Va, Ps, Pb, CA/FA, and Gmb in influencing MF. A more sophisticated plot showing the interaction between different variables and MS was constructed using SHAP, as shown in Fig. 8.

Fig. 8(a) shows the impact of changing the different features on MS. It can be seen that the increases in Gsb, Gmb, and Gmm are associated with an increase in MS. Additionally, increases in Gse, CA/FA, VMA, and Gse levels were associated with a decrease in MS. A more detailed investigation of the interaction between the four most important features and MS and MF is shown in Figs. 9 and 10 for MS and MF, respectively.

The interdependency plots for the variables Va, Pb, Ps, and CA/FA in relation to MS (average response) provide insightful visualizations of how these variables influence the stability measure. Each subplot

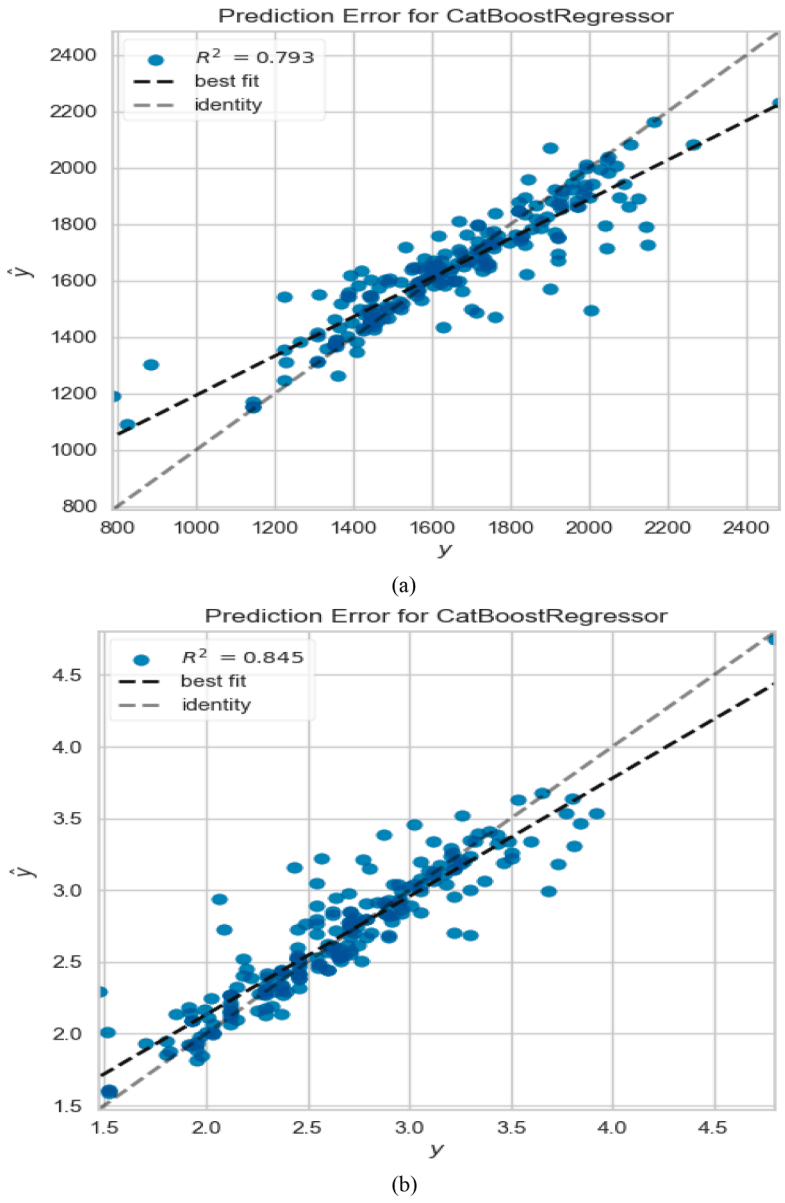


Fig. 6. Model Performance: (a) Stability Prediction, (b) Flow Prediction.

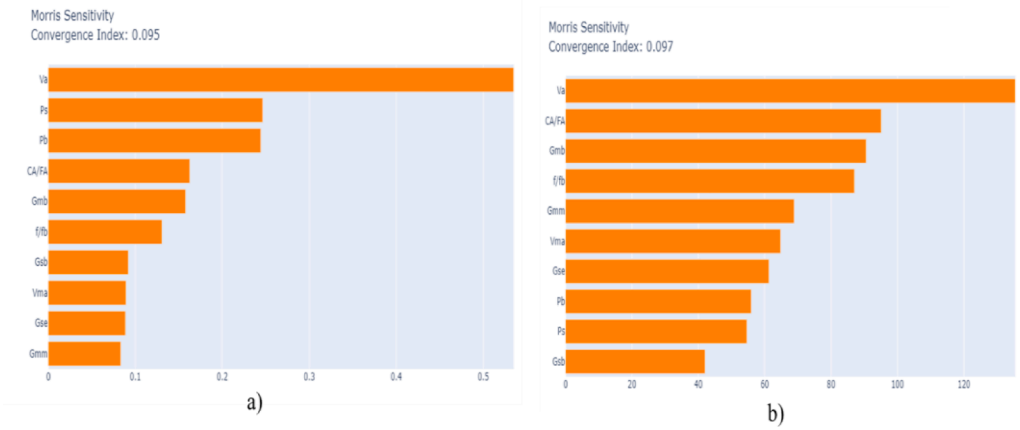


Fig. 7. Morris Sensitivity Results: (a) Marshall Model Results, (b) Flow Model Results.

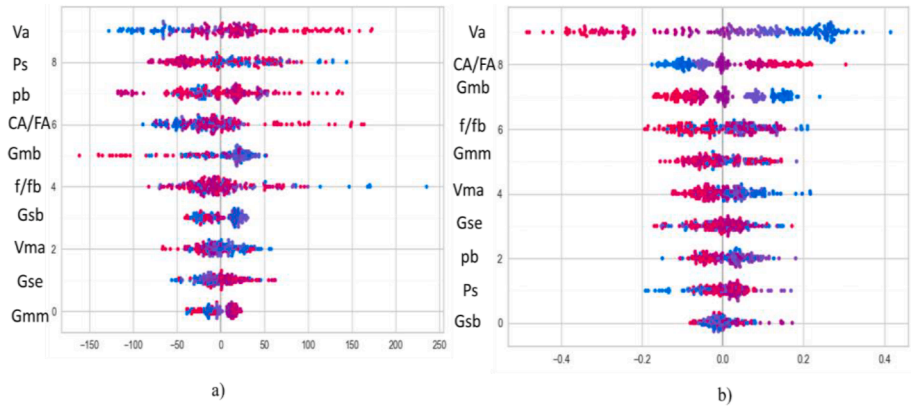


Fig. 8. SHAP: (a) Marshall Stability, (b) Marshall Flow.

consists of a line graph showing the trend of the average response across different values of the variables and a histogram illustrating the density distribution of these values. For Va, the line plot indicates that MS remained relatively stable with slight variations, but there is a minor downward trend as Va increased. The histogram shows that values of around 1.97–2.81 and 5.32–6.15 are more frequent, suggesting that these ranges are significant in the dataset. Similarly, the Pb plot shows a generally stable average response, with slight decreases as the variable increases and higher densities in the range of 4–5.5. For Ps, the average response is consistent, with minor declines noted toward higher values, and the histogram highlights higher frequencies of around 94.5–95.5. Finally, the CA/FA plot shows a stable response, with minor dips around the value of 1. The histogram indicates more frequent occurrences in the range of 0.946–1.14, suggesting that these ranges are critical.

The interdependency plots for MF reveal how different variables—namely, Va, CA/FA, bulk Gmb, and f/fb ratio, influenced the average MF response. For Va, the average MF showed a negative correlation, decreasing from around 3 to 2 as Va increased, with higher density observed in the ranges 1.97–2.81 and 5.32–6.15. The CA/FA ratio plot indicates a relatively stable average MF between 2.5 and 3, with a minor dip around 1.2, and higher densities between 0.965 and 1.16. In the case of Gmb, the average MF showed a slight downward trend from 2.5 to 3 as Gmb increased, with a significant density peak at 2.33. Finally, for f/fb, the average MF remained stable, at around 2.5–3 across different ratios, with higher densities between 0.992 and 1.13. The plots collectively highlight the subtle influences of these variables on MF, which is essential for optimizing asphalt mix designs by identifying prevalent conditions and understanding their impact on flow characteristics. The effects of the most important features on the predicted MS and MF are shown in Figs. 11 and 12, respectively.

Fig. 11 shows the SHAP plot for predicting MS. The figure reveals the contributions of various features to the final prediction of 1541.17, starting from a base value of 1660. The theoretical maximum specific gravity (Gmm) at 2.654 significantly lowered the prediction, pulling it down to 2449, indicating a strong negative impact on stability. Other features, such as Gsb (3.017), GSE (2.748), and the bitumen content by weight of mix (3.5%), also contributed to reducing the stability prediction. Conversely, the CA/FA ratio of 1.273 slightly increased the predicted stability. Overall, the plot highlights that while multiple factors collectively reduced the predicted MS, Gmm had the most substantially negative influence, providing insights into which factors need adjustment to enhance stability.

Fig. 12 shows the SHAP plot for predicting MF. The figure demonstrates the influence of various features on the prediction, starting from a base value of 2.655 and resulting in a predicted value of 2.34. Va, at 3.7, had a significant negative impact, reducing the prediction to 2.34. Other features, including CA/FA at 0.5214, Gse at 2.551, Gmb at 2.327, f/fb at 1.1, Gsb at 2.507, and VMA at 14.3, all contributed to further decreasing the MF prediction. The plot highlights that Va is the most influential factor negatively impacting the predicted MF, while the other factors also contributed to the overall reduction, though to a lesser extent. Understanding these contributions is crucial for adjusting the mix design to achieve the desired MF characteristics.

The successful application of machine-learning models, including CatBoost, LightGBM, XGBoost, and ExtraTrees, for predicting MS and MF produces a promising framework for enhancing the efficiency of pavement design processes. These models yield highly accurate predictions, thereby reducing reliance on time-intensive and costly laboratory tests traditionally required for asphalt mix design. By employing these models, engineers and decision-makers can preselect design variables with increased confidence, thus streamlining the mix design process and minimizing the necessity for repetitive testing. This approach may result in cost reductions and shortened project timelines, as well as facilitate sustainable construction practices through the optimization of resource utilization. The explainable machine-learning techniques employed, such as SHAP analysis, also provide insights into the influence of various mix parameters, enabling engineers to make data-driven decisions when modifying designs or investigating performance failures.

These machine-learning models can be integrated into current pavement design and maintenance practices through incorporation into existing pavement management systems (PMS). By automating the prediction of asphalt mix performance, engineers can expeditiously evaluate various design scenarios and identify optimal solutions with minimal manual intervention. Furthermore, by integrating these models into the design phase, real-time feedback can be provided during construction to assess material performance, allowing for on-site adjustments to optimize pavement quality and longevity. In maintenance practices, these models can be utilized to forecast pavement deterioration and predict the timing of necessary repairs or overlays, enabling more proactive and data-driven maintenance scheduling.

A significant limitation of machine-learning models is their dependence on the quality and quantity of data. The accuracy of predictions is contingent on the availability of high-quality, representative datasets. Consequently, the application of these models may be constrained in regions where historical data on pavement performance are limited or inconsistent. Moreover, given that the data utilized in this study were collected from projects in Jordan, there may be challenges in generalizing the findings to other geographic regions with distinct environmental conditions, traffic loads, or material sources.

4. Conclusions and discussion

In this study, several artificial intelligence techniques were embedded and pretrained using the PyCaret library in Python to predict the MS and MF of different asphalt mixes. These techniques included LightGBM, XGBoost, ExtraTree, and gradient boosting. The databases

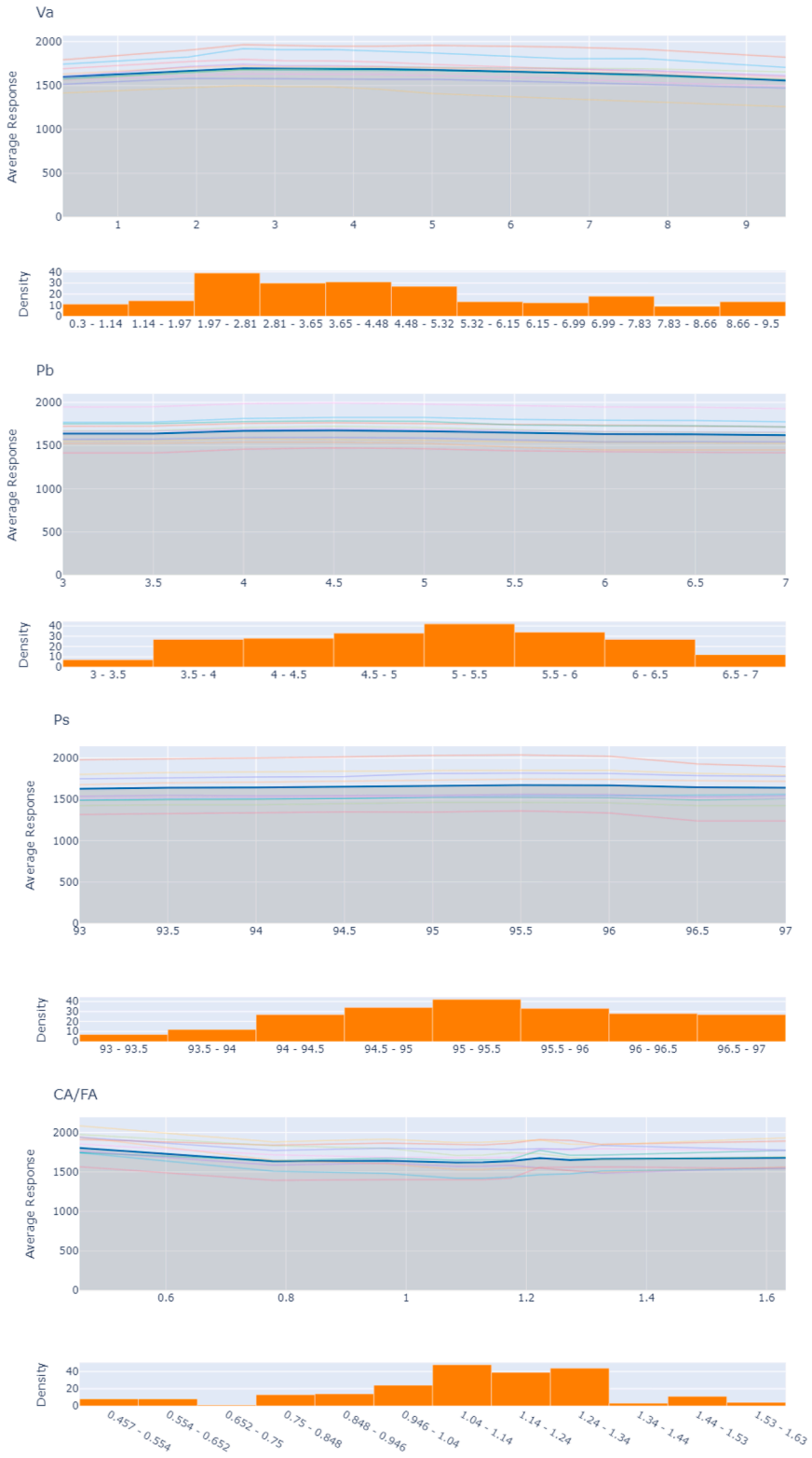


Fig. 9. Interdependencies Relationship for the Predicted MS and the Five Most Important Features.

for MS and MF were constructed from an extensive collection of results from numerous road-construction mixes across Jordan. The CatBoost regression model achieved high R^2 values of 0.793 and 0.845 for MS and MF, respectively, confirming its robustness and reliability. The SHAP values highlighted the critical roles of Gmm and Va, among other features, in influencing the predictions, thereby providing valuable insights for optimizing mix designs. The findings reveal that Gmm had the most considerable negative impact on MS, pulling the prediction down to 2449 from a base value of 1660. Similarly, Va significantly reduced MF predictions to 2.34 from a base value of 2.655. These insights are crucial for making informed adjustments to mix designs, ultimately promoting sustainable construction practices by reducing the need for extensive laboratory testing and improving the efficiency of the mix design process.

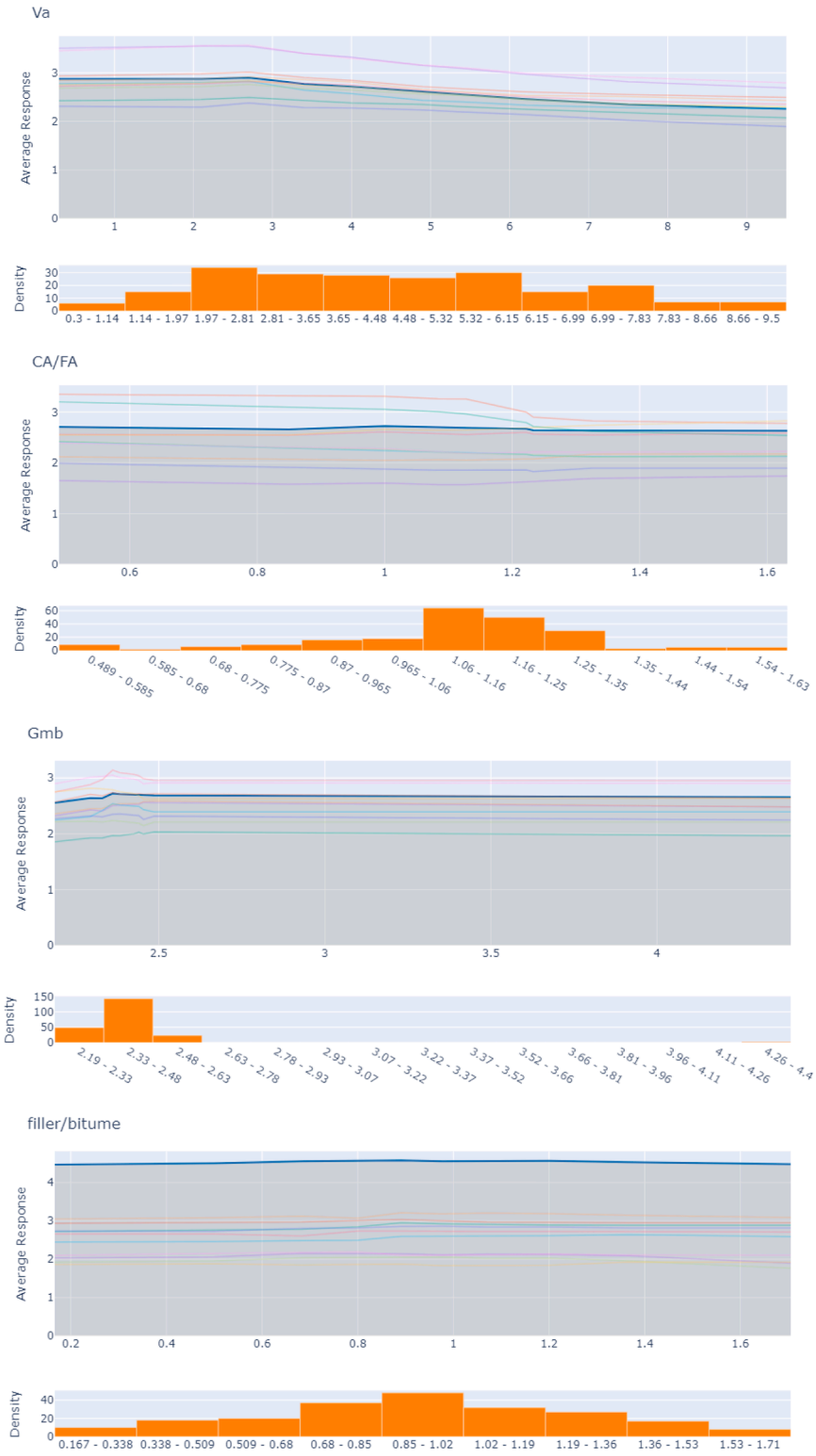


Fig. 10. Interdependency Relationships for the Predicted MF and the Five Most Important Features.

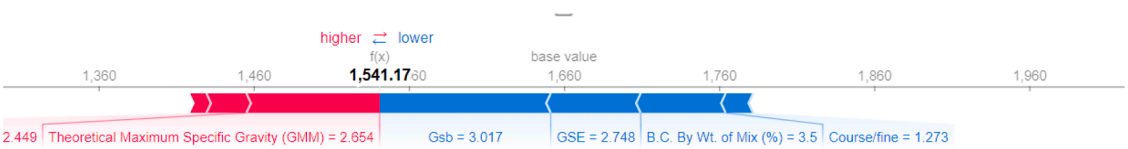


Fig. 11. SHAP Values for Marshall Stability Prediction.

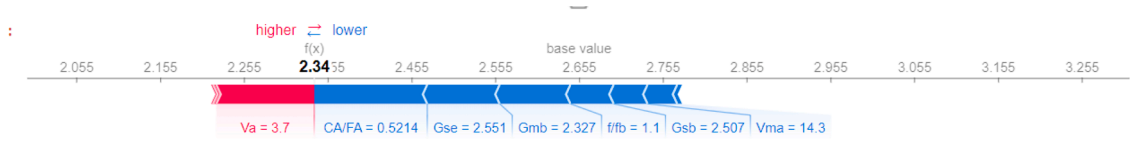


Fig. 12. SHAP Values for Marshall Stability Prediction.

4.1. Limitations

Despite its promising results, this study has several limitations. The dataset, although comprehensive, is specific to construction projects in Jordan, which may limit the generalizability of the findings to other regions with different climatic conditions and for other material properties. Additionally, the study focuses on the CatBoost regression model; while this model showed superior performance, exploring other advanced machine-learning models could provide further improvements. The reliance on historical data also implies that any changes in material sources or testing boundaries could affect the model's accuracy, necessitating continuous updating and validation.

4.2. Future work

Future work should aim to address these limitations by expanding the dataset to include diverse geographical regions and varying environmental conditions, ensuring broader applicability of the model. Exploring other advanced machine-learning algorithms, such as deep-learning models, could further enhance prediction accuracy. Integrating real-time data could also improve the model's responsiveness to changes in material properties and testing conditions. Additionally, developing user-friendly software tools based on these models can facilitate their adoption in the field, providing practitioners with easy access to predictive insights and optimizing the mix design process. Finally, conducting longitudinal studies to validate the long-term performance of the predicted mix designs in real-world conditions would further strengthen the findings and their practical relevance.

Funding statement

The study was not supported by any agency.

Data access statement

Research data are available upon a reasonable request from the corresponding author.

CRediT authorship contribution statement

Ibrahim Asi: Writing – review & editing, Supervision, Project administration, Investigation. **Yusra I. Alhadidi:** Writing – review & editing, Visualization, Investigation, Conceptualization. **Taqwa I. Alhadidi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Data availability

Data will be made available on request.

References

[1] T. Abunama, F. Othman, M. Ansari, A. El-Shafie, Leachate generation rate modeling using artificial intelligence algorithms aided by input optimization method for an MSW landfill, Environ. Sci. Pollut. Res. 26 (2019) 3368–3381.
[2] A. Aksoy, E. Iskender, H.T. Kahraman, Application of the intuitive k-NN Estimator for prediction of the Marshall Test (ASTM D1559) results for asphalt mixtures, Constr. Build. Mater. 34 (2012) 561–569.
[3] M. Ali, PyCaret: An open source, low-code machine learning library in Python. *PyCaret Version, 2*, 2020.
[4] F. Althoey, M.N. Akhter, Z.S. Nagra, H.H. Awan, F. Alanazi, M.A. Khan, M.F. Javed, S.M. Eldin, Y.O. Özkılıç, Prediction models for marshall mix parameters using bio-inspired genetic programming and deep machine learning approaches: A comparative study, Case Stud. Construct. Mater. 18 (2023) e01774.
[5] M. Atakan, K. Yıldız, Prediction of Marshall design parameters of asphalt mixtures via machine learning algorithms based on literature data, Road Mater. Pavement Des. 25 (3) (2024) 454–473.
[6] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, Front. Big. Data 4 (2021) 688969.
[7] M. Chaudhary, N. Saboo, A. Gupta, Development of the Criteria for Optimum Filler–Binder Ratio in an Asphalt Mix Based on Fatigue Performance, J. Mater. Civil Eng. 36 (9) (2024) 04024254.
[8] D. Das, Y. Nishimura, R.P. Vivek, N. Takeda, S.T. Fish, T. Ploetz, S. Chernova, Explain. Activity Recogn. Smart Home Systems (2021), https://doi.org/10.48550/arxiv.2105.09787.
[9] M El Sayed, Effect of changing theoretical maximum specific gravity on asphalt mixture design, Eng. J. 16 (4) (2012) 137–148.
[10] Ersöz, T., & Ersöz, F. (2022). Data Mining and Machine Learning Approaches in Data Science: Predictive Modeling of Traffic Accident Causes. *International Journal of 3d Printing Technologies and Digital Industry*. 10.46519/ij3dptdi.1199614.
[11] T.H. Fadhil, H.A. Mohammed, T.Y. Ahmed, An empirical relationship between asphalt and water absorption of coarse aggregates in HMA, J. Eng. Sustain. Dev. 19 (1) (2015) 132–146.
[12] A. Garcia-Hernandez, L. Wan, S. Dopazo-Hilario, In-silico manufacturing of asphalt concrete, Powder. Technol. 386 (2021) 399–410.
[13] D. Gardete, L. Picado-Santos, S. Capitão, R. Luzia, Asphalt mix design: Discussion on the bulk specific gravity procedure influence on the results obtained from empirical, volumetric, and performance-based methods, Constr. Build. Mater. 342 (2022) 127870.
[14] M.A. Gul, M.K. Islam, H.H. Awan, M. Sohail, A.F. Al Fuhaid, M. Arifuzzaman, H. J. Qureshi, Prediction of Marshall stability and marshall flow of asphalt pavements using supervised machine learning algorithms, Symmetry. (Basel) 14 (11) (2022) 2324.
[15] A. Hamidi, A. Motamed, The effect of filler type and content on rutting resistance of asphaltic materials, International Journal of Pavement Research and Technology 12 (2019) 249–258.
[16] M. Hashemi, P. Shafigh, M.R.Bin Karim, C.D Atis, The effect of coarse to fine aggregate ratio on the fresh and hardened properties of roller-compacted concrete pavement, Constr. Build. Mater. 169 (2018) 553–566.
[17] A. Holzinger, From machine learning to explainable AI, in: 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), 2018, pp. 55–66.
[18] S.K. Jagatheesaperumal, Q.-V. Pham, R. Ruby, Z. Yang, C. Xu, Z. Zhang, Explainable AI Over the Internet of Things (IoT): Overview, State-of-the-Art and Future Directions, Ieee Open Journal of the Communications Society 3 (2022) 2106–2136, https://doi.org/10.1109/ojcoms.2022.3215676.
[19] Y. Jiang, C. Deng, J. Xue, Z. Chen, Investigation into the performance of asphalt mixture designed using different methods, Constr. Build. Mater. 177 (2018) 378–387.
[20] C. Jin, Y. Feng, X. Yang, P. Liu, Z. Ding, M. Oeser, Virtual design of asphalt mixtures using a growth and contact model based on realistic aggregates, Constr. Build. Mater. 320 (2022) 126322.
[21] Y.S. Jweihan, R.J. Alawadi, Y.S. Momani, A.N. Tarawneh, Prediction of marshall test results for dense glasphalt mixtures using artificial neural networks, Front. Built. Environ. 8 (2022) 949167.
[22] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nature Reviews Physics 3 (6) (2021) 422–440.
[23] J.H. Kim, Multicollinearity and misleading statistical results, Korean J. Anesthesiol. 72 (6) (2019) 558–569.
[24] D.P. Kroese, Z. Botev, T. Taimre, Data Science and Machine learning: Mathematical and Statistical Methods, Chapman and Hall/CRC, 2019.
[25] M. Krötzsch, Ontologies for knowledge graphs?. Description Logics, 2017.
[26] R.P. Leandro, K.L. Vasconcelos, L.L.B. Bernucci, Evaluation of the laboratory compaction method on the air voids and the mechanical behavior of hot mix asphalt, Constr. Build. Mater. 156 (2017) 424–434.

[27] Y. Li, L. Wang, Computer-aided procedure for determination of asphalt content in asphalt mixture using discrete element method, International Journal of Pavement Engineering 18 (9) (2017) 765–774.
[28] J. Liu, F. Liu, C. Zheng, D. Zhou, L. Wang, Optimizing asphalt mix design through predicting the rut depth of asphalt pavement using machine learning, Constr. Build. Mater. 356 (2022) 129211.
[29] H. Majidifard, B. Jahangiri, W.G. Buttlar, A.H. Alavi, New machine learning-based prediction models for fracture energy of asphalt mixtures, Measurement 135 (2019) 438–451.
[30] H. Majidifard, B. Jahangiri, P. Rath, L.U. Contreras, W.G. Buttlar, A.H. Alavi, Developing a prediction model for rutting depth of asphalt mixtures using gene expression programming, Constr. Build. Mater. 267 (2021) 120543.
[31] M. Miani, M. Dunnhofer, F. Rondinella, E. Manthos, J. Valentin, C. Micheloni, N. Baldo, Bituminous mixtures experimental data modeling using a hyperparameters-optimized machine learning approach, Applied Sciences 11 (24) (2021) 11710.
[32] M.D. Morris, Factorial sampling plans for preliminary computational experiments, Technometrics. 33 (2) (1991) 161–174.
[33] A.C. Müller, S. Guido, Introduction to Machine Learning With Python: a Guide For Data Scientists, O'Reilly Media, Inc., 2016.
[34] P.I. Nakagawa, L.F. Pires, J.L.R. Moreira, L.O.B. Santos, S. da, F.A. Bukhsh, Semantic Description of Explainable Machine Learning Workflows for Improving Trust, Applied Sciences (2021), https://doi.org/10.3390/app112210804.
[35] Nor, A.K.M., Pedapati, S.R., Muhammad, M., & Leiva, V. (2021). *Explainable Artificial Intelligence for Anomaly Detection and Prognostic of Gas Turbines Using Uncertainty Quantification With Sensor-Related Data*. 10.20944/preprints202109.0034.v2.
[36] F.M.A. Olaniyan, A. Owoseni, Toward Improved Data Quality in Public Health: Analysis of Anomaly Detection Tools applied to HIV/AIDS Data in Africa, in: 2022 IST-Africa Conference (IST-Africa), 2022, pp. 1–9.
[37] H.I. Ozturk, M.E. Kutay, An artificial neural network model for virtual Superpave asphalt mixture design, International Journal of Pavement Engineering 15 (2) (2014) 151–162.
[38] M.R. Pouranian, J.E. Haddock, Determination of voids in the mineral aggregate and aggregate skeleton characteristics of asphalt mixtures using a linear-mixture packing model, Constr. Build. Mater. 188 (2018) 292–304.
[39] S. Rahman, A. Bhasin, A. Smit, Exploring the use of machine learning to predict metrics related to asphalt mixture performance, Constr. Build. Mater. 295 (2021) 123585.
[40] J.-R. Rehse, N. Mehdiyev, P. Fettke, Towards Explainable Process Predictions for Industry 4.0 in the DFKI-Smart-Lego-Factory, Ki - Künstliche Intelligenz 33 (2) (2019) 181–187, https://doi.org/10.1007/s13218-019-00586-1.
[41] M. Rekha, MLmuse: Correlation and Collinearity—How they can make or break a model. Correlation Analysis and Collinearity| Data Science| Multicollinearity| Clairvoyant Blog (Clairvoyantsoft. Com), 2019.
[42] K. Sandamal, Pavement Roughness Prediction Using Explainable and Supervised Machine Learning Technique for Long-Term Performance, Sustainability. 15 (12) (2023) 9617, https://doi.org/10.3390/su15129617.
[43] H. Sebaaly, S. Varma, J.W. Maina, Optimizing asphalt mix design process using artificial neural network and genetic algorithm, Constr. Build. Mater. 168 (2018) 660–670.
[44] M.I. Shah, M.F. Javed, T. Abunama, Proposed formulation of surface water quality and modelling using gene expression, machine learning, and regression techniques, Environmental Science and Pollution Research 28 (2021) 13202–13220.
[45] S. Shihui, Y. Huanan, Analysis of Aggregate Gradation and Packing for Easy Estimation of Hot-Mix-Asphalt Voids in Mineral Aggregate, Journal of Materials in Civil Engineering 23 (5) (2011) 664–672, https://doi.org/10.1061/(ASCE)MT.1943-5533.0000224.
[46] N. Shrestha, Detecting multicollinearity in regression analysis, Am. J. Appl. Math. Stat. 8 (2) (2020) 39–42.
[47] E. Veitch, O.A. Alsos, Human-Centered Explainable Artificial Intelligence for Marine Autonomous Surface Vehicles, J. Mar. Sci. Eng. 9 (11) (2021) 1227, https://doi.org/10.3390/jmse9111227.
[48] J.J. Yzenas, Bulk density, relative density (specific gravity), pore structure, absorption, and surface moisture. Significance of Tests and Properties of Concrete and Concrete-Making Materials, ASTM International, 2006.
[49] C. Zhang, Z. Xiong, C. Lin, Fusing XGBoost and SHAP Models for Maritime Accident Prediction and Causality Interpretability Analysis, J. Mar. Sci. Eng. 10 (8) (2022) 1154, https://doi.org/10.3390/jmse10081154.
[50] S. Jeong, H. Park, A study on predictive factors of multiple citizenship among adolescents: using automated machine learning and SHAP, Koreann J. Educ. Res. 61 (8) (2023) 57–87.