applied sciences

MDPI

Article

Multivariate Analysis for Prediction of Splitting Tensile Strength in Concrete Paving Blocks

Vinicio R. Benalcázar-Rojas [1,*], Wilman J. Yambay-Vallejo [1,2] and Erick P. Herrera-Granda [1]

[1] Universidad Politécnica Estatal del Carchi, Posgrado, Av. Universitaria y Antisana, Tulcán 040101, Carchi, Ecuador; wilman.yambay@upec.edu.ec (W.J.Y.-V.); erick.herrera@upec.edu.ec (E.P.H.-G.)
[2] Facultad de Industrias Agropecuarias y Ciencias Ambientales, Universidad Politécnica Estatal del Carchi, Tulcán 040101, Ecuador
* Correspondence: vinicio.benalcazar@upec.edu.ec; Tel.: +593-982120317

Abstract: Paving blocks are concrete pieces exposed to the weather and subjected to loads and wear. Hence, quality control in their manufacture is essential to guarantee their properties and durability. In Ecuador, the requirements are described in the technical standard "NTE INEN 3040", and tensile splitting strength is a fundamental requirement to guarantee product quality. The objective of the study is to predict the tensile splitting strength using two groups of predictor variables. The first group is the thickness in mm, width in mm, length in mm, mass of the fresh paving block in g, and percentage of water absorption; the second group of predictor variables is the density of the fresh paving block in kg/m^3 and the percentage of water absorption. The data were obtained from a company that can produce 30,000 units per day of rectangular paving blocks with 6 cm thickness. The research involves sampling, analysis of outliers, descriptive and inferential statistics, and the analysis of multivariate models such as multiple linear regression, regression trees, random forests, and neural networks. It is concluded that the multiple linear regression method performs better in predicting the first group of predictor variables with a mean square error (MSE) of 0.110086, followed by the neural network without hidden layers, resulting in an MSE of 0.112198. The best method for the second set of predictors was the neural network without hidden layers, with a mean square error (MSE) of 0.112402, closely followed by the multiple linear regression model, with an MSE of 0.115044.

Keywords: multivariate analysis; prediction of tensile splitting strength; quality in concrete paving blocks; density of the fresh paving block; water absorption of concrete paving blocks; weight of the fresh paving blocks

Citation: Benalcázar-Rojas, V.R.; Yambay-Vallejo, W.J.; Herrera-Granda, E.P. Multivariate Analysis for Prediction of Splitting Tensile Strength in Concrete Paving Blocks. *Appl. Sci.* **2023**, *13*, 10956. https://doi.org/10.3390/app131910956

Academic Editor: Luís Picado Santos

Received: 26 August 2023
Revised: 22 September 2023
Accepted: 27 September 2023
Published: 4 October 2023

1. Introduction

Concrete is a mixture that has revolutionized construction. Standard [1] defines it as a material composed of a binding component with embedded particles and aggregates. However, producing high-quality paving blocks on a large scale poses challenges. The batch-type paving block manufacturing process involves mixing, vibro-compacting, curing, and palletizing. In process control, a company with high production volumes in Quito-Ecuador has quality controls with specialists in raw materials, intermediate goods, and finished goods for each batch of 30,000 paving blocks. The vibro-compacting machine compacts the mixture of cement, water, aggregates, and additives that enters the mold and uses a tamper that shapes the fresh paving blocks, for which the dimensions, the weight of the vibro-compacted paving block, and the water absorption test are means of controlling the product quality initially.

The tensile splitting strength test was described in standard [2], but its scope is limited to concrete cylinders. The Ecuador standard [1] is applied for paving blocks and replaces standard [3], where the compressive strength test is omitted from the tensile splitting strength test. The article described by [4] indicates how the paving blocks break on site in

their daily use, not by compression but by fracture (splitting); for this reason, the tensile splitting strength test was recommended to measure its ability to resist stresses.

The tensile splitting strength is a crucial quality characteristic for customers, indicating the material's breaking point and useful life. However, this property is tested 28 days after manufacture. Predicting tensile strength earlier would enable manufacturers to avoid nonconforming paving blocks due to limited vision regarding this property. This study investigates how tensile strength depends on two groups of predictor variables in the production process. The first group is the thickness in mm, width in mm, length in mm, mass of the fresh paving block in g, and percentage of water absorption; the second group of predictor variables is the density of the fresh paving block in kg/m^3 and the percentage of water absorption. Predictive modelling with statistical validation enables acceptable error when forecasting tensile strength using these explanatory variables.

A statistical sampling is carried out to infer 30,000 units of the rectangular paving block with 6 cm thickness (study population) representing a production batch. Specific objectives are identifying multivariate techniques for this case study, implementing predictive models, and comparing methodologies to determine the best model for predicting tensile splitting strength. To perform accurate statistical analysis predicting a continuous dependent variable from multiple continuous independent variables, we must verify hypotheses, infer the population, take representative samples, process data, identify outliers, build models, and check assumptions.

Related Work

Predicting concrete tensile splitting strength from related properties has been an important research focus. Simple linear regression models provide initial correlations but limited analysis. Several studies used linear and nonlinear regression models to predict splitting strength. The research by [5] used simple linear regression to correlate tensile splitting strength with compressive strength, water/binder ratio, and age. While simple, this limited analysis of related variables. Nonlinear equations proposed increasing tensile strength slower than compression. The investigation in [6] related abrasion and splitting strength to bulk density and ultrasonic pulse velocity, finding that the logarithmic transform of dry bulk specific better predicted splitting strength than the linear model.

The study that was carried out by [4] indicated that the compression test described in Standard NS 6717:1986, which is carried out in the laboratory, does not represent the real behavior in the paving blocks in real conditions of use and is caused by splitting stresses, which generates a fracture by dividing the paving block into two parts. Compressive and tensile strengths were related through regression. Compressive strength is also related to density. The study carried out by [7] elaborated mixes with recycled aggregates, establishing compressive strength's dependence on density. Ref. [8] related density and compression to rubber content through linear and logarithmic regression, with increased rubber reducing both. High compressive strengths occurred alongside high densities. The prediction by multiple linear regression of the compressive strength can be seen in the study carried out by [9], where it is indicated that the compressive and traction strength decreases by incorporating rubber from vehicular transport tires in the paving block mixture.

The relationship of different physical and mechanical properties of concrete paving blocks was investigated by [10], where it is described in terms of physical variables—water absorption, porosity, and specific gravity—and mechanical variables—compressive strength and tensile splitting strength. The regressions of the variables were carried out, where it was found that the most robust correlation coefficients are (a) correlation between the water absorption and the tensile splitting strength, (b) the compressive strength and the water absorption, (c) porosity and compressive strength, (d) porosity and tensile splitting strength. Additionally, it is indicated that water absorption is a physical property that can be easily determined, has a high correlation with performance parameters, and can also be used as a rapid quality control parameter.

The research led by [11] used 200 × 100 × 100 paving blocks manufactured by vibro-compaction; the objective was to determine the changes in density using the DIN12390-7 standard, absorption, freezing–thawing resistance, and tensile splitting strength of paving blocks on the pallet. It is indicated that the changes at the ends of the tray were attributable to the uneven distribution of the compaction and filling of the mixture in the vibro-compacting machine, making the density higher in the center and lower at the ends. Likewise, when the density is lower at the ends, the product's resistance is lower. In the investigation carried out by [12], several concrete mixtures with different densities and water/cement ratios were elaborated to relate them with the resistance at 28 days of age. The increase in the density of the concrete increased the compressive strength with an exponential behavior. The study described by [13] incorporates tea waste ash with different proportions to replace cement, giving us a lower density and compression with a higher proportion of ash. It can be seen in the graphs that the first two mixtures with the minor cement replacement give us the highest compressive strength, and these same samples are the ones with the highest density. In Indonesia, the study carried out by [14] uses different ratios of $NaOH/Na_2SiO_3$ and fly ash as a substitute for cement to predict the performance of paving blocks; higher Na_2SiO_3 content results in lower percentage absorption and higher endurance.

Ref. [15] details that two groups were compared; the first group incorporates only sand, and the second gravel (stone). It was indicated that a high-quality paving block has a high compressive strength and low water absorption percentage. The results show that the resistance of the paving block only with sand gives greater resistance. The elaboration of different mixtures incorporating wheat straw fibers with and without treatment with sodium silicate was studied by [16], where the untreated mixtures increased the percentage of water absorption and decreased the compressive strength and tensile splitting strength. In this study, it can be seen from the graphs that the mixture with the lowest compressive strength and tensile splitting strength is the one with the highest percentage of water absorption. The authors of [17] prepared concrete mixes with different water/cement ratios in cubes, subjecting them to different curing methods. The authors indicated no clear relationship between compressive strength and water absorption.

Simple linear regression is a method that is basic and easy to calculate. However, it can only be accurate if it does not violate the assumption of linearity. Multivariate methods and machine learning are revolutionizing data analysis in scientific research. These powerful tools allow us to explore complex relationships between multiple variables, discover hidden patterns in large datasets, and build superior predictive models. In the study carried out by [18], 40 types of methods for data analysis are summarized and discussed with an approach to pavement engineering for the prediction and classification of variables. They delved into the data analysis, explaining in detail the definition of each one, making the regression models understandable and easy to interpret as linear and nonlinear equations, logistic regression, survival analysis, and stochastic processes, giving the coefficients of regression a quantitative meaning. Supervised machine learning models, such as artificial neural networks, decision trees, support vector machine, and k-nearest neighbors, give the ability to predict and classify large volumes of data. The investigation carried out by [19] to predict the tensile splitting strength in concrete indicates that the compressive strength and tensile strength are essential characterization indices of the concrete, indicating that generally, the compressive strength is much lower than the tensile strength; the study proposes an alternative method to predict the tensile splitting strength by compressive strength using a novel method called GEP, which is a gene expression programming technique based on constantly adapting tree structures.

Different regression methods were studied by [20], such as neural networks and gene expression programming, to predict splitting tensile strength and water absorption using predictor variables such as the amount of cement, amount of ZnO_2 nanoparticles, type of aggregate, content of water, amount of superplasticizer, age and cured type, and number of test attempts with various types of concrete including ZnO_2 nanoparticles. The training

and validation data are separated to test two models with different design parameters. The determination coefficients for the relationship between predicted values and values of the validation set were above 0.9. Ref. [21] applied Random Forest, Support Vector Regression, and XGBoost to predict the resistance of high-performance concrete, finding XGBoost to be the most accurate. The artificial neural network, decision trees, and random forest methods for predicting tensile splitting strength were described in the study by [22], where concrete is used as a recycled aggregate, obtaining a database divided into training and validation. The input variables were the amount of water, cement, superplasticizer, fine aggregate, coarse aggregate, residual coarse aggregate, density, and water absorption. The random forest method showed a higher coefficient of determination when relating the predicted values and the validation base; additionally, it gave lower error values than the other methods.

Multidimensional methods provide a complete understanding of complex phenomena that may go unnoticed in simple models. Nonlinear behaviors can be addressed through machine learning techniques; however, this entails computational costs and requires software for analysis, as well as the limited capacity to graphically represent multiple dimensions in space, since the physical environment is abstracted into three dimensions.

2. Materials and Methods

The paving blocks of the present investigation were obtained from a factory that produces 30,000 units per day in Quito-Ecuador; the paving block model is rectangular with nominal dimensions of 200 mm in length, 100 mm in width, and 60 mm in thickness. The research has a quantitative approach, since all the variables to be analyzed are continuous quantitative.

As a first point, the sample size is estimated using the G* Power software for multiple linear regression of 5 predictors, using a medium range effect size of 0.0363, which allowed us to estimate a sample size of 300 paving stones, with which an estimated power of the test of 0.9502764 was obtained. On production day, the paving blocks for the population of 30,000 units are sampled as the pieces come out of the vibro-compacting machine, the measurements of the length, width, thickness, and mass of the fresh paving block are taken, the paving blocks are marked to guarantee traceability, and the absorption and resistance tests are carried out according to the [1] The database consists of rows representing the analysis individuals of the sampled and numbered paving blocks; the columns represent the analysis variables.

The variables that enter the models of multivariate techniques, also called input, predictor, or independent, are length in mm, width in mm, height in mm, the mass of the fresh paving blocks (each piece fresh from the vibro-compacting process) expressed in grams, and the percentage of absorption based on the NTE INEN 3040 standard. Additionally, since the dimensional variables of the paving blocks and the mass of the fresh product can be reformulated into a new variable called the density of the fresh product expressed in units of (kg/m^3), this was taken into account together with the percentage of water absorption to make the prediction and compare the multivariate methods.

The response variable, also called dependent or output, is the indirect tensile strength in megapascals (MPa), which will depend on the predictor variables, which are grouped into the first group of predictor variables (5): length, width, thickness, mass of the fresh product, and water absorption percentage. The second group of predictor variables were (2) density of the fresh paving block and percentage of water absorption.

Water absorption. The water absorption test is carried out by reference to the NTE INEN 3040 (2016) standard, which indicates that to determine the absorption rate, the paving block must be submerged in potable water at 20 $\pm$ 5 °C minimum for three days, and then the surface must be cleaned of excess water with a moistened cloth, and the moistened paving block must be weighed to a constant mass. In the same way, an oven is used to find the mass of the dry paving block, and it is placed for a minimum period of 3 days at 105 $\pm$ 5 °C until constant mass. The calculation is made by the difference

between the saturated and dry mass divided by the dry mass. This would correspond to the percentage of the maximum mass of water that the paving block has absorbed from the dry state to the saturated state. The calculation formula is as follows:

$$W_a = \frac{M_1 - M_2}{M_2} \quad (1)$$

M_1 is the mass of the specimen saturated with water, expressed in grams.
M_2 is the final mass of the dry specimen, expressed in grams.

Fresh paving block weight (mass). It is the mass of the paving block expressed in grams taken just after it leaves the mold in the vibro-compaction process.

Thickness (height), length, and width of the paving block. The height of the paving block, or thickness, is the distance between the lower and upper face. In practice, the height variation is given by the vertical mobility of the mechanical parts in the vibro-compactor, where the plate of tamping compresses the mixture into the mold. The width and length of the paving blocks are taken from one end to the other.

Tensile splitting strength. The tensile splitting strength test is carried out using a hydraulic press, giving the measured load at failure in newtons, and the strength is calculated by applying the following formula indicated in the standard NTE INEN 3040 (2016).

$$T = 0.637 \times k \times \frac{P}{S} \quad (2)$$

where T is the paving block strength in MPa, P is the measured load at failure in newtons, S is the area of failure plane in mm^2 that results from the multiplication of the measured failure length and the thickness at the failure plane of the paving block, and k = 0.87 for a thickness of 60 mm.

Mahalanobis Distance. The importance of identifying outlier data in a database is that these can distort the statistical analysis, and therefore, the distance to the centroid and the shape are taken into account; the Mahalanobis distance takes these two premises into account [23]. The study by [24] indicates that in the multivariate field with Gaussian data, the Mahalanobis distance follows a chi-square distribution, where p means degrees of freedom and represents the number of variables. The Mahalanobis distance measures the amount of the standard deviation of an observation or individual from the mean of a distribution, considering correlations for multivariate analysis. The Mahalanobis distance transforms to a Euclidean distance when the covariance matrix is the identity matrix [25]. A multivariate normal distribution is defined as follows:

$$f(X) = \left(\frac{1}{2\pi}\right)^{p/2} * |\Sigma|^{-1/2} * exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}, \quad (3)$$

where Σ is the covariance matrix, and μ is the mean vector. If X is a vector with p variables which follows a multivariate normal distribution $X \sim N_p(\mu, \Sigma)$, then the Mahalanobis distance square D^2 follows a chi-squared distribution with p degrees of freedom $D^2 \sim X_p^2$. Mahalanobis represents the distance between each data point and its center of mass and is defined by the following formula:

$$D^2 = (X-\mu)^T \Sigma^{-1} (X-\mu) \quad (4)$$

Simple linear regression. Simple linear regression allows one to relate two variables: variable Y, called response or dependent, and variable X, predictor or explanatory. The regression of the two random variables is given by the expected value of Y when X takes a specific value ($X = x$). If we consider linear regression with intercept β_0, slope β_1, and e_i

represents the random error of Y_i, Ref. [26] explains that the residuals $\hat{e}_i$ are $y_i - \hat{y}_i$, where $\hat{y}_i$ is the fitted value of y.

$$Y_i = E(Y|X = x) + e_i = \beta_0 + \beta_1 x + e_i \tag{5}$$

The popular method for obtaining β_1 and β_0 is *RSS* ordinary least squares to minimize the difference between the observed and predicted values. The minimization is carried out by differentiating RSS concerning the coefficients b_0 and b_1 and setting it equal to 0.

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - b_0 + b_1 x_i)^2 \tag{6}$$

The structural assumptions of the regression model are the linearity that explains that Y depends on x through linear regression, and the homoscedasticity indicates that the variance of the errors when $X = x$ must be common, better explained as $Var(e|X = x) = \sigma^2$; the normality assumption indicates that the errors must follow a normal distribution with 0 mean and variance σ^2, and finally the independence of the errors. The inference of the linear regression under the previous assumptions $\hat{\beta}_1$ follows a normal distribution with mean β_1 and variance $\left(\sigma^2/SXX\right)$ where $SXX = \sum_{i=1}^{n} (x_i - \overline{x})^2$; if we consider to be σ^2 unknown, then the test statistic follows a distribution of T-student with $n - 2$ degrees of freedom, where H_o: $\beta_1 = 0$ is the null hypothesis and H_a: $\beta_1 \neq 0$ is the alternative hypothesis. T is described by the following expression:

$$T = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SXX}} \sim t_{n-2}, \tag{7}$$

Similarly, for β_0, the T statistic follows a T-student distribution with n-2 degrees of freedom, where the null hypothesis is H_o: $\beta_0 = 0$ and the alternative hypothesis is H_a: $\beta_0 \neq 0$.

$$T = \frac{\hat{\beta}_0 - \beta_0}{s/\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SXX}}} \sim t_{n-2}, \tag{8}$$

The analysis of variance allows us to decompose the variability by analyzing the mean of Y, the predicted and observed points; the total variability is separated into the sum of the variability explained by the model plus the unexplained variability or error. The total sum of squares is $SST = SYY = \sum_i^n (y_i - \overline{y})^2$ and $SST = SSreg + RSS$, where $SSreg$ is the sum of squares of the regression $(SSreg) = \sum_i^n (\hat{y}_i - \overline{y})^2$ and RSS is the sum of squares of the residuals $RSS = \sum_i^n (y_i - \hat{y}_i)^2$.

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}) \tag{9}$$

$$SST = SSreg + RSS \tag{10}$$

The F statistic follows an F distribution with 1 and n-2 degrees of freedom, where the null hypothesis is H_o: $\beta_1 = 0$ and the alternative hypothesis is H_a : $\beta_1 \neq 0$. It can be seen that if the null hypothesis is rejected, then Y depends on X.

$$F = \frac{SSreg/1}{\frac{RSS}{(n-2)}} \sim F_{1,\, n-2}, \tag{11}$$

The coefficient of determination in linear regression is given by the following:

$$R^2 = \frac{SSreg}{SST} \tag{12}$$

Multiple Linear Regression (MLR). According to [26], the response variable (Y) in MLR is predicted and related to multiple explanatory or predictor variables, where the expectation of Y when each variable X takes a specific value is represented as follows:

$$E(Y|X_1 = x_1, X_2 = x_2, .., X_p = x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \tag{13}$$

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + e_i \tag{14}$$

The sum of squares for the multivariate case is

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 + b_1 x_{1i} + \cdots + b_p x_{pi})^2 \tag{15}$$

Multiple linear regression is denoted as

$$Y = X\,\beta + e \tag{16}$$

Each term expressed in vectors and matrices indicates the following:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{pmatrix}$$

The estimated coefficients of each term can be calculated by linear algebra calculus, where the term $X\left(X^T X\right)^{-1} X^T$ is defined as a Hat Matrix (H) and the residual maker matrix as M, which is equal to $I_n - H$, where I_n is the identity matrix, and the projection $\hat{Y}$ is equal to $H * Y$, so in the regression hyperplane, $\hat{Y}$ is a transformation of Y.

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{17}$$

If the errors follow normal distribution with constant variance, then the T statistic follows a student's t distribution with n-p-1 degrees of freedom and is given by

$$T_i = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \sim t_{n-p-1} \tag{18}$$

The term $se(\hat{\beta})$ is the estimated standard deviation of $\hat{\beta}_i$, where the null hypothesis indicates that H_o: $\beta_i = 0$ and the alternative hypothesis is H_a: $\beta_i \neq 0$. In the analysis of variance in the multivariate case, as in the case of simple linear regression, the total variability is equal to the variability explained by the model plus the unexplained variability or error. The F statistic follows an F distribution with p and n-p-1 degrees of freedom, where the null hypothesis is Ho: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$ and the alternative hypothesis indicates that H_a: *at least some of the* $\beta_i \neq 0$.

$$F = \frac{SSreg/p}{\frac{RSS}{(n-p-1)}} \sim F_{p,\, n-p-1} \tag{19}$$

Adding the number of predictors increases R^2, so R^2_{adj} is used.

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-p)} \tag{20}$$

Regression trees. A decision tree is an algorithm in machine learning that can be used in regression and classification; that is, a white box where they are intuitive and easy to interpret. For the regression case, the tree, instead of predicting a class, predicts

a value that is the average value across the training instances of the node. Instead of minimizing impurity, the regression tree minimizes the mean squared error (MSE) [27]. The cost function for regression is

$$J(k, t_k) = \frac{m_{left}}{m} MSE_{left} + \frac{m_{right}}{m} MSE_{right} \quad (21)$$

where m is the number of instances to the left or right, and MSE is the mean square error.

$$MSE = \sum_{i \epsilon\ node} \left(\hat{y}_{node} - y^{(i)} \right)^2 \quad (22)$$

$$\hat{y}_{node} = \frac{1}{m_{node}} \sum_{i \epsilon\ node} y^{(i)} \quad (23)$$

The limits on the decision trees are perpendicular to the axis (orthogonal) and are sensitive to variations in training. The regression trees, according to [28], generate divisions of the database into more homogeneous groups, with a set of "if" and "then" conditions being easily interpretable with different predictors. A disadvantage that could occur is instability with minor changes in the data. The oldest and most widely used technique is CART by [29], whose methodology is to find the predictor and the dividing value of the base whose squared error is the smallest.

$$SSE = \sum_{i \epsilon S1} (y_i - \overline{y}_1)^2 + \sum_{i \epsilon S2} (y_i - \overline{y}_2)^2 \quad (24)$$

$\overline{y}_1$ represents the subgroup mean S_1, $\overline{y}_2$ represents the subgroup mean S_2, and the division process continues within the sets S_1 and S_2 up to a stopping criterion. The predictor's relative importance can be calculated using SSE, where the predictors higher up the tree or that are more frequent are the most important.

Random forests. In the research by [30], an algorithm called the random forest allows for predicting and reducing overfitting. The procedure consists of choosing the number of tree models to be built from 1 to m, obtaining an initial sample, and then training the tree model for each division; the predictors are randomly selected, the best one is chosen, and finally, the stopping criteria are used. Each tree model generates a prediction, and the m predictions are averaged to generate the final prediction. By randomly choosing the k variables in each division, their correlation decreases. Random forests are computationally more efficient tree by tree, and the predictors' importance can also be seen through the permutation or impurity methodology [28]. The tree bagging procedure reduces the prediction variance. The ensemble method is the algorithm that analyzes the predictions as a whole, obtaining the predictions of each individual tree with different random subsets [27]. Analyzing the predictions together will yield better results than just one prediction. Random forests are trained by bagging with max_samples.

Principal component analysis. It is a dimension reduction technique that occupies the orthogonal transformation so that a group of correlated n-dimensional variables can maintain their variability information in other uncorrelated k-dimensional ones. The general process consists first of data standardization so that the base has a mean of zero and a variance of one. The covariance matrix, correlation matrix, eigenvectors, and eigenvalues are calculated. The first eigenvectors representing the most significant variability are chosen [31]. Research carried out by [32] indicates that this technique was developed by Karl Pearson and Harold Hotelling independently. The technique linearly transforms multivariate data into a new uncorrelated set of variables. The eigenvectors are vectors that do not change position when a data transformation occurs and represent the axis of maximum variance, called the principal component.

According [33], principal components are commonly defined as the matrix multiplication between the eigenvectors of the correlation matrix (A) and the standardized variables (X^*).

$$z = A^T X^* \quad (25)$$

The principal components calculated by covariance have a drawback, and it is the sensitivity to the units of measurement, for which it is carried out using the correlation matrix with the standardized variables, since each variable has a different unit of measurement. Also, the sizes of the variances of the principal components have the same implications in correlation matrices as in covariance matrices. One of the properties of the principal components using the correlation matrix is that they do not depend on the absolute values of the correlation.

According to the study in [34], principal component analysis can determine the number of hidden layers in artificial neural networks, which represents sufficient variability for statistical analysis. On the other hand, the study by [35] shows the number of hidden layers in neural networks through principal component analysis to predict a continuous variable, giving good results through quality measures with optimal performance.

Artificial neural networks. Neural networks were inspired by the biological capacity of the brain. The perceptron is a different neuron called a threshold logic unit, where the inputs are numbers just like the outputs, and each connection has a weight [27]. A fully connected layer has the outputs $h_{W,b}(X)$, where X is the input matrix (instance rows and feature columns), W is the weight matrix (rows per input neurons and columns per artificial neuron), b is the polarization vector (connection weights of the bias neuron and the artificial neurons), and is the activation function.

$$h_{W,b}(X) = (X\,W + b) \tag{26}$$

Learning has a rule; the perceptron connections are strengthened when the error is reduced, receiving one instance at a time and making the predictions. Perceptron learning is performed by ${w_{i,j}}^{(next\ step)}$:

$${w_{i,j}}^{(next\ step)} = w_{i,j} + \eta\ \left(y_j - \hat{y}_j\right)x_i \tag{27}$$

where $w_{i,j}$ is the weight of the connection between the input and output of the neurons, x_i is the input value of the instance, $\hat{y}_j$ is the output value of the instance, y_j is the target output value, and η is the learning rate. The perceptron convergence theorem tells us that the algorithm converges to a solution if the instances are linearly separable. The multilayer perceptron, MLP, has an input layer (lower layer), hidden layers, and an output layer (upper layer). If the artificial neural network (ANN) has more than one hidden layer, it is called a deep neural network (DNN). An algorithm called gradient descent was created to calculate the gradients automatically with one pass forward and another pass backwards; the process is repeated until converging to the solution. According to [36], the sigmoid neuron has weights and a bias occupying the sigmoid function defined as $\sigma(z) = 1/(1 + e^{-z})$, where $z = wx + b$ and the bias $b = 1/\left(1 + e^{-\sum_j w_j x_j - b}\right)$ is the introduced bias. Considering the cost function to evaluate the model and quantify how well the objective is achieved, we have the following:

$$C(w, b) = \frac{1}{2n}\sum_x \|y(x) - a\|^2, \tag{28}$$

where w represents the weights of the network, b are all the biases, n is the total training inputs, a is the vector of outputs when inputting an x, $y(x)$ is the output desired, x sums over the training inputs, and C is the cost function. As the cost function approaches 0 and as $y(x)$ approaches the output a, gradient descent allows minimization of the cost function, where ΔC can be written as $\Delta C \approx \nabla C\ \Delta v$, where ∇C is the gradient vector and relates the changes of C to changing v, so Δv is the vector of changes in position, and m is the number of variables.

$$\nabla C \equiv \left(\frac{\partial C}{\partial v_1}, \ldots, \frac{\partial C}{\partial v_m}\right)^T \tag{29}$$

Gradient descent repeatedly computes ∇C, looking like small steps in the direction C decreases the most. The backpropagation algorithm gives information on how to change

the weights and biases in the behavior of the neural network. The notation to use is l for the l-th layer, k is the k-th neuron of the l-th layer minus one $(l-1)$, j is the j-th neuron of the l-th layer, w_{jk}^{l} is the weight of the connection of the l-th layer for the j-th neuron and k-th neuron of the layer $(l-1)$, b_{j}^{l} is the bias of the j-th neuron for the l-th layer, and a_{j}^{l} is the activation of the l-th layer for the j-th neuron, where $a_{j}^{l}=\sigma\left(\sum_{k} w_{jk}^{l}\, a_{k}^{l-1}+b_{j}^{l}\right)$ and $z_{j}^{l}=\sum_{k} w_{jk}^{l}\, a_{k}^{l-1}+b_{j}^{l}$; for matrix and vector notation it is $a^{l}=\sigma\left(w^{l}\, a^{l-1}+b^{l}\right)=\sigma\left(z^{l}\right)$, $z^{l}=w^{l}\, a^{l-1}+b^{l}$, and for the cost function $C=\frac{1}{2n}\sum_{x}\left\|y(x)-a^{L}(x)\right\|^{2}$. The desired output is expressed as $y(x)$, n is the number of training instances or examples, and $a^{L}=a^{L}(x)$ is the output vector of activations when x is entered. The Hadamard product uses $\odot$ for denoting the multiplication of the elements of two vectors; the error in the j-th neuron, in the output layer, is

$$\delta_{j}^{L}=\frac{\partial C}{\partial a_{j}^{L}}\sigma'(z_{j}^{L}), \tag{30}$$

A weight will learn slowly if the output neuron is saturated or if the input neuron has low activation. The rate of change of cost concerning bias is $\partial C/\partial b_{j}^{l}$, and the rate of change of cost concerning weight is $\partial C/\partial w_{jk}^{l}$. The summary backpropagation equations are $\delta^{L}=\nabla_{a}C\odot\sigma'\left(z^{L}\right)$, $\delta^{l}=\left(\left(w^{l+1}\right)^{T}\,\delta^{l+1}\right)\odot\sigma'\left(z^{l}\right)$, $\partial C/\partial b_{j}^{l}=\delta_{j}^{l}$, and $\partial C/\partial w_{jk}^{l}=a_{k}^{l-1}\;\delta_{j}^{l}$. The backpropagation algorithm consists first in establishing the corresponding activation in the input layer, second we calculate $z^{l}=w^{l}a^{l-1}+b^{l}$ and $a^{l}=\sigma(z_{l})$, third we calculate the output error by calculating the vector $\delta^{L}=\nabla_{a}C\;\odot\sigma'\left(z^{L}\right)$, fourth we calculate the backpropagation error $\delta^{l}=\left(\left(w^{l+1}\right)^{T}\,\delta^{l+1}\right)\odot\sigma'\left(z^{l}\right)$, and lastly we calculate the gradient of the cost function. It will be for the weights $\partial C/\partial w_{jk}^{l}=a_{k}^{l-1}\;\delta_{j}^{l}$, and the bias is $\partial C/\partial b_{j}^{l}=\delta_{j}^{l}$. An intuitive way to see the rate of change of C concerning the weights in the network is

$$\frac{\partial C}{\partial w_{jk}^{l}}=\sum\nolimits_{mnp\ldots q}\frac{\partial C}{\partial a_{m}^{L}}\,\frac{\partial\, a_{m}^{L}}{\partial\, a_{n}^{L-1}}\,\frac{\partial\, a_{n}^{L-1}}{\partial a_{p}^{L-2}}\cdots\frac{\partial a_{q}^{l+1}}{\partial a_{j}^{l}}\,\frac{\partial a_{j}^{l}}{\partial w_{jk}^{l}} \tag{31}$$

Overfitting. Overfitting is a substantial problem in statistical modeling that can compromise the integrity of the reported findings. Ref. [27] indicates that it occurs when the model works very well in training but performs poorly with new instances and does not allow generalization. If the training set is noisy or small, it is possible for the models to detect patterns of the noise itself, since very complex models detect negative patterns.

The ways to avoid overfitting are to simplify the model, regularize, and stop training at an optimal point.

3. Results

The database is obtained from measurements on 300 sample paving blocks; the tensile splitting strength is the response variable, and the models consider two groups of predictor variables, the first group of five variables (length, width, thickness, mass of the fresh paving block, and water absorption percentage) and the second group of two variables (density of the fresh paving block and percentage of absorption). As the first point in the data processing, an analysis of missing data is carried out, resulting in the database being complete; therefore, it does not require any imputation method. Outliers are determined by Mahalanobis distances, which represent the distance between each data point and its center of mass. For the first group of predictor variables, the square of the distance is calculated following a chi-square distribution with six degrees of freedom. The data belonging to the area under the curve of 99.9% of the distribution are preserved, evidencing seven records outside that represent 2.3% of the total database, which must be excluded for the analysis, resulting in a database of analysis that is made up of 293 records that enter the multivariate

regression analysis of the first group of predictors. Similarly, for the multivariate regression analysis for the second group of predictors with three degrees of freedom, four records outside 99.9% of the distribution were detected, which are excluded and represent 1.3% of the total database, so the analysis database contains 296 records for the second group of predictor variables. A total of 80% of the database was separated for training records and 20% for validation of the models.

The Table 1 shows the descriptive statistics in the sample of each variable, the measures of central tendency such as the mean and median, and the measures of dispersion such as the standard deviation, the range, and the coefficient of variation. Additionally, a histogram is included, where the values of the variable are grouped into intervals, and each rectangular bar indicates the frequency of the data in each interval, the red line indicates the density function for the behavior of the data in the sample, and at the bottom is the boxplot diagram and the points that indicate the data with its dispersion.

Table 1. Descriptive statistics of the variables.

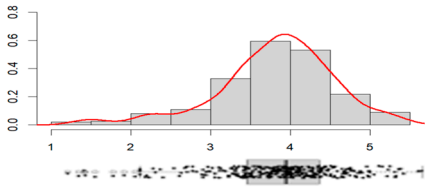
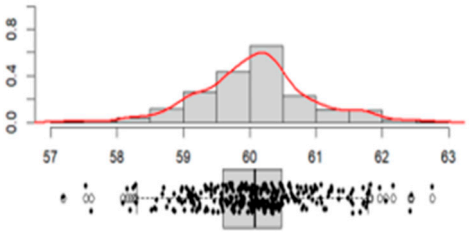
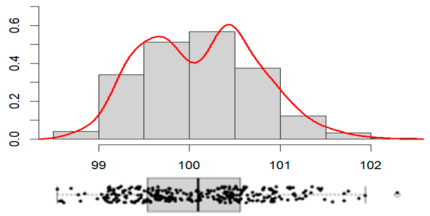
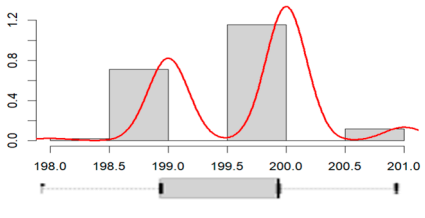
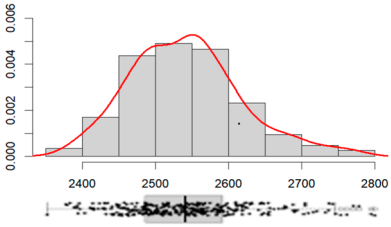
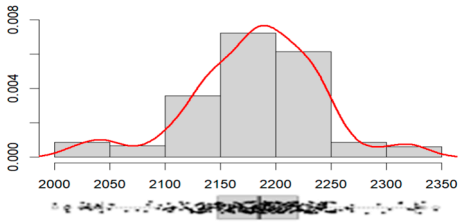
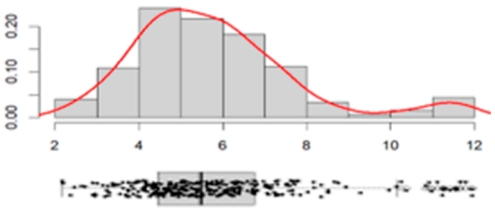
Variable	Measures			Histogram, Density Function, Boxplot (Below), and Dots of Data
Tensile splitting strength, MPa	Max:_____	5.47	MPa	
	Min:_____	1.31	MPa	
	Mean:_____	3.82	MPa	
	Median:_____	3.89	MPa	
	Standard deviation:_____	0.74	MPa	
	Variance:_____	0.55	MPa^2	
	Range:_____	4.16	MPa	
	Coefficient of variation:____	19.42	%	
Thickness, mm	Max:_____	62.75	mm	
	Min:_____	57.19	mm	
	Mean: _____	60.08	mm	
	Median: _____	60.08	mm	
	Standard deviation:_____	0.85	mm	
	Variance:_____	0.72	mm^2	
	Range:_____	5.56	mm	
	Coefficient of variation:____	1.41	%	
Width, mm	Max:_____	102.03	mm	
	Min:_____	98.77	mm	
	Mean:_____	100.13	mm	
	Median: _____	100.14	mm	
	Standard deviation:_____	0.63	mm	
	Variance:_____	0.39	mm^2	
	Range: _____	3.26	mm	
	Coefficient of variation:____	0.63	%	
Length, mm	Max:_____	201.00	mm	
	Min:_____	198.00	mm	
	Mean:_____	199.68	mm	
	Median:_____	200.00	mm	
	Standard deviation:_____	0.60	mm	
	Variance:_____	0.35	mm^2	
	Range:_____	3.00	mm	
	Coefficient of variation:____	0.30	%	

Table 1. *Cont.*

Variable	Measures			Histogram, Density Function, Boxplot (Below), and Dots of Data
Mass of fresh paving block, g	Max:_____	2780.30	g	
	Min:_____	2363.70	g	
	Mean:_____	2541.55	g	
	Median:_____	2539.60	g	
	Standard deviation:_____	77.19	g	
	Variance:_____	5958.60	g^2	
	Range:_____	416.60	g	
	Coefficient of variation:_____	3.04	%	
Density of fresh paving block, kg/m^3	Max:_____	2385.27	kg/m^3	
	Min:_____	1993.58	kg/m^3	
	Mean:_____	2181.57	kg/m^3	
	Median:_____	2186.00	kg/m^3	
	Standard deviation:_____	61.53	kg/m^3	
	Variance:_____	3785.66	$(kg/m^3)^2$	
	Range: _____	391.69	kg/m^3	
	Coefficient of variation:____	2.82	%	
Percentage of water absorption, g/g%	Max:_____	13.64	g/g%	
	Min:_____	2.12	g/g%	
	Mean: _____	5.81	g/g%	
	Median:_____	5.46	g/g%	
	Standard deviation: _____	2.05	g/g%	
	Variance:_____	4.18	$(g/g\%)^2$	
	Range:_____	11.52	g/g%	
	Coefficient of variation: ____	35.20	%	

The Figure 1 shows us the dot plot of all the variables that can be related two by two and the correlation coefficient that measures the intensity of the linear relationship of the two variables, which is positively higher, with values close to 1 (direct relationship) or −1 (inverse relationship); diagonal graphs show the density function that indicates the probability that the variable takes the values in a specific interval. It is observed that the response variable tensile splitting strength has a high linear correlation with the density of the fresh paver (R^2 = 0.834) and the percentage of water absorption (R^2 = 0.843), indicating the linear nature of the data for multivariate models.

The variables are plotted in three dimensions in Figures 2 and 3, where the variable (y) is the tensile splitting strength, which is also represented through color to be able to distinguish the location of the points in the graph; the points in green indicate a high tensile strength, followed by the yellow and red dots. The grouping of the data with high tensile splitting strength in a specific area and the gradual decrease to the red points initially indicate a linear behavior of the data through a prediction plane.

3.1. Multiple Linear Regression (MLR)

3.1.1. Multiple Linear Regression Model for the First Group of Predictors (Thickness, Width, Length, Mass of Fresh Paving Block, and Percentage of Water Absorption)

The multiple linear regression model for the first group of predictors shows us a non-significant p-value of 0.22736 in the T-test of the coefficient of the width variable. Therefore, the null hypothesis is not rejected $\beta_{i\ (width)} = 0$, and there is no statistical evidence to affirm that $\beta_{i\ (width)} \neq 0$. The discarding of the width variable indicated that it does not significantly influence the prediction of tensile splitting strength. The following model is carried out with the variables (thickness, length, mass of fresh paving block, and the percentage of water absorption), giving rise to significant results (p-value < 0.05); in all the T-tests of the coefficients for a confidence level of 95% the null hypothesis is rejected $\beta_i = 0$, there being significant statistical evidence to affirm that $\beta_i \neq 0$, so the

predictor variables do influence the response variable. The resulting adjusted coefficient of determination is 0.7974. In the *F*-test, the *p*-value is much less than 0.05, so for a confidence level of 95%, the null hypothesis is rejected, there being significant evidence to affirm that *at least some of the* $\beta_i \neq 0$, obtaining a mean square error of 0.110086. The structural assumptions are verified, and Figure 4 illustrates the adjustment to new values for the prediction of tensile splitting strength vs. the current values of the test dataset; the model can be seen below, where the variable with the most significant inverse influence is the thickness followed by the percentage of water absorption.

$$\begin{aligned} \textit{Tensile splitting strength} = \quad & (29.911784) + \\ & (-0.244570)\ \textit{Thickness} + \\ & (-0.108435)\ \textit{Length} + \\ & (0.004428)\ \textit{Mass\_of\_fresh\_paving\_block} + \\ & (-0.174059)\ \textit{Percentage\_of\_water\_absorption} \end{aligned} \tag{32}$$

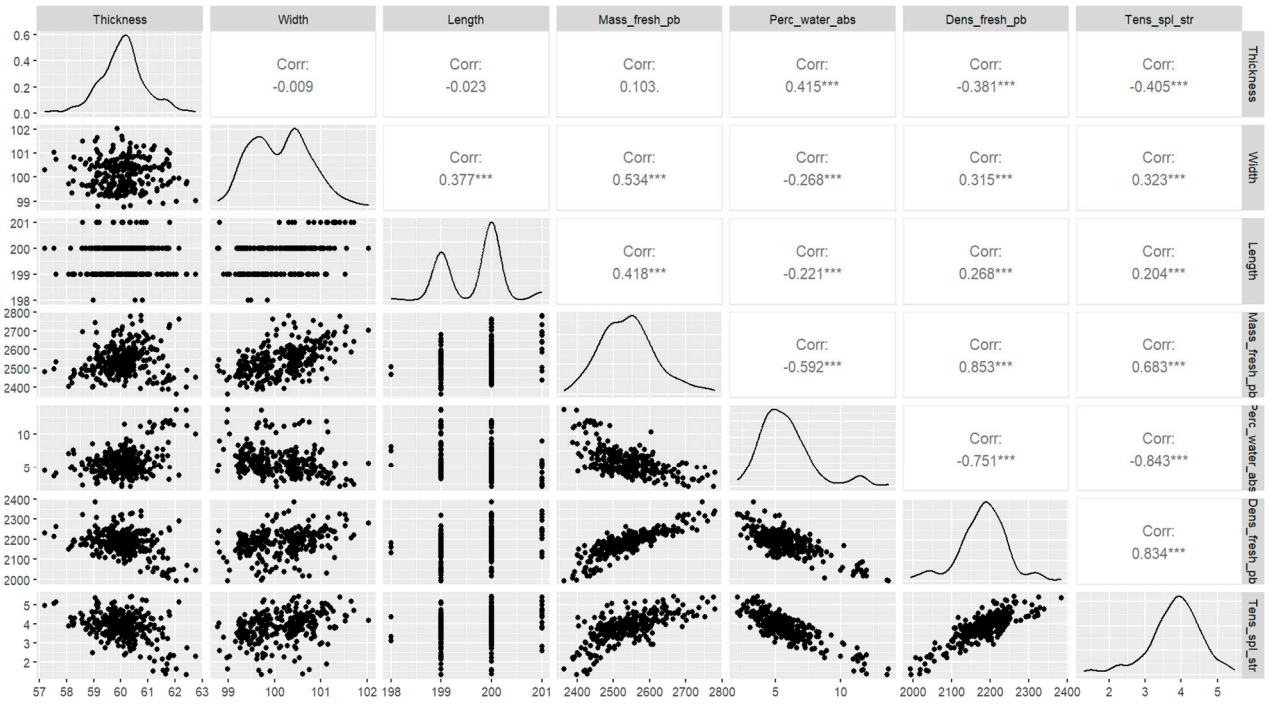


Figure 1. The density function of the variables and correlation coefficient between variables.

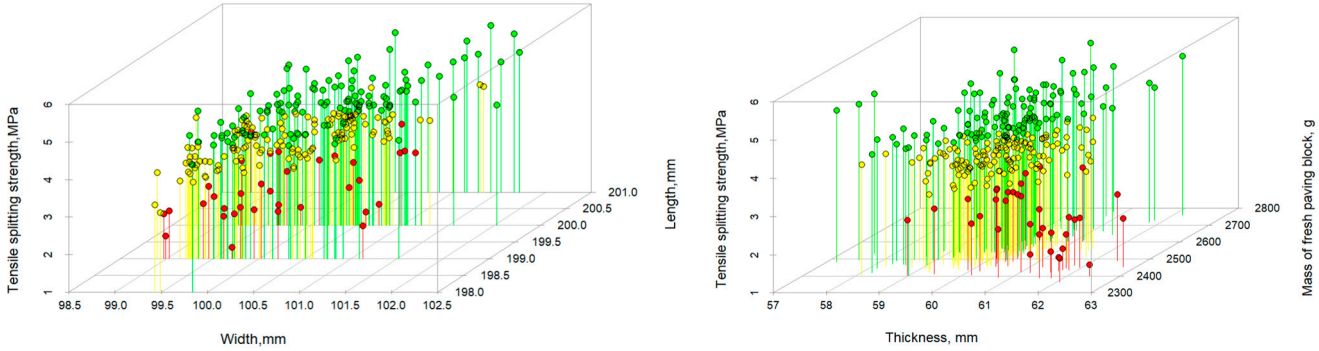


Figure 2. Dot plot of the variables in 3 dimensions. On the left, the tensile splitting strength, length, and width. The tensile splitting strength, thickness, and mass of the fresh paving block are on the right. The green, yellow and red dots represent the high, medium and low tensile splitting strength tensile strength respectively.

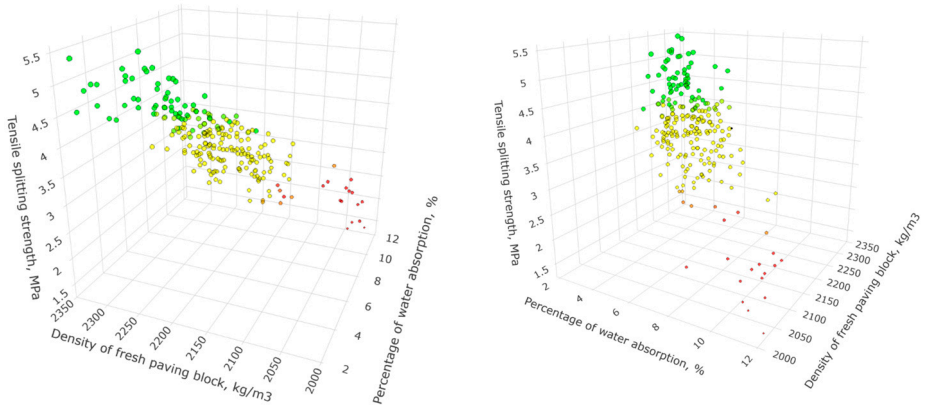


Figure 3. Three-dimensional dot plot for the following variables: tensile splitting strength, density of fresh paving block, and percentage of water absorption at different angles. The green, yellow and red dots represent the high, medium and low tensile splitting strength tensile strength respectively.

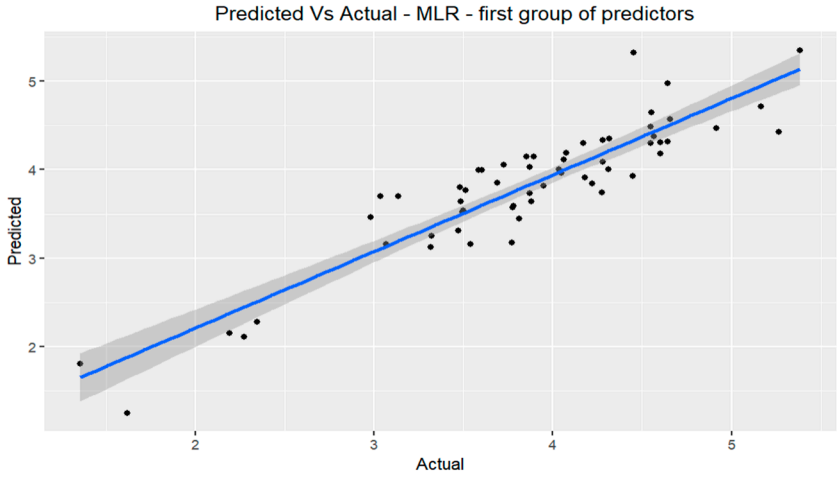


Figure 4. Prediction of the observations vs. their actual value from the evaluation database for the first group of predictors in MLR.

3.1.2. Multiple Linear Regression Model for the Second Group of Predictors (Density of Fresh Paving Block and Percentage of Water Absorption)

When carrying out the model with the predictors of the density of the fresh paving block and the percentage of water absorption, a p-value of less than 0.05 is evidenced for the two variables in the T-test of the coefficients; for a confidence level of 95%, the null hypothesis is rejected, $\beta_i = 0$ exhibiting significant statistical evidence to affirm that $\beta_i \neq 0$. In the F-test, the p-value is less than 0.05, so for a confidence level of 95%, the null hypothesis is rejected, that is, $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$, there being significant evidence to affirm that *at least some of the* $\beta_i \neq 0$; the adjusted coefficient of determination is 0.7897 and results in a mean square error of 0.115044.

The verification of the structural assumptions is carried out by statistical tests on the residuals. Linearity, homoscedasticity, and normality are verified, having a low variance inflation factor. The 3D graph of the tensile splitting strength prediction model is presented using the predictor variables, the density of the fresh paving block, and the percentage of water absorption; the prediction plane can be seen in Figure 5, where the linear behavior is evidenced in three dimensions. Additionally, Figure 6 illustrates the prediction of the tensile splitting strength vs. the actual values of the test dataset, evidencing the adjustment to

new values. The model can be seen as follows, where the variable with the most significant inverse influence is the percentage of water absorption:

$$\begin{aligned} Tensile\ splitting\ strength = \quad & (-7.9314769) + \\ & (\ 0.0058441)\ Density\_of\_fresh\ \_paving\_block + \\ & (-0.1741685)\ Percentage\_of\_water\_absorption \end{aligned} \tag{33}$$

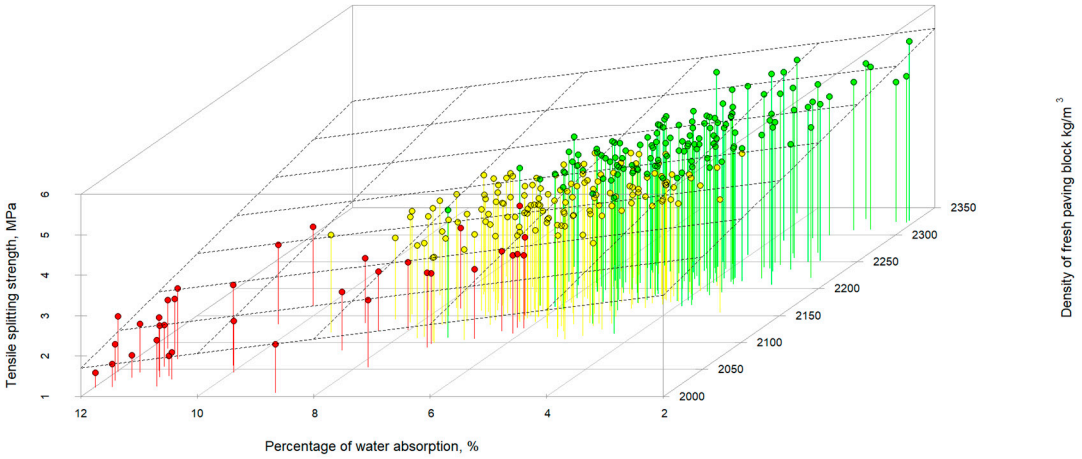


Figure 5. Three-dimensional representation of the plane of the multiple linear regression model for predicting tensile splitting strength with the predictor variables density of the fresh paving block and percentage of water absorption. The green, yellow and red dots represent the high, medium and low tensile splitting strength tensile strength respectively.

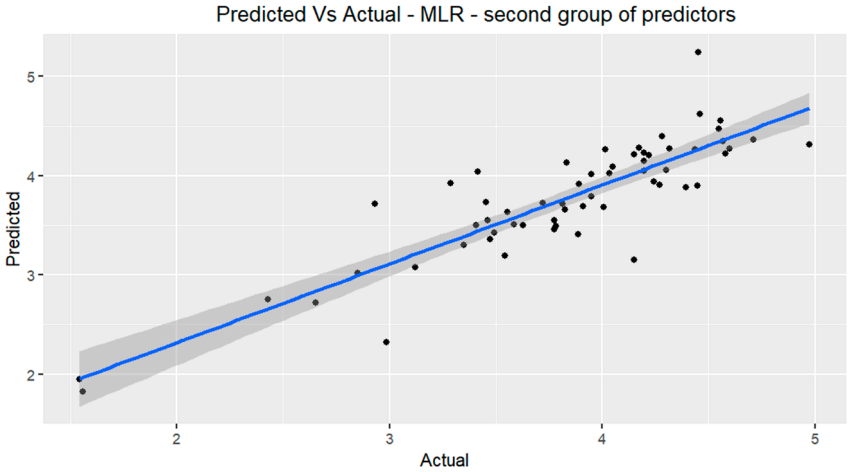


Figure 6. Prediction of the observations vs. their actual value from the evaluation database for the second group of predictors in MLR.

3.2. Regression Trees

3.2.1. Regression Tree Model for the First Group of Predictors (Thickness, Width, Length, Mass of the Fresh Paving Block, and Percentage of Water Absorption)

A regression tree is created to be treated by cross-validation and find the optimal size of terminal nodes to reduce the validation error, resulting in 10 terminal nodes; the resulting mean squared error is 0.165174, and the diagram of Figure 7 shows the tree splits and their model conditions according to the predictor variables, and a route is created for the new values entered into the model, which is practical in the applicability of the prediction. However, the limitation and rigidity in terms of predicted values is evident. For the regression trees of the first group of predictors, Figure 8 shows a low test error.

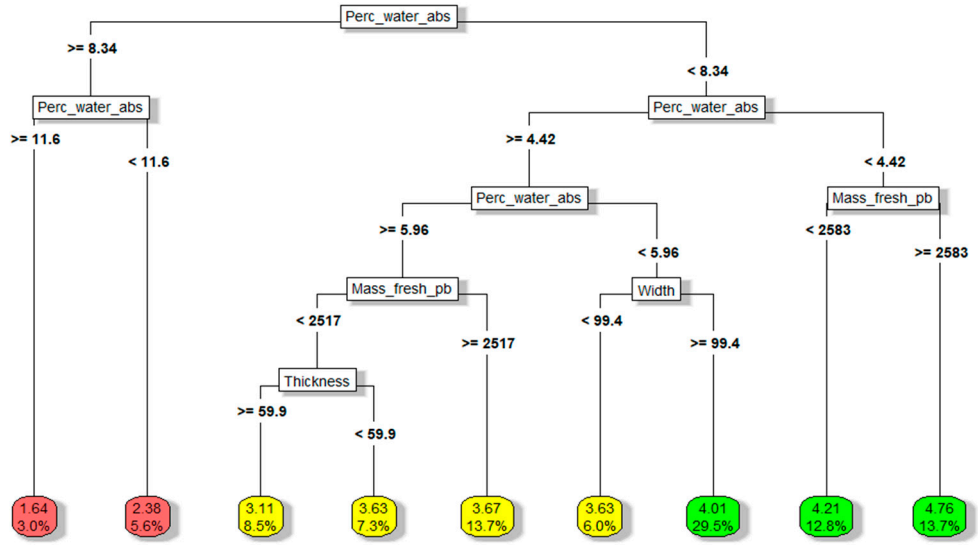


Figure 7. Representation of the regression tree for the first group of predictors (thickness, width, length, mass of the fresh paving block, and percentage of water absorption). The results in green, yellow and red represent high, medium and low tensile strength, respectively.

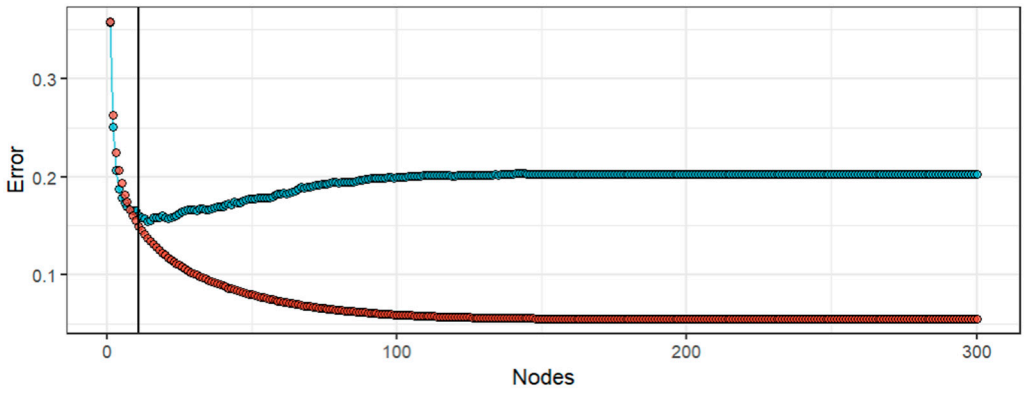


Figure 8. Representation of the test error (blue) and training error (red) for different nodes of the regression tree model in the first group of predictors.

3.2.2. Regression Tree Model for the Second Group of Predictors (Density of Fresh Paving Block and Absorption Percentage)

The regression tree model to predict the tensile splitting strength through the variables density of the fresh paving block and percentage of water absorption is carried out through cross-validation, where it is found that the optimal size to minimize the error is nine terminal nodes; Figure 9 shows the architecture of the regression tree with the divisions according to the conditions of the model, which makes the path to follow intuitive according to the input values of the predictor variables. Figure 10 shows the model's planar projection (regression surface) in three dimensions to predict the tensile splitting strength through the predictor variables, and each step is a prediction value. The higher steps show greater resistance to tensile splitting, corresponding to a higher density and lower percentage of absorption. Figure 11 shows the training of the model to find the optimal number of nodes to identify the stopping point where the overfitting begins, allowing the data to be well-generalized. The resulting mean square error is 0.139050.

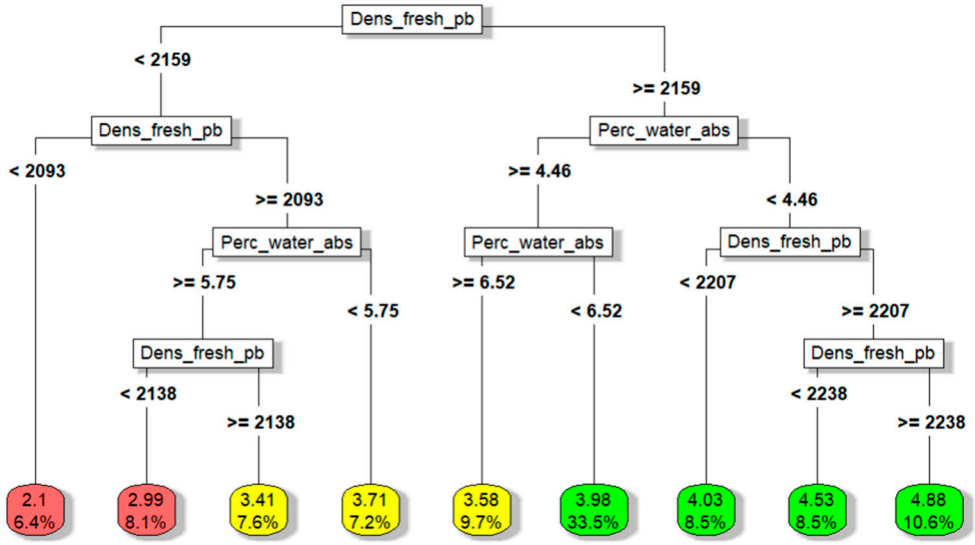


Figure 9. Representation of the regression tree for the second group of predictors (density of the fresh paving block and percentage of water absorption). The results in green, yellow and red represent high, medium and low tensile strength, respectively.

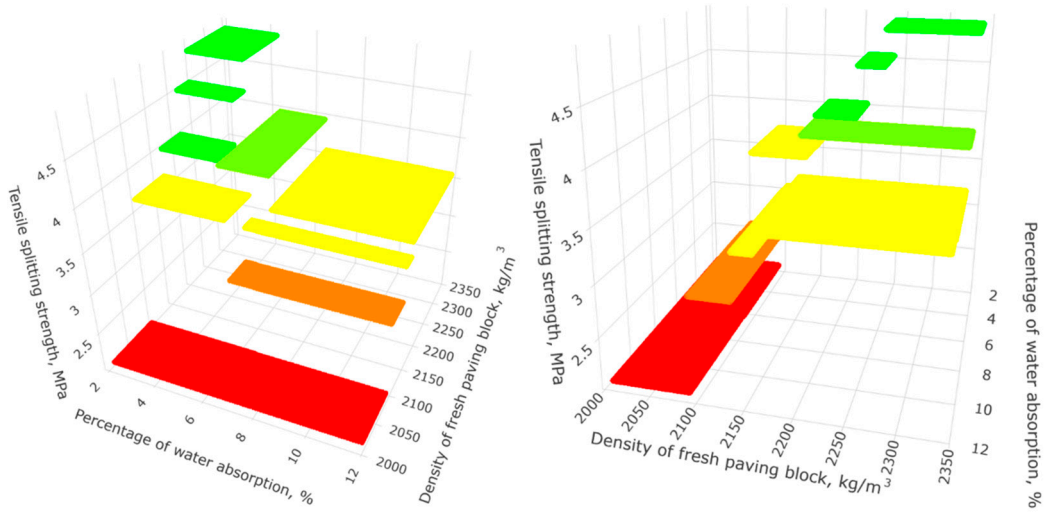


Figure 10. Three-dimensional planar representation of the regression tree model for the second group of predictors (density of the fresh paving block and percentage of water absorption). The green, yellow and red steps represent the high, medium and low indirect tensile strength respectively.

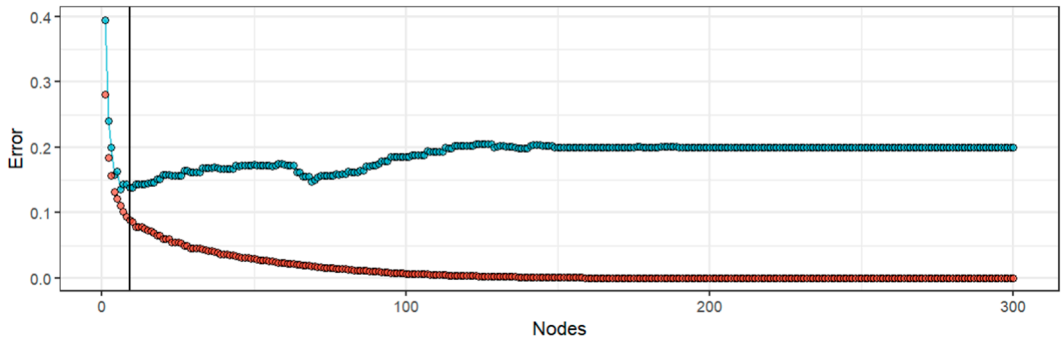


Figure 11. Representation of the test error (blue) and training error (red) for different nodes of the regression tree model in the second group of predictors..

3.3. Random Forest

3.3.1. Random Forest Model for the First Group of Predictors (Thickness, Width, Length, Mass of the Fresh Paving Block, and Percentage of Water Absorption)

The random forest model is created with the predictor variables of the first group (thickness, width, length, mass of the fresh paving block, and percentage of water absorption). The resulting optimal hyperparameters using the cross-validation method are 278 for the number of trees, and the number of predictor variables randomly chosen for each division is 4. The result of the model performance gives a mean square error of 0.115392. The importance of the predictors by permutation can be seen in Figure 12; the impact of each variable on the performance of the model is measured when the values of the variable are randomized, where the most important variable is the percentage of absorption, followed by the mass of the fresh paver, the thickness, width, and length. Figure 13 shows the model error for different numbers of trees in the training dataset and the test dataset, where it is observed that a few trees generate high errors and as the number increases, it stabilizes sufficiently at 278 trees.

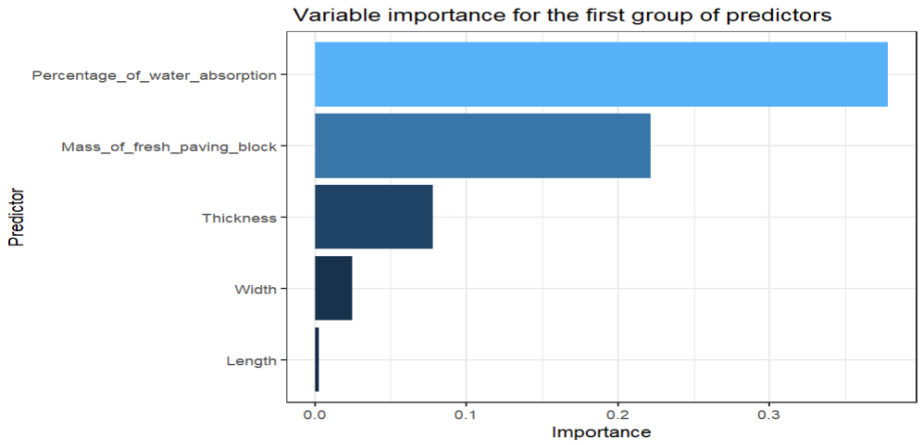


Figure 12. Variable importance for random forest model for the first group of predictors.

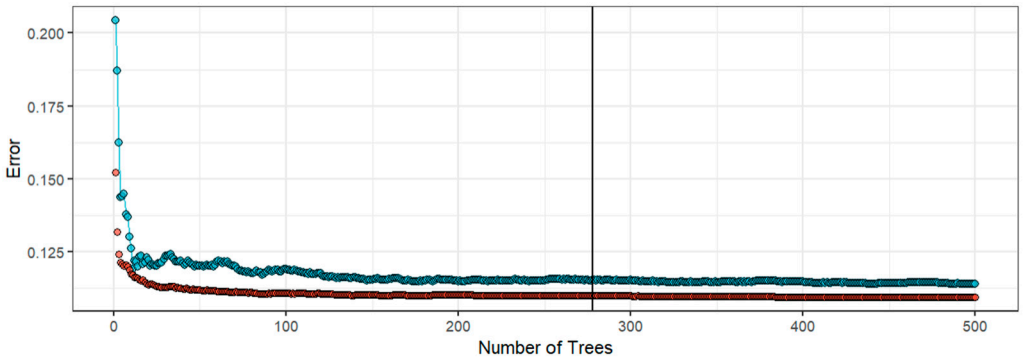


Figure 13. Representation of the test error (blue) and training error (red) for different number of trees of the random forest model in the first group of predictors.

3.3.2. Random Forest Model for the Second Group of Predictors (Density of the Fresh Paving Block and Percentage of Water Absorption)

For the prediction with the variable density and percentage of water absorption, an optimal number of trees of 144 is obtained by cross-validation; The importance of each predictor is shown in Figure 14, where the percentage of water absorption is the most important variable, followed by the density of the fresh paving block. Figure 15 indicates the prediction of the tensile splitting strength with the use of the predictor variables in three dimensions, where it can be noted that the random forest model has many more steps than the regression tree model, which gives it greater flexibility and better fit in the behavior of

the data than a single tree, shown in Figure 10. Figure 16 shows the errors of the model for the training dataset and test dataset, two predictors as the number of variables for each division, and the resulting performance of the model gives a mean square error of 0.125097.

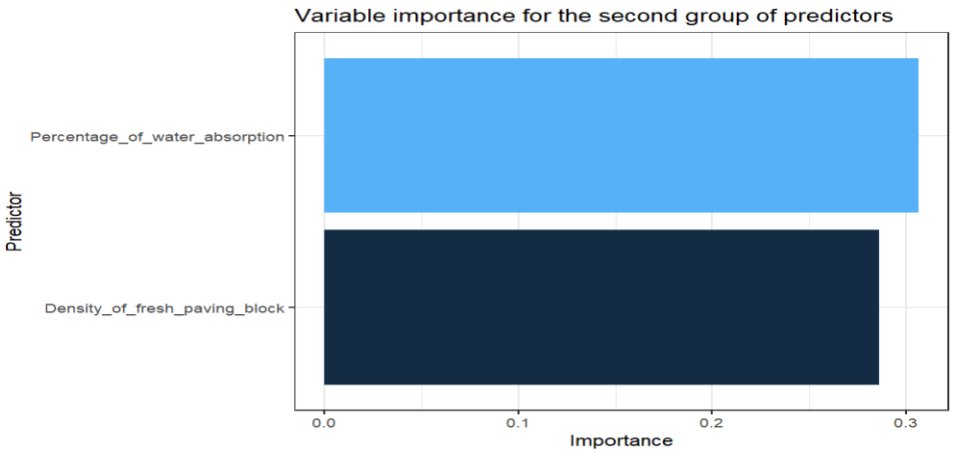


Figure 14. Variable importance for random forest model for the second group of predictors.

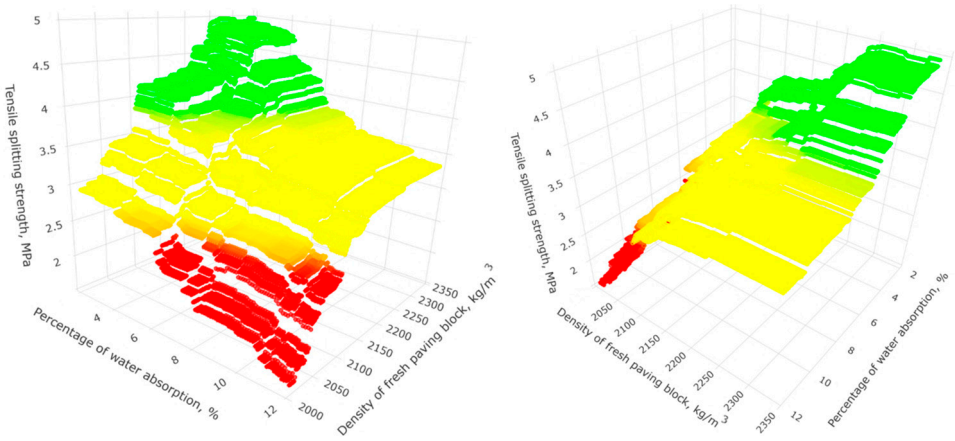


Figure 15. Three-dimensional representation at different angles for predicting the tensile splitting strength response variable using the random forest for the second group of predictor variables. On the left and right the same prediction model is represented at different angles, the small green, yellow and red steps represent the high, medium and low indirect tensile strength, respectively.

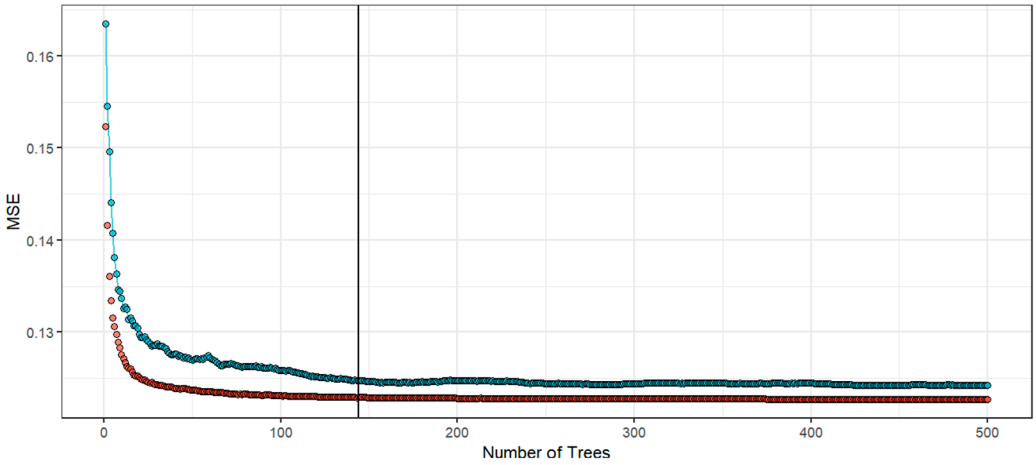


Figure 16. Representation of the test error (blue) and training error (red) for different number of trees of the random forest model in the second group of predictors.

3.4. Neural Networks

3.4.1. Regression Using Neural Networks for the First Group of Predictors (Thickness, Width, Length, Mass of the Fresh Paving Block, and Percentage of Water Absorption)

An artificial neural network model without hidden layers is made, with a normalized layer to be able to eliminate drawbacks regarding the units of measurement of the variables in different scales; an output neuron is taken into consideration since the variable to be predicted (tensile splitting strength) is continuous. It is trained with 100 epochs, and the learning rate is 0.1 for each learning stage, taking into account the gradient descent. The model performance or loss is a mean square error of 0.112198 using the ELU (Exponential Linear Unit) activation function for the neural network with no hidden layer. Figure 17 shows the neural network training calculating the loss for different epochs; the red line indicates the errors for the training dataset, and the blue line for the test dataset. Principal component analysis is applied to specify the number of optimal hidden layers in the neural network. Figure 18 shows that the cumulative variance proportion reaches 0.9999 with two principal components. So, two hidden layers are needed to explain 99.99% of the variability.

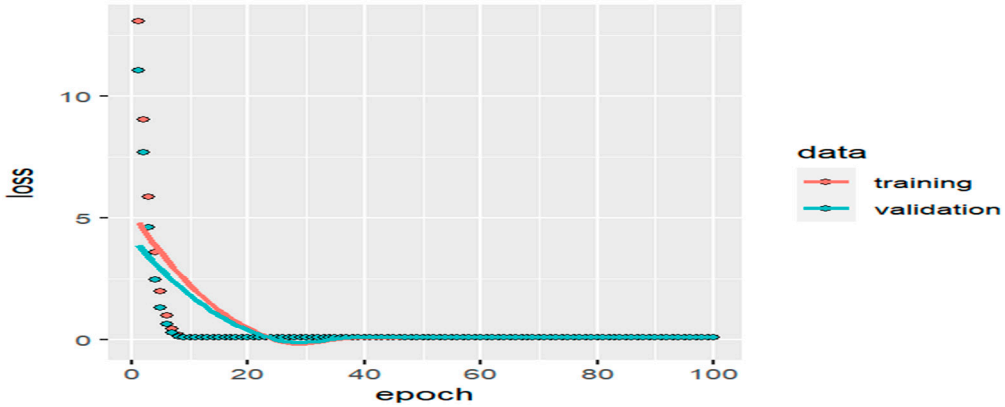


Figure 17. Training the artificial neural network with no hidden layer for the first group of predictors.

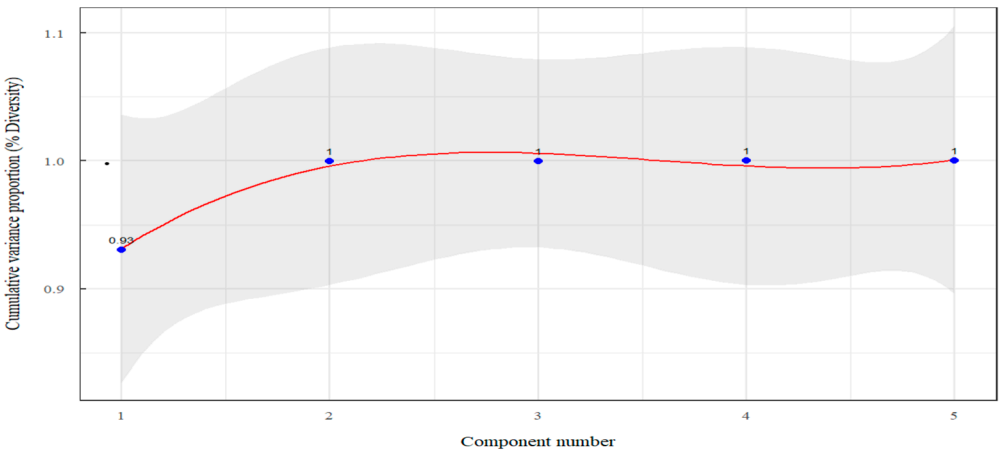


Figure 18. Cumulative variance proportion according to the number of principal components for the first group of predictors. The values in the graph indicate the cumulative variance for different numbers of principal components.

For the first hidden layer, the number of neurons is determined by iterating from 2 to 10 neurons, calculating the mean square error (loss), which is lower with 10 neurons. The training process stabilizes with 40 epochs, which can be evidenced in Figure 19, whose result of the model for one hidden layer of 10 neurons through the ELU activation function gives a mean square error of 0.114271; the structure of the model can be seen in Figure 20. For the second hidden layer, the optimal number of neurons is defined by iterating the

results of the mean square error with the activation function ELU; the optimal number of neurons is ten, and the training process is visualized in Figure 21, where the error for the training dataset is in red and blue for the test dataset, the mean square error of the model is 0.156214, and the architecture is visualized in Figure 22. The variables enter the model, creating paths through interconnected networks of nodes that represent neurons, each node processes the input, generating an output through the activation function, and during training, the weights are adjusted to minimize the prediction error.

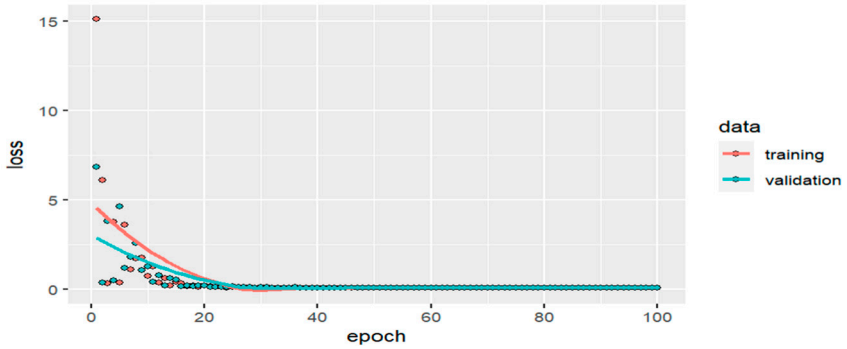


Figure 19. Training of the artificial neural network for one hidden layer with ten neurons for the first group of predictors.

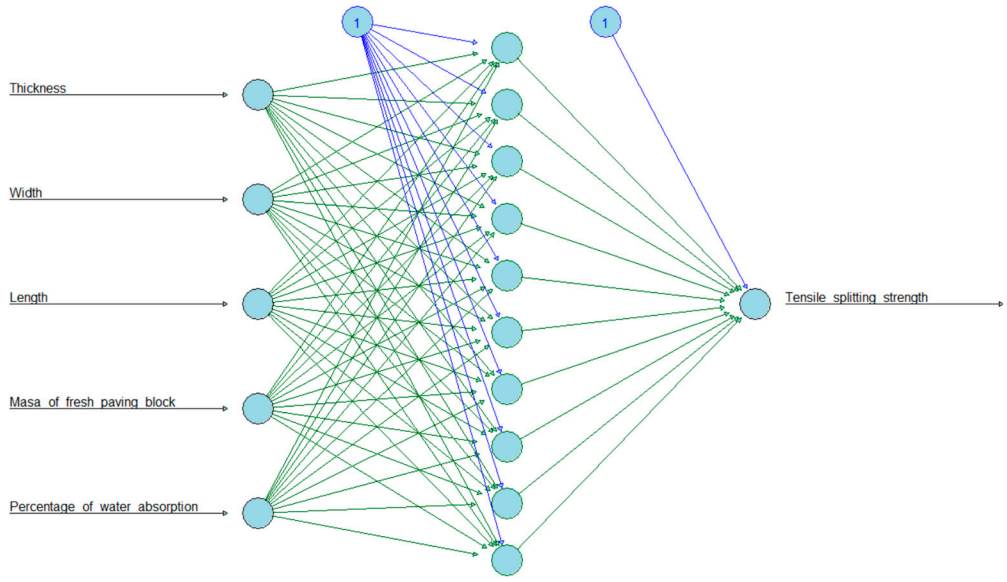


Figure 20. The architecture of the artificial neural network for one hidden layer with ten neurons for the first group of predictors. The circles represent neurons, and the lines are the connections in the neural network.

3.4.2. Regression Using Neural Networks for the Second Group of Predictors (Density of the Fresh Paving Blocks and Percentage of Water Absorption)

The neural network model without hidden layers, only with the normalization layer and one output neuron, using an ELU activation function, is trained with 100 epochs, giving a mean square error performance of 0.112402. Figure 23 shows the training with the errors of the training dataset and test dataset. Using principal component analysis, it is determined that one hidden layer explains 99.9% of the variability; this can be seen in Figure 24, and the number of neurons is determined by iteration from 2 to 10 using the ELU activation function. The model's performance results in a mean square error of 0.116783 with ten neurons for one hidden layer.

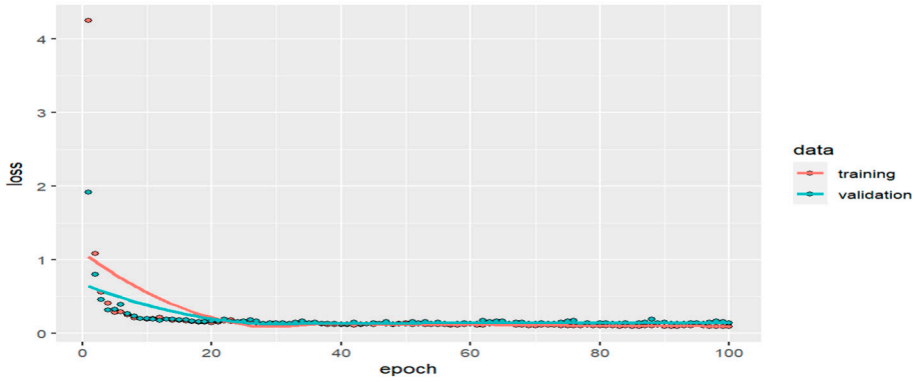


Figure 21. Training of the artificial neural network for two hidden layers for the first group of predictors.

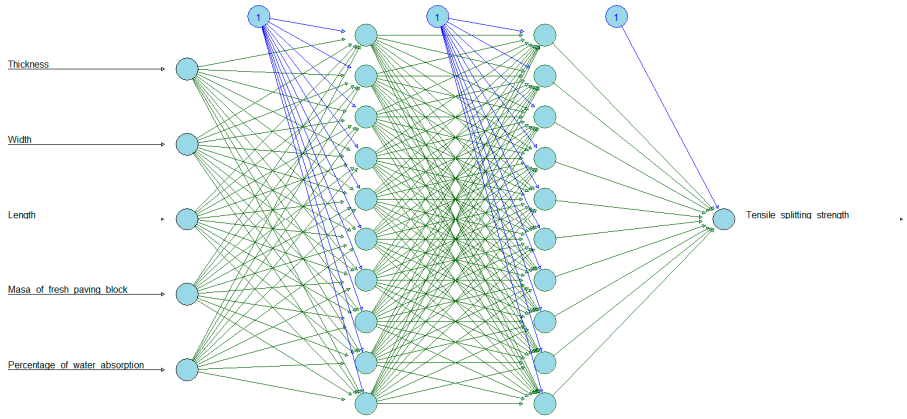


Figure 22. The architecture of the artificial neural network for two hidden layers for the first group of predictors. The circles represent neurons, and the lines are the connections in the neural network.

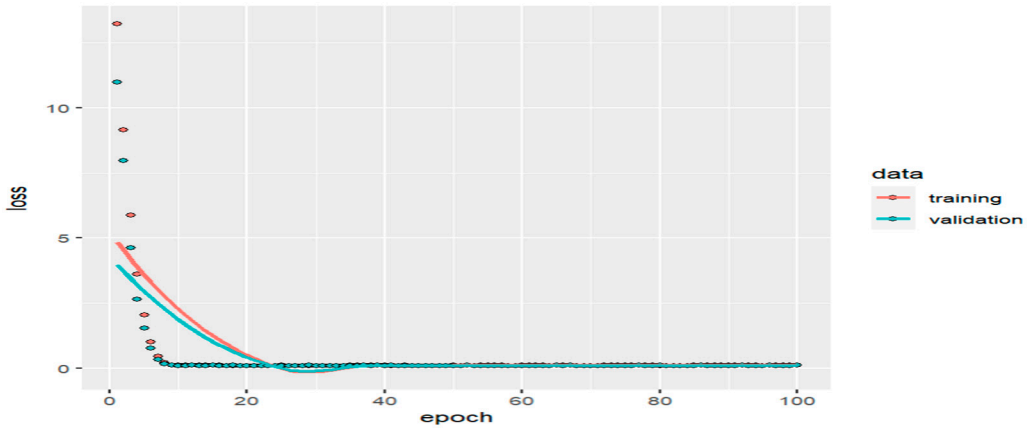


Figure 23. Training of the artificial neural network without hidden layers for the second group of predictors.

Figure 25 indicates the prediction plane of the tensile splitting strength and the predictor variables percentage of water absorption and density of fresh paving block for the neural network model without hidden layers, showing a plane with a slight curvature. On the contrary, the prediction plane shown in Figure 26 with 1 hidden layer of 10 neurons presents more complexity, since the plane is more flexible, with more curvatures, and this is caused by the fact that there are more neural connections, whose structure is shown in Figure 27.

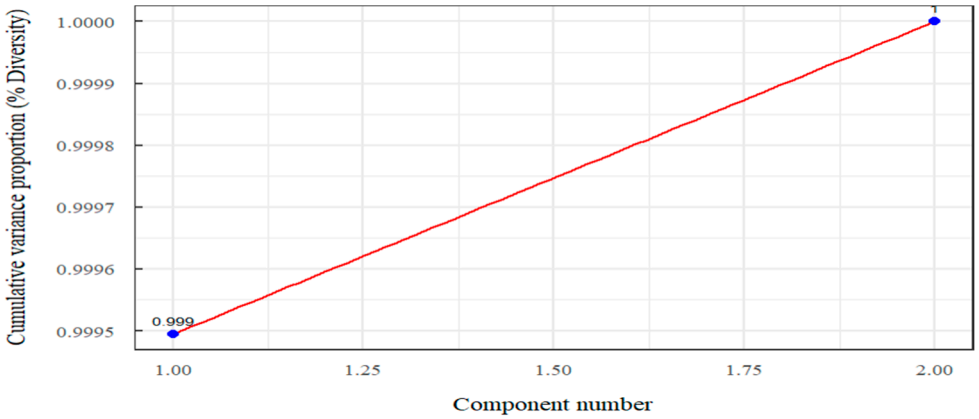


Figure 24. Cumulative variance proportion according to the number of principal components for the second group of predictors.

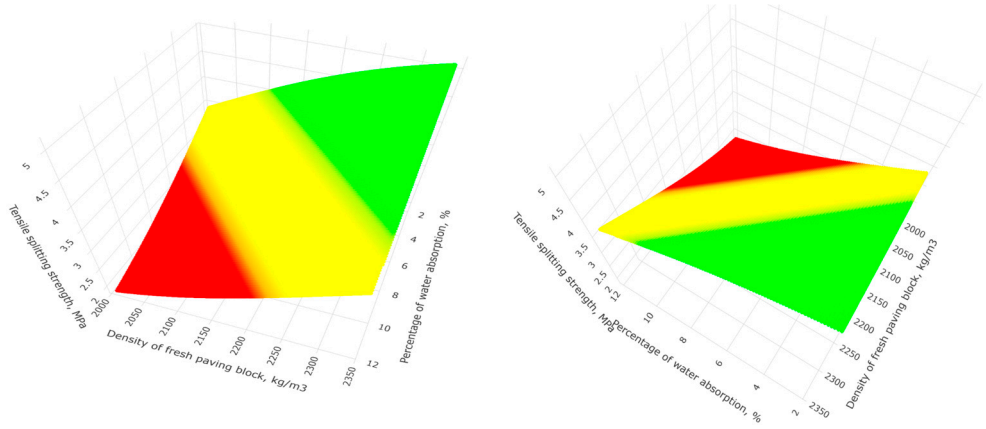


Figure 25. Three-dimensional representation at different angles of the nonlinear prediction of the tensile splitting strength response variable using artificial neural networks for the second group of predictor variables without hidden layers.

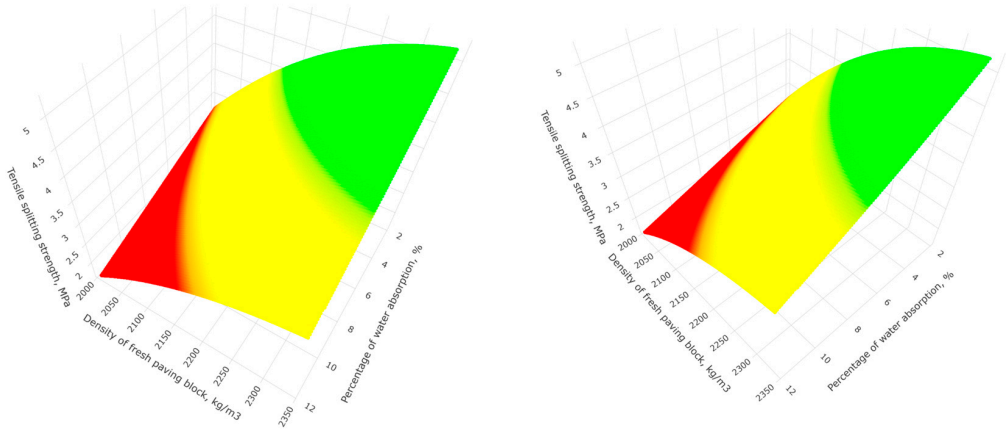


Figure 26. *Cont.*

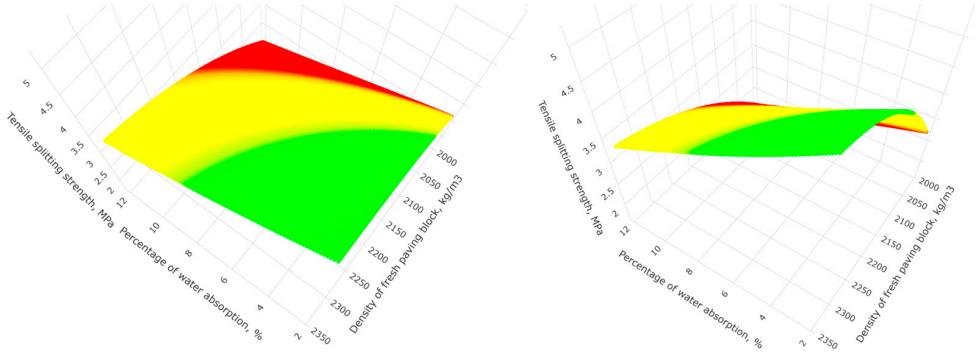


Figure 26. Three-dimensional representation at different angles of the nonlinear prediction of tensile splitting strength response variable using artificial neural networks for the second group of predictor variables with one hidden layer of 10 neurons.

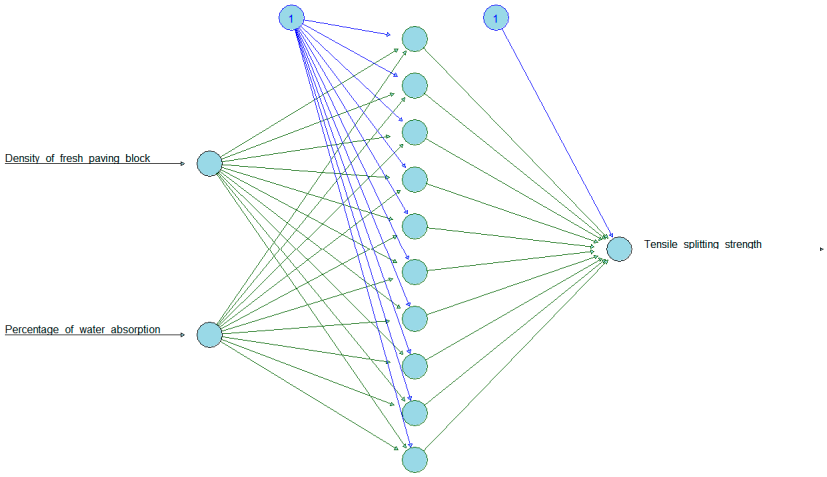


Figure 27. The architecture of the artificial neural network for predicting the tensile splitting strength for the second group of predictors (density of the fresh paving block and percentage of water absorption). The circles represent neurons, and the lines are the connections in the neural network.

Table 2 shows each model performance using the mean square error for the two groups of predictors.

Table 2. Performance of models for each group of predictors using the mean square error (MSE).

MODEL	MSE (Thickness, Length, Width, Mass of the Fresh Paving Block, and Percentage of Water Absorption)	MSE (Density of the Fresh Paving Block and Percentage of Water Absorption)
Multiple Linear Regression	**0.110086**	0.115044
Regression tree	0.165174	0.139050
Random forest	0.115392	0.125097
Neural network (without layers)	0.112198	**0.112402**
Neural network (1 layer)	0.114271	0.116783
Neural network (2 layers)	0.156214	NA

Values in bold indicate the lowest error of the methods for the two groups of predictors, and NA means not applicable.

4. Discussion

Based on the results and literature, it can be noted that regression trees are not a robust technique from the data for the prediction of tensile splitting strength. However, the neural network allows nonlinear behaviors to be learned, and based on the study carried out, quality can be guaranteed in terms of tensile splitting strength, knowing the explanatory variables in the production process, which is a contribution compared to related works, such as those that relate the components of the mixture, or a predictor variable which explains in a limited way the population behavior of the tensile splitting strength.

Tensile splitting strength prediction using neural networks was studied by [22], resulting in a mean square error of 0.141; neural network models should be compared with different layers and numbers of neurons for better modelling, explaining the activation function used. Table 2 indicates a mean square error of 0.112198 for the neural network model without layers analyzed in the first group of predictors.

The graphs presented in this research are a contribution to the understanding of the prediction method used and the behavior of the data, as in Figure 5, where the prediction of the response variable tensile splitting strength forms a plane whose value point moves three-dimensionally according to Equation (33), where the predictor variables create the projected dimension through linear behavior, which is very characteristic of multiple linear regression. Figure 10 allows us to graphically understand in three dimensions the conditions of the second group of predictor variables in the regression tree, where each rung predicts the tensile splitting strength. The prediction using random forest is represented in Figure 15 with an optimal number of 144 trees. Figures 25 and 26 show the nonlinear behavior for the prediction using artificial neural networks for the second group of predictor variables.

It is observed in Equation (32) for the first group of predictors that the variable with the most significant influence for the prediction in MLR is the thickness followed by the percentage of water absorption; for random forests, the importance is shown in Figure 12, with the percentage of water absorption coming in first place, followed by the mass of the fresh paving block, the thickness, width, and length. For the second group of predictors, Equation (33) indicates a more significant influence on the percentage of water absorption, showing this importance in Figure 14, followed by the density of the fresh paving block.

It is evident in Figure 1 that the density of the fresh paving block and the percentage of water absorption has a high correlation with the tensile splitting strength and indicates that the behavior of the data is linear for the prediction, which can be distinguished in Figures 3 and 5, which show the three-dimensional representation with its projection plane using multiple linear regression, and can be confirmed in Figures 18 and 24, where it can be seen that one principal component explains more than 90% of the accumulated variance in the dataset for the two groups of predictors.

The overfitting was controlled by simplifying the model, regularizing and stopping the training just in time, taking into account the principle of parsimony; a low test error is achieved, generalizing the predictions well. Figures 1–3 indicate the behavior of the data. For the regression trees of the first group of predictors, Figure 8 shows a low test error, with eight nodes, and is the optimal training stopping point to avoid overfitting, since with more nodes and higher model complexity, the errors of training decrease, but validation errors increase, which makes new data poorly predicted. For the second group of predictors, a similar analysis with nine nodes can be seen in Figure 11. For the random forest model, the number of optimal trees can be seen in Figure 13, where it stabilizes as the number of trees increases, remaining at similar values along the horizontal axis; this event occurs for the training and validation data. Similarly, for neural networks, Figures 17, 19, 21 and 23 show a point at which it stabilizes for a number of optimal epochs, indicating no overfitting problem.

Multiple linear regression is a classic method that must comply with the structural assumptions. The regression trees have low predictive power with limited prediction values. The set of trees generates a random forest, which significantly improves this characteristic but with little interpretability, since it is considered a black box where information

enters and a result comes out. Neural networks are also considered black boxes, but their prediction capacity can be applied to non-linear behaviors.

The scope of the study focused on concrete pavers with dimensions of 20 cm × 10 cm × 6 cm from a specific manufacturing process in a company in Quito-Ecuador. Based on the research, future studies may expand the sample to different factories, including specific variables in the statistical control of processes and quality in different production conditions. This study lays a solid foundation for developing multivariate predictions in concrete pavers, broadening the perspective on quality control and optimization in the industrial production of precast vibro-compacted concrete.

5. Conclusions

The study proposes different models to predict the tensile splitting strength through the explanatory variables in the concrete paving block production process and the percentage of water absorption in a company in Quito-Ecuador. The first group of predictor variables includes thickness, width, length, mass of the fresh paving block, and percentage of water absorption. The second group of predictor variables includes the density of the fresh paving block and the percentage of water absorption. The *R* programming language is used to carry out descriptive and inferential statistical analysis, multivariate models, and three-dimensional graphs, with the advantage and freedom that programming generates to deepen the investigation, allowing one to understand the behavior of the data and models. Additionally, *Python* is used with the Anaconda distribution to use the *Keras* and *TensorFlow* packages with the *articulate* library.

The variable with the greatest influence on the prediction of indirect tensile strength is the percentage of water absorption, which significantly influenced the random forest study shown in Figures 12 and 14, which is corroborated by Equation (33) of multiple linear regression, where the coefficient is greater. The study allowed us to determine the capacity of the developed models, errors, and their practical advantages, with the conclusion that the multiple linear regression makes it easier to apply the values in the equation to obtain the punctual prediction with simple calculations; the regression tree allowed us to follow a path-specific conditional according to the values of predictors, while random forests require the use of software for their application, and neural networks, having greater flexibility to learn behaviors, such as nonlinear ones, require software, due to its high predictive capacity.

Table 2 shows that the best model to predict the tensile splitting strength in the first group of predictors is multiple linear regression, with a mean square error (MSE) of 0.110086 and an adjusted coefficient of determination of 0.7974, followed by the neural network without hidden layers, with an MSE of 0.112198. The best model for the second group of predictors is the neural network without hidden layers, with a mean square error (MSE) of 0.112402, followed by multiple linear regression, with an MSE of 0.115044 and an adjusted coefficient of determination of 0.7897. The worst method used for prediction is the regression tree, with a mean square error of 0.165174 for the first group of predictors and 0.139050 for the second. Also, the regression tree method has the worst ability to avoid overfitting and greater risk of suffering from it, which can be seen in Figures 8 and 11. Therefore, it is concluded that it is possible to predict the tensile splitting strength through the predictor variables of the first and second groups, allowing us to determine in advance the results inferred to the population from the production process and, with the water absorption test, to guarantee the quality of tensile splitting strength of the paving block.

Author Contributions: Conceptualization, V.R.B.-R., W.J.Y.-V. and E.P.H.-G.; methodology, W.J.Y.-V. and E.P.H.-G.; software, V.R.B.-R. and E.P.H.-G.; validation, W.J.Y.-V. and E.P.H.-G.; formal analysis, V.R.B.-R., W.J.Y.-V. and E.P.H.-G.; investigation, V.R.B.-R.; writing—review and editing, V.R.B.-R.; visualization, V.R.B.-R.; supervision, W.J.Y.-V. and E.P.H.-G.; project administration, W.J.Y.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Carchi State Polytechnic University (www.upec.edu.ec).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is unavailable due to privacy or ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *NTE INEN 3040*; Adoquines de Hormigón. Requisitos y Métodos de Ensayo. Instituto Ecuatoriano de Normalizacion: Quito, Ecuador, 2016.
2. *ASTM C496*; Method for Splitting Tensile Strength of Cylindrical Concrete Specimens. ASTM: West Conshohocken, PA, USA, 2002.
3. *INEN 1485*; Determinación de la Resistencia a la Compresión. Instituto Ecuatoriano de Normalizacion: Quito, Ecuador, 1986.
4. Purwanto, P.; Priastiwi, Y. Testing of concrete paving blocks the bs en 1338:2003 british and european standard code. *Teknik* **2008**, *29*.
5. Zain, M.F.M.; Mahmud, H.B.; Ilham, A.; Faizal, M. Prediction of splitting tensile strength of high-performance concrete. *Cem. Concr. Res.* **2002**, *32*, 1251–1258. [CrossRef]
6. Haktanir, T.; Arı, K. Splitting strength and abrasion resistance of concrete paving blocks as a function of dry bulk specific gravity and ultrasonic pulse velocity. *Mater. Constr. Mater. Constr.* **2005**, *55*, 5–12. [CrossRef]
7. Poon, C.-S.; Chan, D. Effects of contaminants on the properties of concrete paving blocks prepared with recycled concrete aggregates. *Constr. Build. Mater.* **2007**, *21*, 164–175. [CrossRef]
8. Ling, T.-C. Prediction of density and compressive strength for rubberized concrete blocks. *Constr. Build. Mater.* **2011**, *25*, 4303–4306. [CrossRef]
9. Ohemeng, E.A.; Yalley, P.P.K. Models for predicting the density and compressive strength of rubberized concrete pavement blocks. *Constr. Build. Mater.* **2013**, *47*, 656–661. [CrossRef]
10. Dervishi, F.; Luga, E. Relation between Physical and Mechanical Properties of Concrete Paving Blocks. In Proceedings of the 2nd International Congress on Roads, Tirana, Albania, 24–25 September 2015.
11. Skripkiunas, G.; Girskas, G.; Malaiškienė, J.; Šemelis, E. Variation of Characteristics of Vibropressed Concrete Pavement Blocks. *Constr. Sci.* **2014**, *15*. [CrossRef]
12. Wong, S.H.; Shek, P.N.; Saggaff, A.; Tahir, M.M.; Lee, Y.H. Compressive strength prediction of lightweight foamed concrete with various densities. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *620*, 012043. [CrossRef]
13. Caronge, M.A.; Lando, A.T.; Djamaluddin, I.; Tjaronge, M.W.; Runtulalo, D. Development of eco-friendly paving block incorporating co-burning palm oil-processed tea waste ash. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *419*, 012158. [CrossRef]
14. Jonbi, J.; Fulazzaky, M.A. Modeling the water absorption and compressive strength of geopolymer paving block: An empirical approach. *Measurement* **2020**, *158*, 107695. [CrossRef]
15. Mudjanarko, S.W.; Julianto, E.; Harmanto, D.; Wiwoho, F.P. Addition of Gravel in the Manufacture of Paving Block with Water Absorption Capability. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *498*, 012031. [CrossRef]
16. Al-Kheetan, M.J. Properties of lightweight pedestrian paving blocks incorporating wheat straw: Micro-to macro-scale investigation. *Results Eng.* **2022**, *16*, 100758. [CrossRef]
17. Zhang, S.P.; Zong, L. Evaluation of relationship between water absorption and durability of concrete materials. *Adv. Mater. Sci. Eng.* **2014**, *2014*, 650373. [CrossRef]
18. Dong, Q.; Chen, X.; Dong, S.; Ni, F. Data Analysis in Pavement Engineering: An Overview. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 22020–22039. [CrossRef]
19. Saridemir, M. Empirical modeling of splitting tensile strength from cylinder compressive strength of concrete by genetic programming. *Expert Syst. Appl* **2011**, *38*, 14257–14268. [CrossRef]
20. Nazari, A.; Azimzadegan, T. Prediction the effects of ZnO_2 nanoparticles on splitting tensile strength and water absorption of high strength concrete. *Mater. Res.* **2012**, *15*, 440–454. [CrossRef]
21. Liu, Y. High-Performance Concrete Strength Prediction Based on Machine Learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 5802217. [CrossRef]
22. Amin, M.N.; Ahmad, A.; Khan, K.; Ahmad, W.; Nazar, S.; Faraz, M.I.; Alabdullah, A.A. Split Tensile Strength Prediction of Recycled Aggregate-Based Sustainable Concrete Using Artificial Intelligence Methods. *Materials* **2022**, *15*, 4296. [CrossRef]
23. Cabana, E.; Lillo, R.E.; Laniado, H. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Stat. Pap.* **2021**, *62*, 1583–1609. [CrossRef]
24. Gnanadesikan, R.; Kettenring, J.R. Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics* **1972**, *28*, 81–124. [CrossRef]
25. Ghorbani, H. Mahalanobis distance and its application for detecting multivariate outliers. *Facta Univ. Ser. Math. Inform.* **2019**, *34*, 583. [CrossRef]
26. Sheather, S. *A Modern Approach to Regression with R*; Springer: New York, NY, USA, 2009. [CrossRef]

27. Geron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2019.
28. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013. [CrossRef]
29. Breiman, L.; Friefman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: New York, NY, USA, 1984.
30. Breiman, L. Random Forests. In *Machine Learning*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001.
31. Reddy, G.T.; Reddy, M.; Lakshmanna, K.; Kaluri, R.; Rajput, D.; Srivastava, G.; Baker, T. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* **2020**, *8*, 54776–54788. [CrossRef]
32. Frost, H.R. Eigenvectors from Eigenvalues Sparse Principal Component Analysis. *J. Comput. Graph. Stat.* **2022**, *31*, 486–501. [CrossRef]
33. Jolliffe, I.T. *Principal Component Analysis*, 2nd ed.; Springer: New York, NY, USA, 2002.
34. Ibnu Choldun R, M.; Santoso, J.; Surendro, K. Determining the number of hidden layers in neural network by using principal component analysis. In *Advances in Intelligent Systems and Computing*; Springer: Cham, Switzerland, 2020; pp. 490–500. [CrossRef]
35. Rachmatullah, M.I.C.; Santoso, J.; Surendro, K. Determining the number of hidden layer and hidden neuron of neural network for wind speed prediction. *PeerJ Comput. Sci.* **2021**, *7*, e724. [CrossRef] [PubMed]
36. Mielsen, M. *Neural Networks and Deep Learning*; Springer: Cham, Switzerland, 2019.