Machine Learning-Based Optimization Of The Mix Design Of Lightweight Concrete For Enhanced Mechanical Properties

Authors

- **Ayyan Iqbal**
  · maiqbal2
  University of Illinois

- **Shayan Khan**
  · shayank491
  University of Illinois

- **Dhatrika Varma Borukati**
  · DeeVarma24
  University of Illinois

1. Project Proposal

The escalating global population drives the increasing demand for concrete, thereby fostering the development and adoption of Lightweight Aggregate (LWA)-based Lightweight Concrete (LWC). The widespread availability of LWAs, coupled with straightforward and conventional casting techniques, has facilitated industry-wide acceptance [1]. LWC has found extensive applications in lightweight infill panels, structural concrete, and precast concrete. Notably, LWC achieves comparable compressive strength to traditional concrete in specific scenarios, albeit with a 25-35% reduction in density [2]. This reduction yields additional benefits, including minimized foundation steel requirements, lower transportation costs, and decreased construction expenditures, rendering LWC a promising solution for sustainable and cost-effective infrastructure development. A significant obstacle in the widespread adoption of Lightweight Concrete (LWC) lies in its intricate mix design process. Unlike Normal-Weight Concrete (NWC), which relies on established codes and iterative fine-tuning, LWC lacks standardized design guidelines. Furthermore, optimizing LWC's density while maintaining compressive and tensile strength poses a substantial challenge due to its sensitive nature, where minor mix design adjustments drastically impact mechanical properties. The complexity is compounded by the varied shapes, sizes, and densities of LWAs, which significantly influence the mix design. In contrast, NWC aggregates exhibit relatively consistent properties. To address this challenge, a machine learning (ML) framework can be employed to predict LWA concrete's mechanical properties, including compressive strength, tensile strength, and density. The development of user-friendly tools, leveraging these ML models, would facilitate iterative design optimization and trial-and-error experimentation for researchers working on specialized LWC mix designs. This predictive tool would not only streamline LWC mix design hence enhancing accuracy in mechanical property prediction but expedite the development of tailored LWC solutions for specific applications. By integrating ML and materials science, this innovative approach would overcome existing design complexities and unlock LWC's full potential. For the CEE-492 semester project, our team objectives are to develop and compare the performance of Artificial Neural Networks (ANN), Gaussian Process Regression (GPR), and Decision Trees in predicting the mechanical properties of Lightweight Concrete (LWC), specifically density, compressive strength, and tensile strength. Initially, relevant data from published online articles would be collected, followed by data preprocessing to ensure consistency and quality. Next, we identify 10 influential input parameters governing LWC mix design through a comprehensive literature review of the latest published review articles. An exploratory data analysis (EDA) is then conducted to uncover trends and relationships within the data. Subsequently, the preprocessed data is divided into training (~70-80%) and testing sets (~20-30%). The training data is then normalized and scaled to optimize model performance. Then the team aims to train the ANN, GPR, and Decision Tree models on the training data, fine-tuning hyperparameters through cross-validation and grid search. Model evaluation is performed on the testing data using the statistical performance indicators i.e., such as mean squared error (MSE), R-squared (R^2), and mean absolute error (MAE). Finally, we compare the performance of the trained models and select the best-performing algorithm. The formulas for the performance metrics are mentioned below in Table 1.

Table 1. Mathematical formulation of the statistical performance indicators used in the report.

Equations
$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(k_{act} - k_{pre}\right)^2}$$
$$R^2 = 1 - \frac{\sum\left(k_{act} - k_{pre}\right)^2}{\sum\left(k_{act} - \bar{k_{pre}}\right)^2}$$

Equations
$$MAE = \frac{1}{m}\sum_{i=1}^{m}\|k_{act} - k_{pre}\|$$
$$MSE = \frac{1}{m}\sum_{i=1}^{m}(k_{act} - k_{pre})^2$$
$$VAF = \left(1 - \frac{\mathrm{var}(k_{act} - k_{pre})^2}{\mathrm{var}(k_{act})}\right) \times 100$$

A longstanding controversy surrounds the efficacy and reliability of Machine Learning (ML) and Artificial Intelligence (AI)–based models, with critics labeling them as “black boxes” that merely identify patterns without providing meaningful insights. To address concerns regarding overfitting and model interpretability, we aim to explain or results by employing local explanation techniques, specifically Partial Dependence Plots (PDP) and Shapley Additive Explanations (SHAP). These methods decipher the relationships between individual input parameters and the model’s output, demystifying the “black box” nature of ML models, validating their reliability and accuracy, and identifying potential biases.

1.1. Dataset description

The data set attached has been collected by the team members from all the scholarly articles from Scopus. The search query used for finding articles was “{Lightweight} AND {concrete} AND {aggregate} AND {strength} AND {density} AND {ML}”. The authors have collected 500 data points from over 50 articles. The data set has the quantities of Cement, sand, fly ash (FA), the density of lightweight aggregate, water absorption of lightweight aggregate, superplasticizer, curing time, and the amount of normal aggregate (normal agg.), as input parameters while the compressive strength, split tensile strength, and density of the concrete were taken as output parameters. The first test columns of the dataset correspond to inputs while the last three correspond to output. All the quantities were normalized by the cement quantity before the start of the analysis. The input and output parameters along with their units have been mentioned below in Table 2 as well.

Table 2. Input and output parameters of the dataset along with their units.

Parameters	Categories (I/O)
Cement (kg/m^3)	I
Fine agg. (kg/m^3)	I
w/b	I
LW agg. (kg/m^3)	I
LW agg. density (kg/m^3)	I
LW agg. water absorption (%)	I
NW agg. (kg/m^3)	I
HRWR (% of binder)	I
Curing Time (days)	I
Fly Ash (kg/m^3)	I
Compressive Strength of LW concrete (MPa)	O

Parameters	Categories (I/O)
Split Tensile Strength of LW concrete (MPa)	O
Density of LW concrete (kg/m^3)	O

- I = Input
- O = Output
- LW = Lightweight
- NW = Normal weight
- w/b = water to binder ratio
- HRWR = High range water reducer

2. Statistical distribution of dataset

Concrete is widely regarded as the most complex composite material, comprising various ingredients and exhibiting diverse properties that make its behavior challenging to predict. Its composition can vary significantly depending on application, environmental conditions, and performance requirements. Figure 1 illustrates the statistical distribution of input parameters for concrete compositions, revealing diverse applications and formulations contributing to complex datasets. Certain parameters, such as water-to-cement (W/C) ratio (0.35-0.5), superplasticizer content (typically ≤1% of binder weight), water absorption of aggregates (~2% of aggregate weight), and curing time (predominantly 28 days), exhibit narrow interquartile ranges, indicating relatively fixed proportions in typical cement-based mixtures [3,4,5,6]. In contrast, lightweight aggregate types and corresponding variations in concrete compositions display larger interquartile ranges, reflecting the broad range of available materials, underscoring the complexity of concrete's widespread use as the world's most utilized composite material [7].

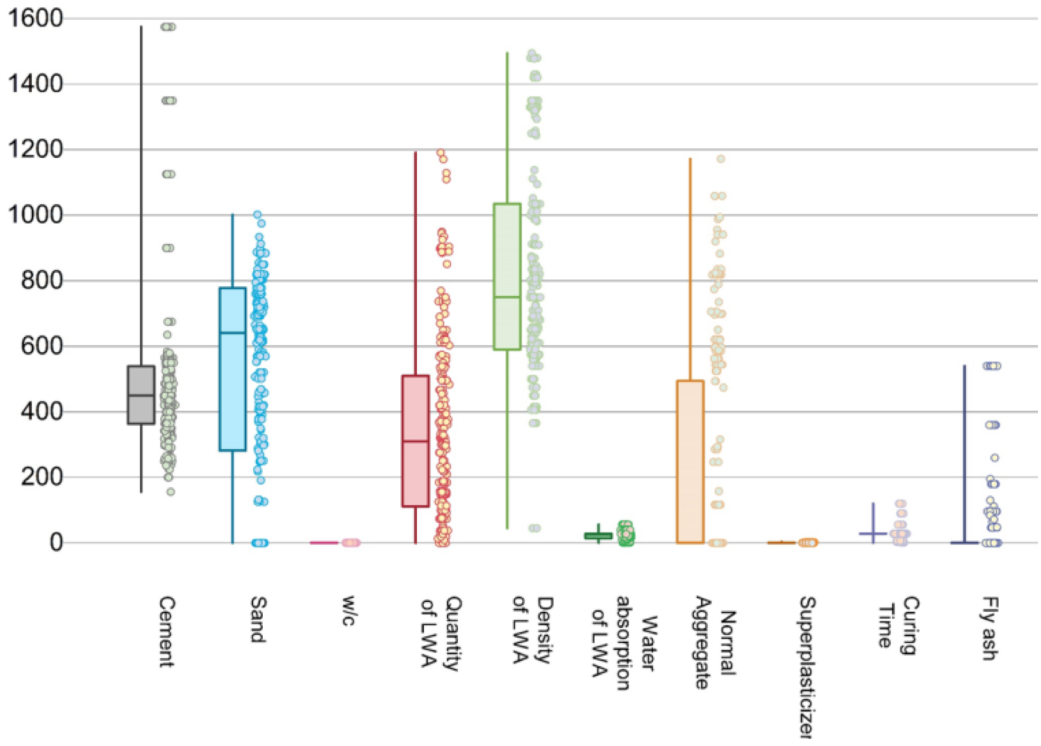

Figure 1: Statistical distribution of the input parameters of the dataset compiled from articles.

The authors have comprehensively compiled a dataset encompassing a wide range of lightweight aggregates from existing literature. These aggregates, derived from industrial waste materials or naturally occurring sources, exhibit spatial variability due to regional differences in availability. Consequently, a diverse array of lightweight aggregates is utilized globally. Figure 2 illustrates the various types of aggregates incorporated in this study. Notably, the dataset reveals that clay-based Lightweight Expanded Clay Aggregate (LECA) predominates, reflecting clay's abundance as a raw material for artificial aggregate production [8]. Furthermore, polystyrene, a prevalent waste material, emerges as a primary source of artificial lightweight aggregates in the dataset [9].

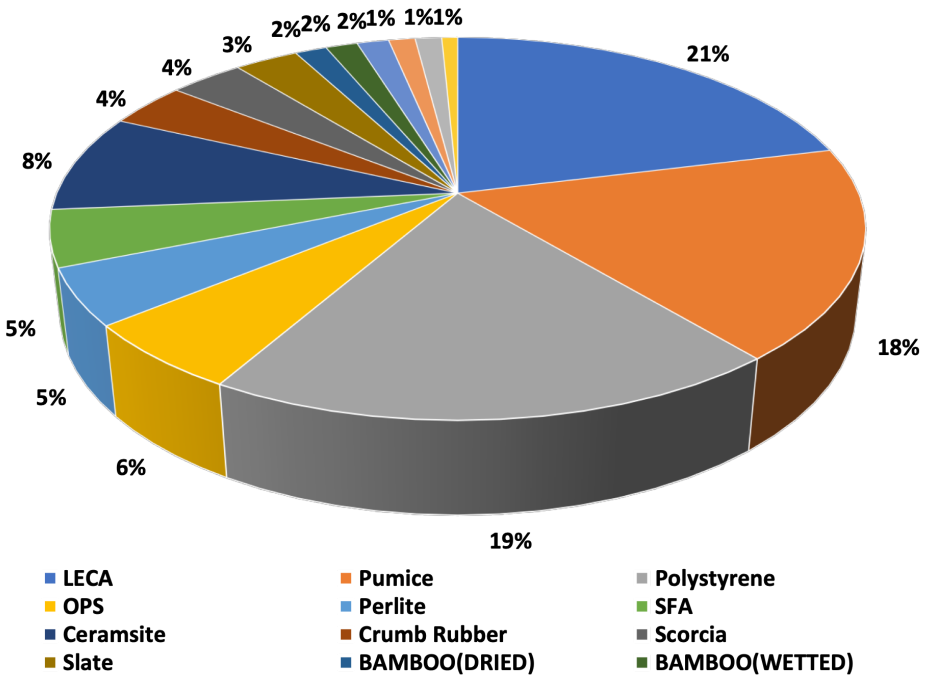

Figure 2: Types of lightweight aggregates used by researchers in the article from which data has been obtained.

Figure 3 illustrates the statistical distribution of compressive strength, tensile strength, and concrete density. The results show that the mean tensile strength is approximately one-tenth of the mean compressive strength (~30 MPa), aligning with established conventions (e.g., ACI codes) [10]. The average density of 1700 kg/m^3 reflects the prevalence of expanded clay aggregate and polystyrene-based concretes since their density lies in this range, validating the dataset's accuracy and reliability for further analysis [11,12].

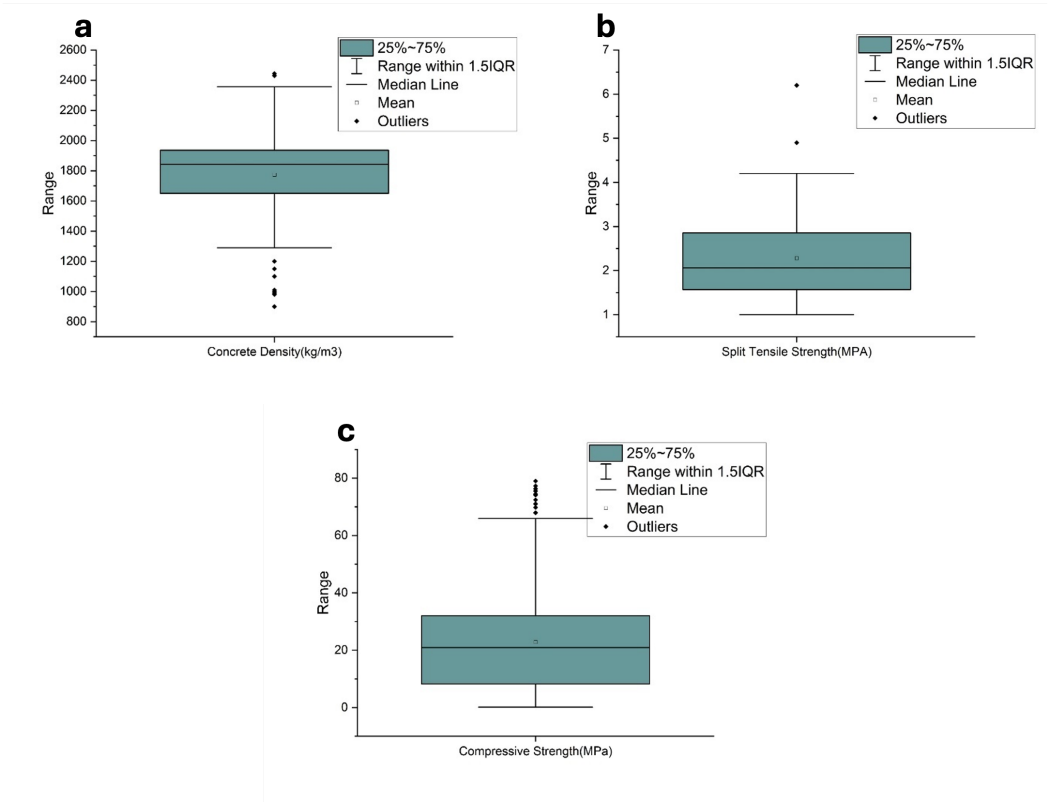

Figure 3: Statistical distributions of the output parameters of the dataset compiled from articles.

2.1. Dataset cleaning and splitting

To maintain accuracy, completeness, consistency, relevance, and validity, the dataset underwent a rigorous cleaning process, a crucial step before applying any machine learning algorithm. This process ensures that the data is processable and enables effective learning for accurate output. Given that the dataset was compiled from X diverse articles sourced from online libraries, there was a high likelihood of human error. However, since most of the dataset was collected by our team, missing values were nonexistent, eliminating the need for removal or imputation. Data cleaning addressed potential issues such as data entry errors, equipment malfunctions, or incomplete surveys. Outliers were identified, and upon examination, most were found to originate from articles published by sources with questionable academic reputations [13,14,15,16,17,18,19]. These outliers were subsequently trimmed to prevent biased analysis and ensure data integrity. Following data cleaning, the refined dataset was split into training (80%) and testing sets (20%). Summary statistics for both are presented in Table 3 and Table 4, providing a comprehensive foundation for reliable model development and evaluation.

Table 3. Summary statistics of dataset set aside for ML model training.

Parameters	Minimum	Maximum	Median	Mode	SD	Type
Cement (kg/m^3)	156	1500	467	480	378.42	I
Fine agg. (kg/m^3)	0	1193	664	0	330.15	I
w/b	0.15	0.80	0.45	0.5	0.08	I
LW agg. (kg/m^3)	23.80	1191	308	37	297.28	I
LW agg. density (kg/m^3)	415	1489	783	575	357.65	I
LW agg. water absorption (%)	0.92	58.30	25.20	40	13.83	I
NW agg. (kg/m^3)	0	1326	0	0	353.98	I
HRWR (% of binder)	0	3	0	0	0.70	I
Curing Time (days)	1	120	28	28	14.27	I
Fly Ash (kg/m^3)	0	540	0	0	117.61	I
Compressive Strength of LW concrete (MPa)	2.03	79	24.58	25	16.68	O
Split Tensile Strength of LW concrete (MPa)	1	7	3.5	3	2	O
Density of LW concrete (kg/m^3)	900	2500	1855	1800	366	O

Table 4. Summary statistics of dataset set aside for ML model testing.

Parameters	Minimum	Maximum	Median	Mode	SD	Type
Cement (kg/m^3)	139	1350	384	450	197.70	I
Fine agg. (kg/m^3)	0	1178	630	0	294.92	I
w/b	0.23	0.8	0.42	0.35	0.07	I
LW agg. (kg/m^3)	0	950	155	0	270.29	I
LW agg. density (kg/m^3)	406	1480	750	610	320.11	I

Parameters	Minimum	Maximum	Median	Mode	SD	Type
LW agg. water absorption (%)	0.92	56	20.5	20.5	13.54	I
NW agg. (kg/m^3)	0	941.2	0	0	282.15	I
HRWR (% of binder)	0	2.5	0.5	0	0.687	I
Curing Time (days)	1	120	28	28	23.4	I
Fly Ash (kg/m^3)	0	540	0	0	111.38	I
Compressive Strength of LW concrete (MPa)	4.28	65.14	27	25	15.54	O
Split Tensile Strength of LW concrete (MPa)	1.2	6.7	3.6	3.1	2.2	O
Density of LW concrete (kg/m^3)	950	2670	1755	1650	354	O

2.2. Dataset cleaning and splitting

Data normalization is a standardization technique for transforming variables to have a common scale. When working with data from different sources or formats, there can be variations in how it is represented, such as differences in units of measurement, data formats, and data structures, making it difficult to compare variables or perform statistical analysis. Data standardization is a crucial step in such cases. The major challenge faced after data collection is processing the raw data to make it compatible with the ML models used. For instance, there was a considerable difference in our dataset between the numerical values of cement, w/c, and normal aggregate used. This difference adversely affected the accuracy of our model. This issue was tackled using the data normalization technique. Data normalization means transforming data into the unit sphere or scaling down the actual values to numerical indexes between 0 and 1. It leads to data cleansing and convergence and significantly enhances the model's efficiency. It also improves data execution by reducing the data set's redundancy. The governing Equation 1 taken into consideration for data normalization is mentioned below, where the normalized value of a certain input variable is a function of the actual, minimum, and maximum values of that variable in the data set. The normalized dataset has been plotted in Figure 4 below

$$y = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (1)$$

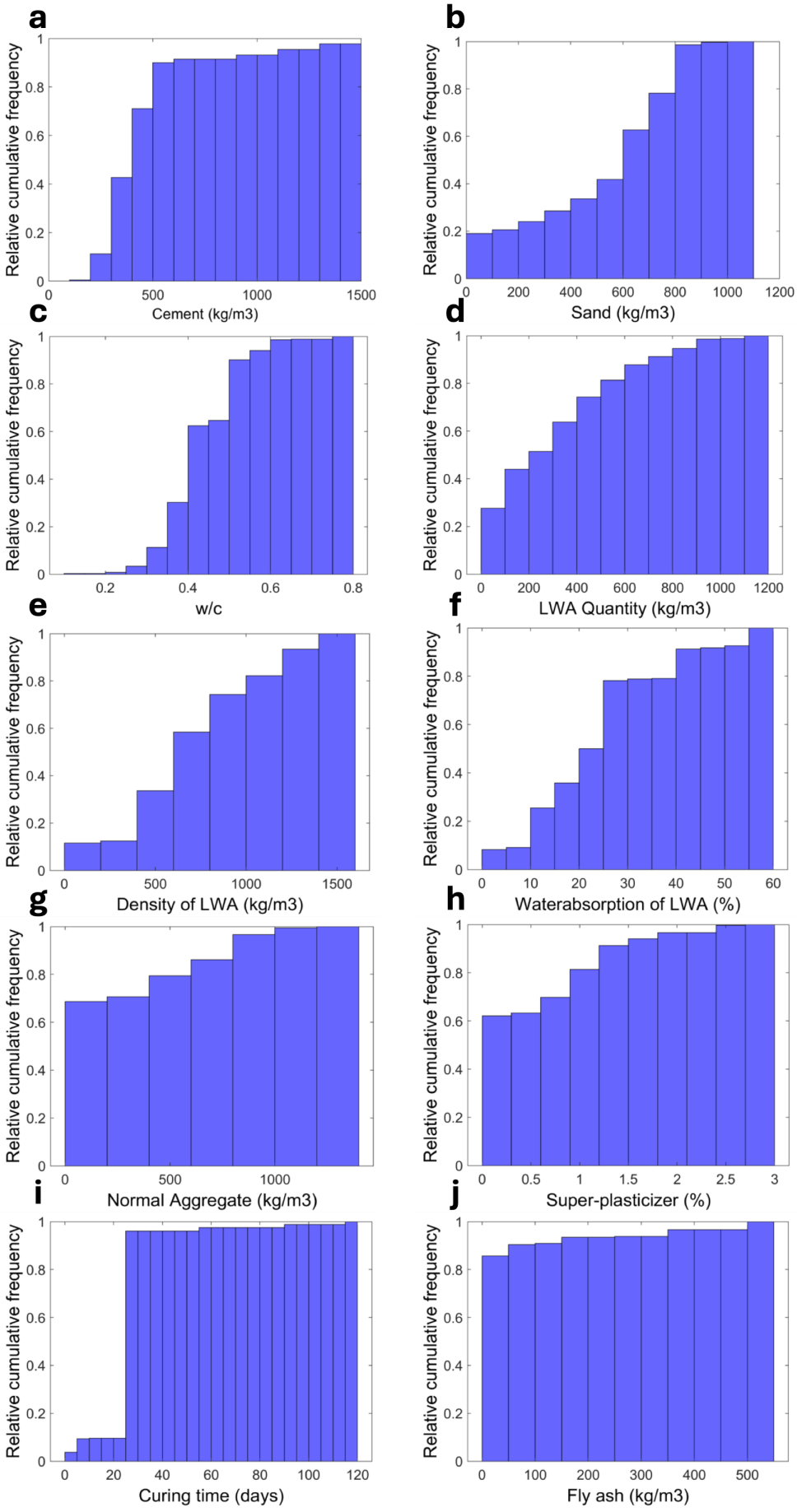

Figure 4: Statistical distribution of the parameters of the dataset compiled after standardization.

2.3. Correlation of input parameters with output parameters

The preprocessing phase proceeded with the plotting of Pearson correlation matrices Figure 5 to elucidate relationships between dependent and independent variables. This comprehensive graph displays pairwise correlations through Pearson correlation coefficients, ranging from -1 to +1. Diagonal entries show perfect correlation (1), while non-diagonal entries exhibit coefficients between -1 and +1, indicating varying degrees of correlation. In the graph template used for this article, blue the color shows the direction of the correlation, and the size of the circle shows magnitude/extent of correlation. Positive coefficients signify direct relationships, and negative coefficients indicate inverse relationships. This analysis, mathematically expressed through the Pearson correlation coefficient (r) Equation 2, provides valuable insights into variable interactions, informing the development of a robust ML model.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2}$$

The correlation matrix of each input with a particular output has been shown separately for a deeper understanding. Keeping the number of input and output parameters in consideration in a single correlation matrix made the visualization difficult. From the correlation matrices, it is evident that the LWA density was the primary factor controlling compressive strength. As per S.A.Khan et al. [20], the difference between normal-weight concrete and lightweight concrete is that lightweight concrete fails due to the failure of aggregates, not matrix so failure concrete with increased lightweight aggregate density had increased compressive strengths as well. Also, an increase in the water-cement ratio causes a decrease in compressive strength, which can also be seen in the graph. Similarly the total fines content or total normal weight aggregate content has a good positive correlation with split tensile strength and concrete density which is also supported by the literature.

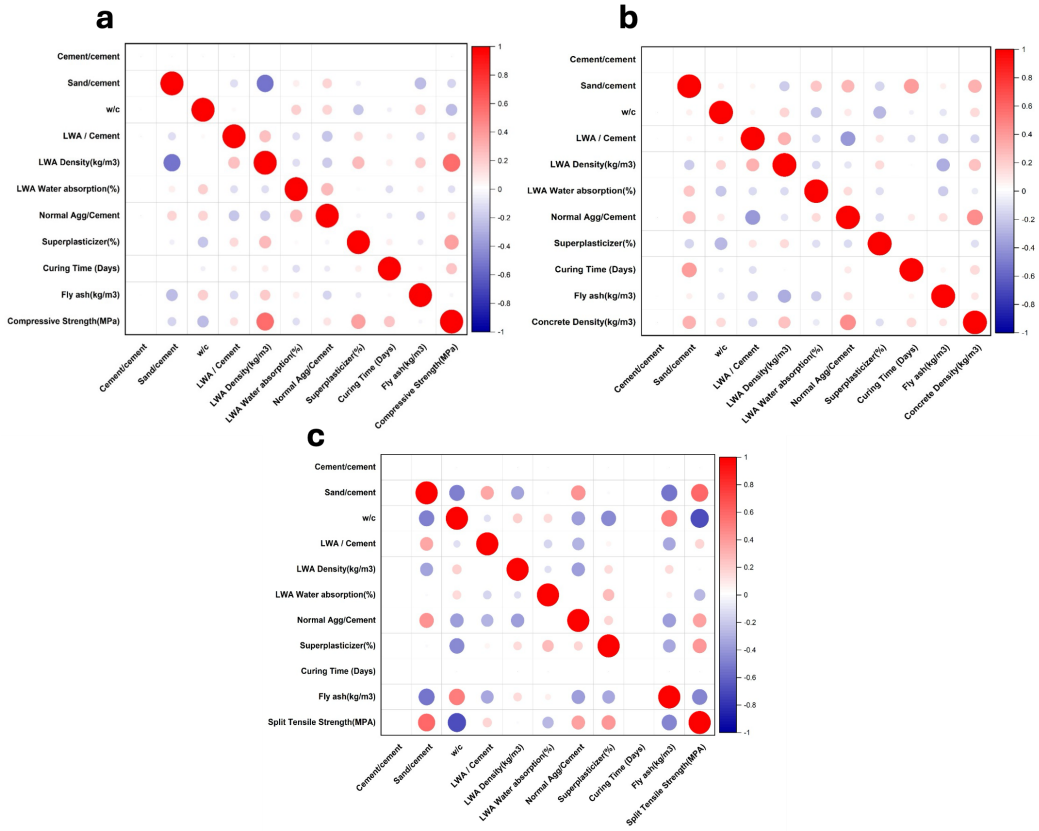

Figure 5: Pearson correlation matrices of input parameters with each output parameter.

Apart from the data cleaning standardization and visualization discussed above, the authors of the report did not require any data augmentation since the number of data points seemed enough for the types of models they intended to train. However, the authors might perform some feature engineering by combining some of the highly correlated input parameters if the statistical performance indicators do not meet their expectations upon training of the model hence changes will be made in the subsequent report

1. Methodology

This report aims to train, validate, and fine-tune Artificial Neural Networks (ANN), Gaussian Process Regression (GPR), and Decision Tree models on the training data for predicting concrete properties. The objectives are to leverage the unique strengths of each model, with ANN capturing non-linear relationships and patterns, Decision Trees providing interpretable results and feature selection, and GPR quantifying uncertainty and handling sparse data. By combining these models, the challenges in concrete property prediction, including non-linear relationships, variability, and limited data, can be effectively addressed.

Additionally, this report addresses the longstanding controversy surrounding the efficacy and reliability of Machine Learning (ML) and Artificial Intelligence (AI) based models, often labeled as "black boxes" that merely identify patterns without providing meaningful insights. To alleviate concerns regarding overfitting and model interpretability, local explanation techniques, specifically Partial Dependence Plots (PDP) and Shapley Additive Explanations (SHAP), will be employed to decipher the relationships between individual input parameters and the model's output. This approach ensures model interpretability, validity, reliability, and identification of potential biases, ultimately demonstrating the efficacy of ML models in concrete property prediction.

3.1. Artificial neural network

Artificial Neural Networks (ANNs) are complex computational models inspired by biological neural networks. They process input data, generate output, and adapt through backpropagation training. Proven effective in various domains, ANNs excel in classification, regression, forecasting, and clustering tasks. The implemented ANN model architecture has been depicted in Figure 6 below.

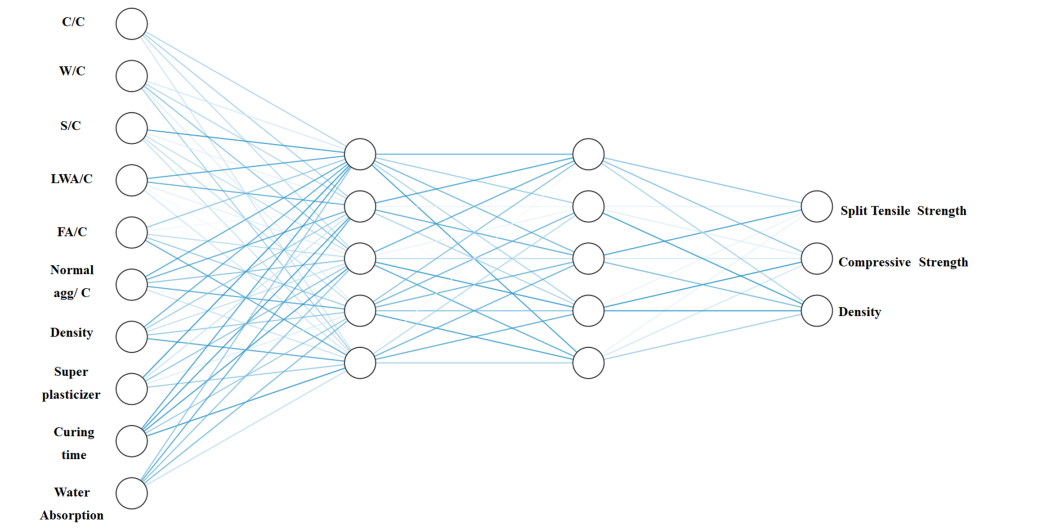

Figure 6: Model architecture of ANN.

3.2. Decision Tree

Decision Trees, a supervised machine learning approach, effectively predicts concrete's mechanical characteristics by modeling complex relationships between input data and output labels. The tree-like structure of Decision Trees provides transparency into prediction outcomes. As a valuable alternative to traditional methods, Decision Trees are a helpful tool for forecasting concrete's mechanical properties, as illustrated in Figure 7.

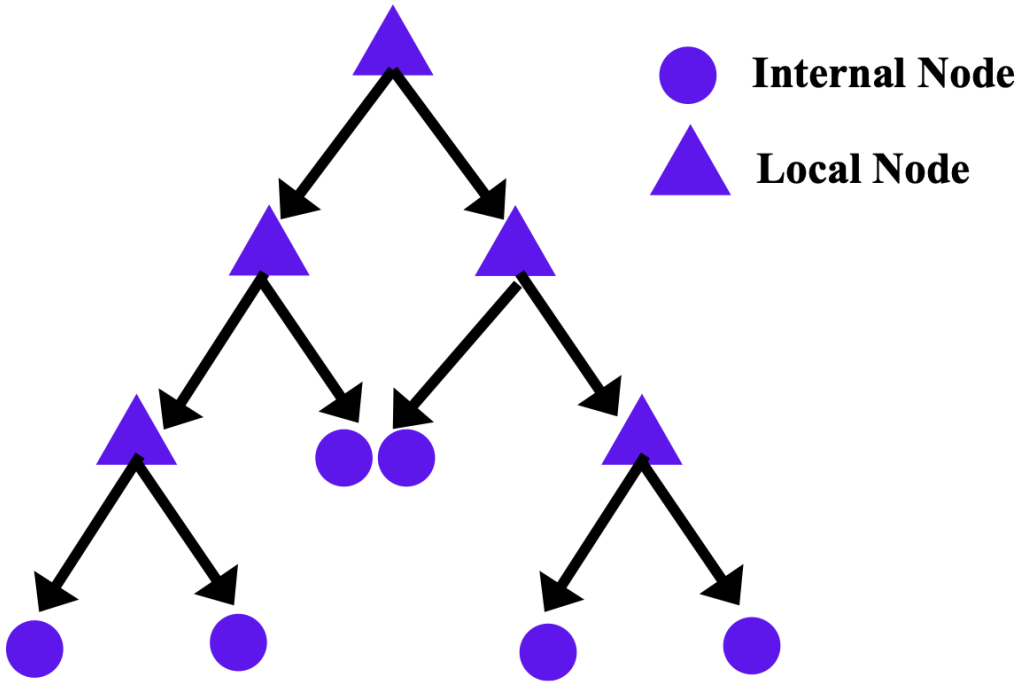

Figure 7: Working mechanism/flowchart of the Decision tree.

3.3. Gaussian Process of Regression

Gaussian Process Regression (GPR) is a supervised machine learning technique using Bayesian inference for predictions. As a non-parametric method [60, 61], GPR excels with limited data, modeling complex relationships between inputs and outputs. Ideal for predicting concrete's mechanical properties, GPR offers accuracy and versatility, making it suitable for diverse applications, as shown in Figure 8.

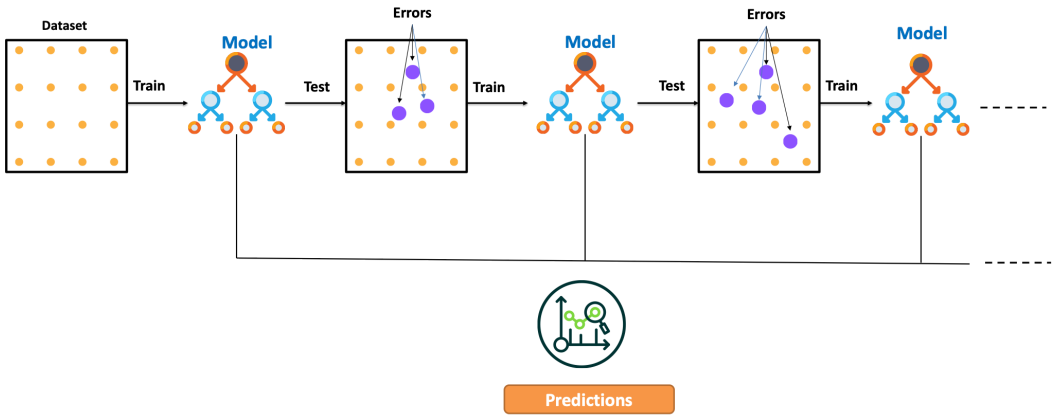

Figure 8: Working mechanism/flowchart of GPR.

In the Figure 9, a complete overview of the whole project is depicted pictorially.

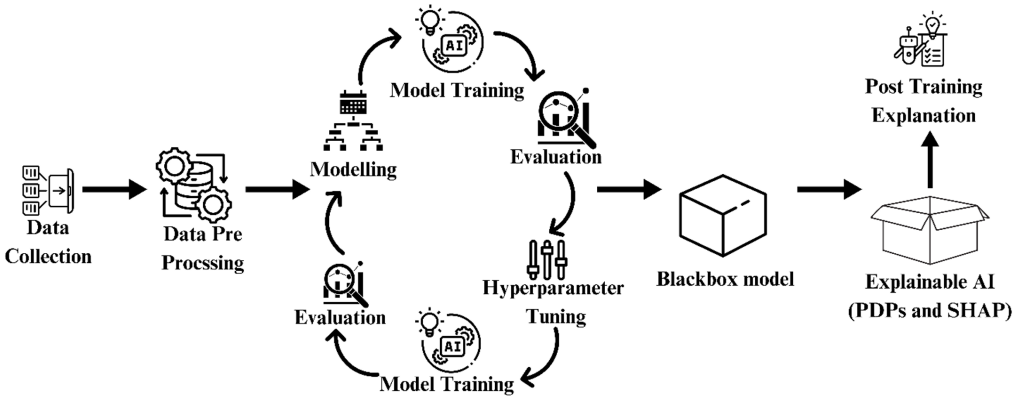

Figure 9: A flowchart explaining the sequence of tasks in the project.

2. Exploratory Data Analysis

Concrete is widely regarded as the most complex composite material, comprising various ingredients and exhibiting diverse properties that make its behavior challenging to predict. Its composition can vary significantly depending on application, environmental conditions, and performance requirements. Figure 1 illustrates the statistical distribution of input parameters for concrete compositions, revealing diverse applications and formulations contributing to complex datasets. Certain parameters, such as water-to-cement (W/C) ratio (0.35-0.5), superplasticizer content (typically ≤1% of binder weight), water absorption of aggregates (~2% of aggregate weight), and curing time (predominantly 28 days), exhibit narrow interquartile ranges, indicating relatively fixed proportions in typical cement-based mixtures [3,4,5,6]. In contrast, lightweight aggregate types and corresponding variations in concrete compositions display larger interquartile ranges, reflecting the broad range of available materials, underscoring the complexity of concrete's widespread use as the world's most utilized composite material [7]..

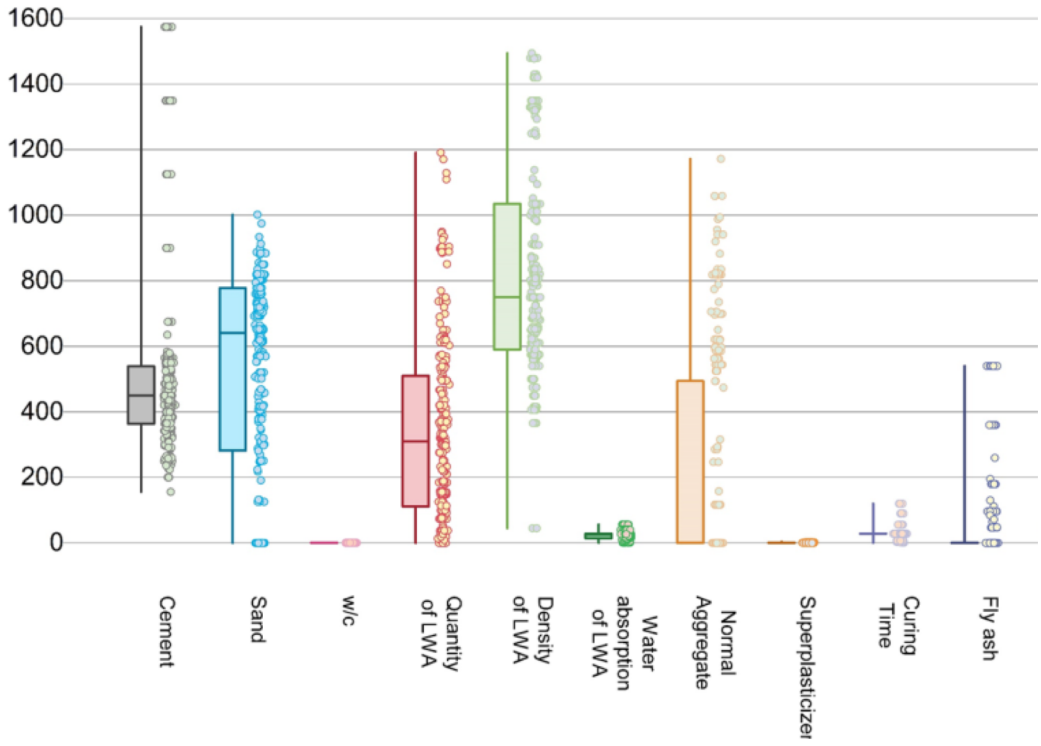

Figure 1: Statistical distribution of the input parameters of the dataset compiled from articles.

The authors have comprehensively compiled a dataset encompassing a wide range of lightweight aggregates from existing literature. These aggregates, derived from industrial waste materials or naturally occurring sources, exhibit spatial variability due to regional differences in availability. Consequently, a diverse array of lightweight aggregates is utilized globally. Figure 2 illustrates the various types of aggregates incorporated in this study. Notably, the dataset reveals that clay-based Lightweight Expanded Clay Aggregate (LECA) predominates, reflecting clay's abundance as a raw material for artificial aggregate production [8]. Furthermore, polystyrene, a prevalent waste material, emerges as a primary source of artificial lightweight aggregates in the dataset [9].

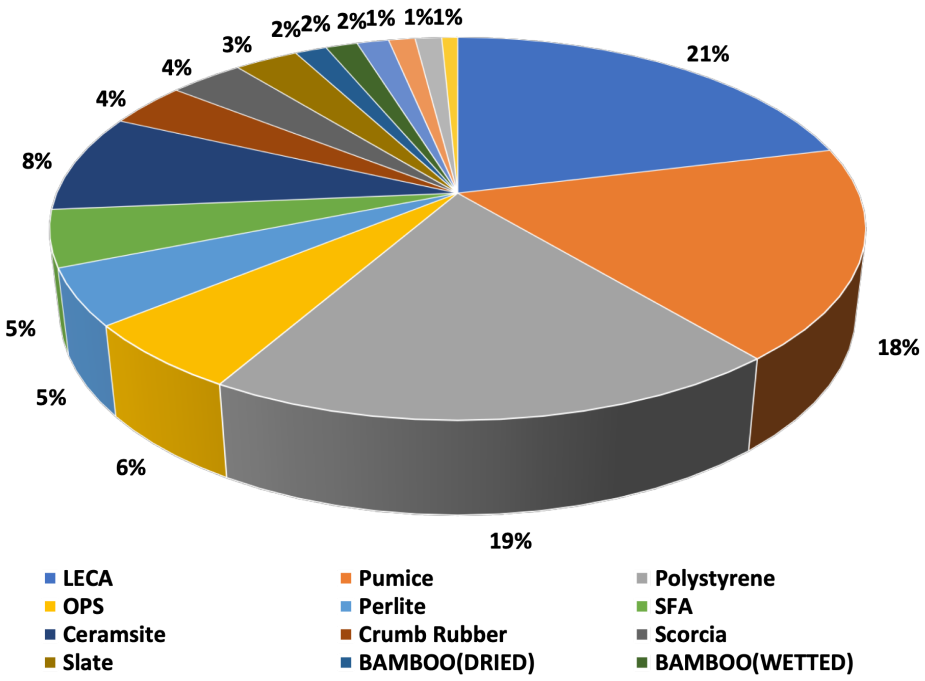

Figure 2: Types of lightweight aggregates used by researchers in the article from which data has been obtained.

Figure 3 illustrates the statistical distribution of compressive strength, tensile strength, and concrete density. The results show that the mean tensile strength is approximately one-tenth of the mean compressive strength (~30 MPa), aligning with established conventions (e.g., ACI codes) [10]. The average density of 1700 kg/m^3 reflects the prevalence of expanded clay aggregate and polystyrene-based concretes since their density lies in this range, validating the dataset's accuracy and reliability for further analysis [11,12].

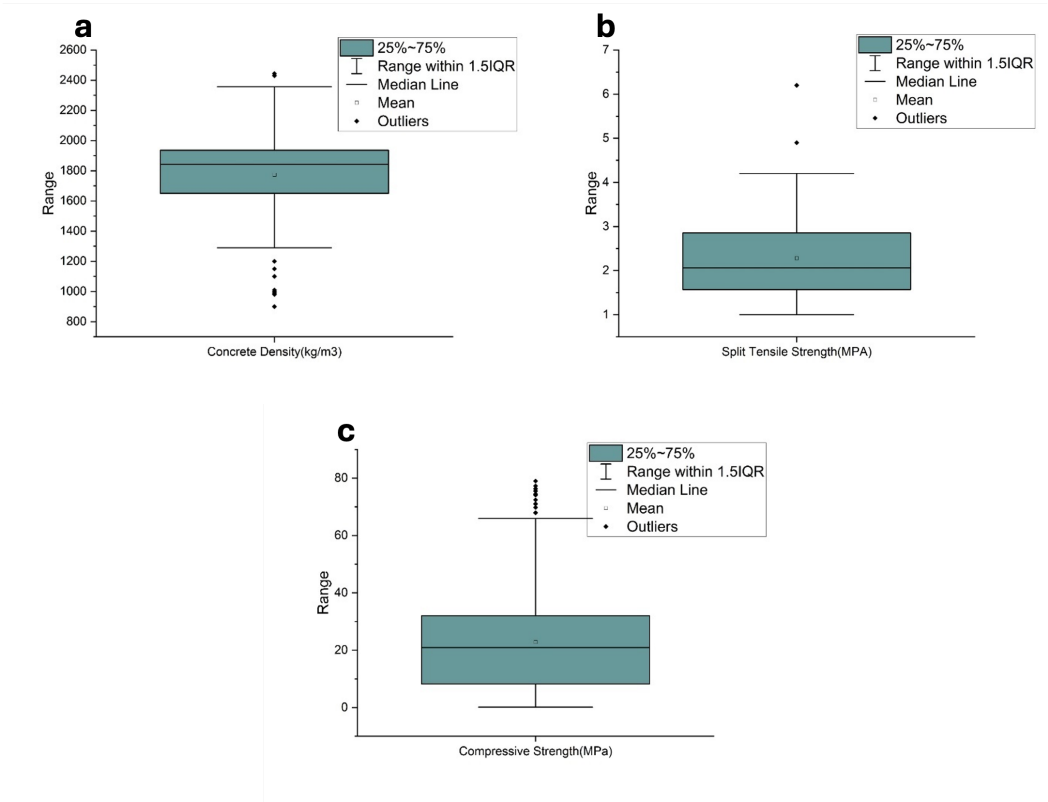

Figure 3: Statistical distributions of the output parameters of the dataset compiled from articles.

2.1. Dataset cleaning and splitting

To maintain accuracy, completeness, consistency, relevance, and validity, the dataset underwent a rigorous cleaning process, a crucial step before applying any machine learning algorithm. This process ensures that the data is processable and enables effective learning for accurate output. Given that the dataset was compiled from X diverse articles sourced from online libraries, there was a high likelihood of human error. However, since most of the dataset was collected by our team, missing values were nonexistent, eliminating the need for removal or imputation. Data cleaning addressed potential issues such as data entry errors, equipment malfunctions, or incomplete surveys. Outliers were identified, and upon examination, most were found to originate from articles published by sources with questionable academic reputations [13,14,15,16,17,18,19]. These outliers were subsequently trimmed to prevent biased analysis and ensure data integrity. Following data cleaning, the refined dataset was split into training (80%) and testing sets (20%). Summary statistics for both are presented in Table 3 and Table 4, providing a comprehensive foundation for reliable model development and evaluation.

Table 3. Summary statistics of dataset set aside for ML model training.

Parameters	Minimum	Maximum	Median	Mode	SD	Type
Cement (kg/m^3)	156	1500	467	480	378.42	I
Fine agg. (kg/m^3)	0	1193	664	0	330.15	I
w/b	0.15	0.80	0.45	0.5	0.08	I
LW agg. (kg/m^3)	23.80	1191	308	37	297.28	I
LW agg. density (kg/m^3)	415	1489	783	575	357.65	I
LW agg. water absorption (%)	0.92	58.30	25.20	40	13.83	I
NW agg. (kg/m^3)	0	1326	0	0	353.98	I
HRWR (% of binder)	0	3	0	0	0.70	I
Curing Time (days)	1	120	28	28	14.27	I
Fly Ash (kg/m^3)	0	540	0	0	117.61	I
Compressive Strength of LW concrete (MPa)	2.03	79	24.58	25	16.68	O
Split Tensile Strength of LW concrete (MPa)	1	7	3.5	3	2	O
Density of LW concrete (kg/m^3)	900	2500	1855	1800	366	O

Table 4. Summary statistics of dataset set aside for ML model testing.

Parameters	Minimum	Maximum	Median	Mode	SD	Type
Cement (kg/m^3)	139	1350	384	450	197.70	I
Fine agg. (kg/m^3)	0	1178	630	0	294.92	I
w/b	0.23	0.8	0.42	0.35	0.07	I
LW agg. (kg/m^3)	0	950	155	0	270.29	I
LW agg. density (kg/m^3)	406	1480	750	610	320.11	I

Parameters	Minimum	Maximum	Median	Mode	SD	Type
LW agg. water absorption (%)	0.92	56	20.5	20.5	13.54	I
NW agg. (kg/m^3)	0	941.2	0	0	282.15	I
HRWR (% of binder)	0	2.5	0.5	0	0.687	I
Curing Time (days)	1	120	28	28	23.4	I
Fly Ash (kg/m^3)	0	540	0	0	111.38	I
Compressive Strength of LW concrete (MPa)	4.28	65.14	27	25	15.54	O
Split Tensile Strength of LW concrete (MPa)	1.2	6.7	3.6	3.1	2.2	O
Density of LW concrete (kg/m^3)	950	2670	1755	1650	354	O

2.2. Dataset cleaning and splitting

Data normalization is a standardization technique for transforming variables to have a common scale. When working with data from different sources or formats, there can be variations in how it is represented, such as differences in units of measurement, data formats, and data structures, making it difficult to compare variables or perform statistical analysis. Data standardization is a crucial step in such cases. The major challenge faced after data collection is processing the raw data to make it compatible with the ML models used. For instance, there was a considerable difference in our dataset between the numerical values of cement, w/c, and normal aggregate used. This difference adversely affected the accuracy of our model. This issue was tackled using the data normalization technique. Data normalization means transforming data into the unit sphere or scaling down the actual values to numerical indexes between 0 and 1. It leads to data cleansing and convergence and significantly enhances the model's efficiency. It also improves data execution by reducing the data set's redundancy. The governing Equation 1 taken into consideration for data normalization is mentioned below, where the normalized value of a certain input variable is a function of the actual, minimum, and maximum values of that variable in the data set. The normalized dataset has been plotted in Figure 4 below

$$y = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \quad (3)$$

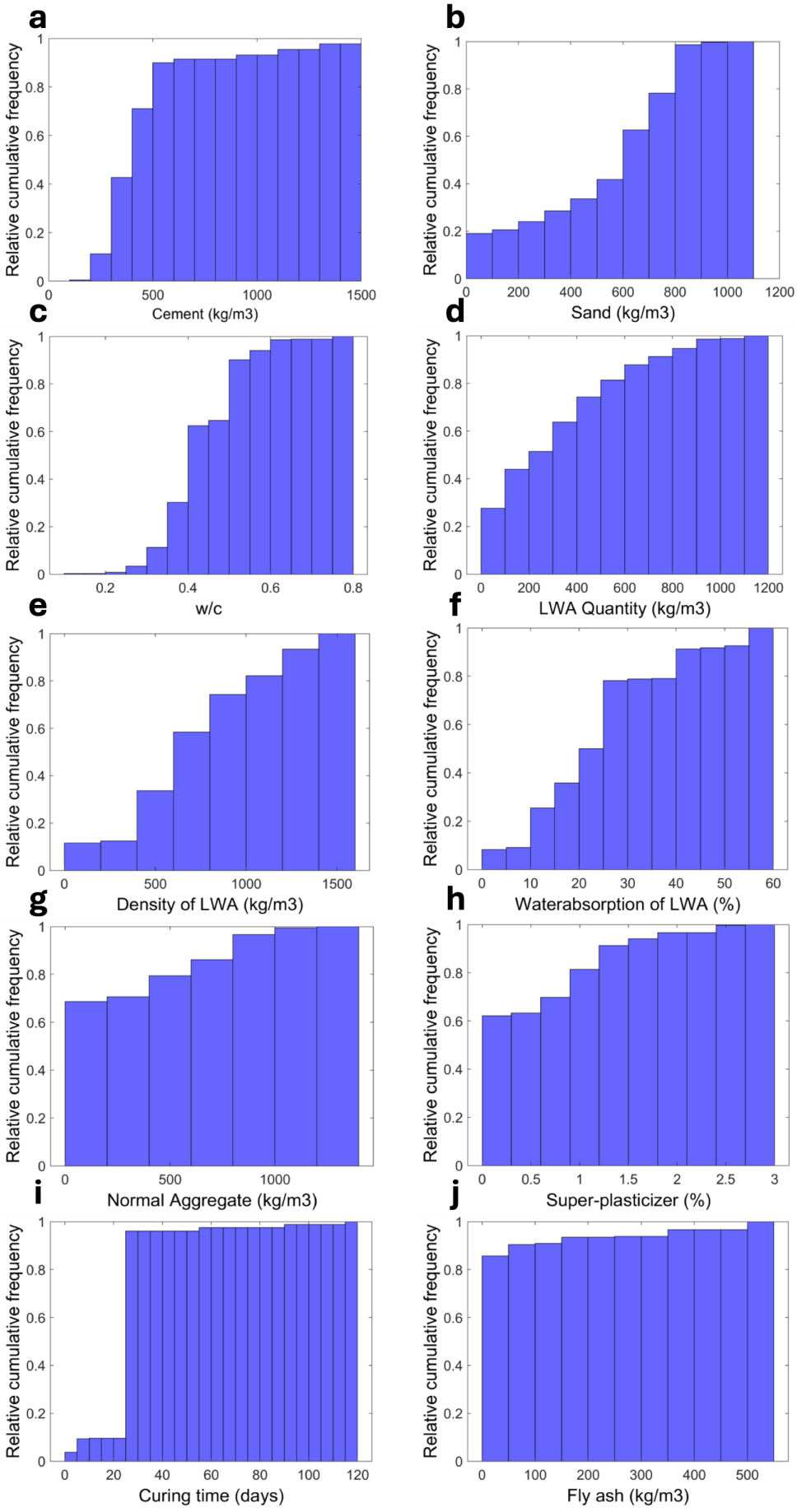

Figure 4: Statistical distribution of the parameters of the dataset compiled after standardization.

2.3. Correlation of input parameters with output parameters

The preprocessing phase proceeded with the plotting of Pearson correlation matrices Figure 5 to elucidate relationships between dependent and independent variables. This comprehensive graph displays pairwise correlations through Pearson correlation coefficients, ranging from -1 to +1. Diagonal entries show perfect correlation (1), while non-diagonal entries exhibit coefficients between -1 and +1, indicating varying degrees of correlation. In the graph template used for this article, blue the color shows the direction of the correlation, and the size of the circle shows magnitude/extent of correlation. Positive coefficients signify direct relationships, and negative coefficients indicate inverse relationships. This analysis, mathematically expressed through the Pearson correlation coefficient (r) Equation 2, provides valuable insights into variable interactions, informing the development of a robust ML model.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{4}$$

The correlation matrix of each input with a particular output has been shown separately for a deeper understanding. Keeping the number of input and output parameters in consideration in a single correlation matrix made the visualization difficult. From the correlation matrices, it is evident that the LWA density was the primary factor controlling compressive strength. As per S.A.Khan et al. [20], the difference between normal-weight concrete and lightweight concrete is that lightweight concrete fails due to the failure of aggregates, not matrix so failure concrete with increased lightweight aggregate density had increased compressive strengths as well. Also, an increase in the water-cement ratio causes a decrease in compressive strength, which can also be seen in the graph. Similarly the total fines content or total normal weight aggregate content has a good positive correlation with split tensile strength and concrete density which is also supported by the literature.

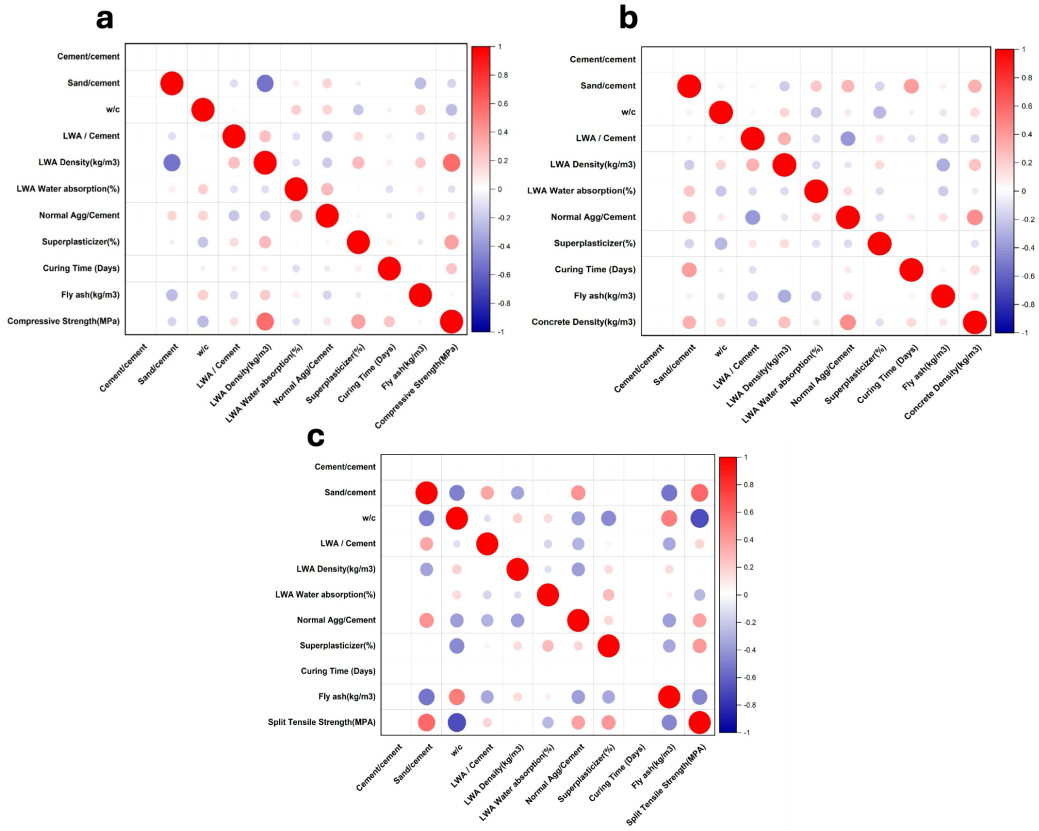

Figure 5: Pearson correlation matrices of input parameters with each output parameter.

Apart from the data cleaning standardization and visualization discussed above, the authors of the report did not require any data augmentation since the number of data points seemed enough for the types of models they intended to train. However, the authors might perform some feature engineering by combining some of the highly correlated input parameters if the statistical performance indicators do not meet their expectations upon training of the model hence changes will be made in the subsequent report

3. Predictive Modeling

This report aims to train, validate, and fine-tune Artificial Neural Networks (ANN), Gaussian Process Regression (GPR), and Decision Tree models on the training data for predicting concrete properties. The objectives are to leverage the unique strengths of each model, with ANN capturing non-linear relationships and patterns, Decision Trees providing interpretable results and feature selection, and GPR quantifying uncertainty and handling sparse data. By combining these models, the challenges in concrete property prediction, including non-linear relationships, variability, and limited data, can be effectively addressed.

Additionally, this report addresses the longstanding controversy surrounding the efficacy and reliability of Machine Learning (ML) and Artificial Intelligence (AI) based models, often labeled as "black boxes" that merely identify patterns without providing meaningful insights. To alleviate concerns regarding overfitting and model interpretability, local explanation techniques, specifically Partial Dependence Plots (PDP) and Shapley Additive Explanations (SHAP), will be employed to decipher the relationships between individual input parameters and the model's output. This approach ensures model interpretability, validity, reliability, and identification of potential biases, ultimately demonstrating the efficacy of ML models in concrete property prediction.

3.1. Artificial neural network

Artificial Neural Networks (ANNs) are complex computational models inspired by biological neural networks. They process input data, generate output, and adapt through backpropagation training. Proven effective in various domains, ANNs excel in classification, regression, forecasting, and clustering tasks. The implemented ANN model architecture has been depicted in Figure 6 below.

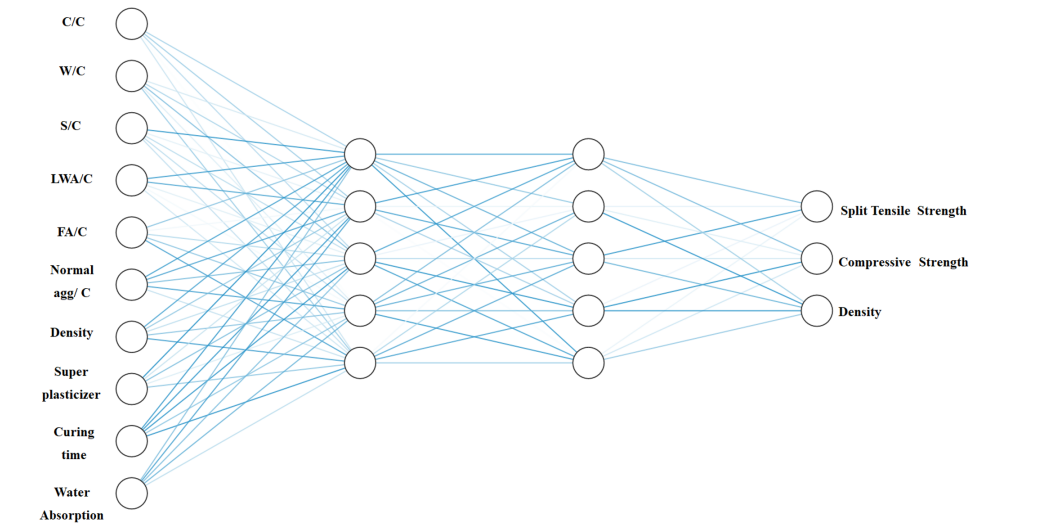

Figure 6: Model architecture of ANN.

3.2. Decision Tree

Decision Trees, a supervised machine learning approach, effectively predicts concrete's mechanical characteristics by modeling complex relationships between input data and output labels. The tree-like structure of Decision Trees provides transparency into prediction outcomes. As a valuable alternative to traditional methods, Decision Trees are a helpful tool for forecasting concrete's mechanical properties, as illustrated in Figure 7.

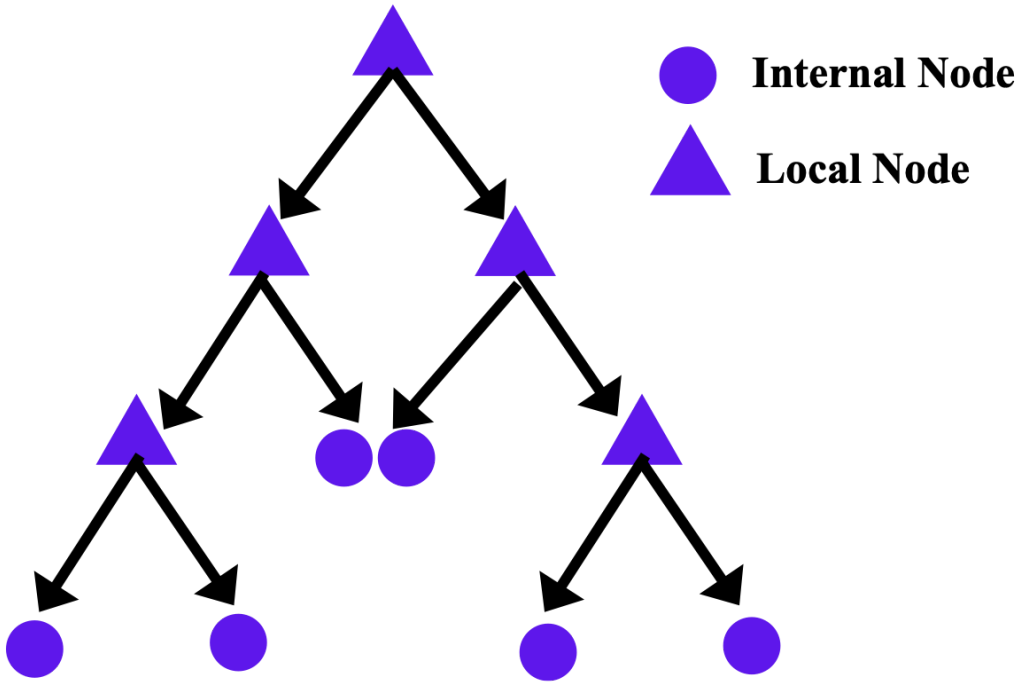

Figure 7: Working mechanism/flowchart of the Decision tree.

3.3. Gaussian Process of Regression

Gaussian Process Regression (GPR) is a supervised machine learning technique using Bayesian inference for predictions. As a non-parametric method [60, 61], GPR excels with limited data, modeling complex relationships between inputs and outputs. Ideal for predicting concrete's mechanical properties, GPR offers accuracy and versatility, making it suitable for diverse applications, as shown in Figure 8.

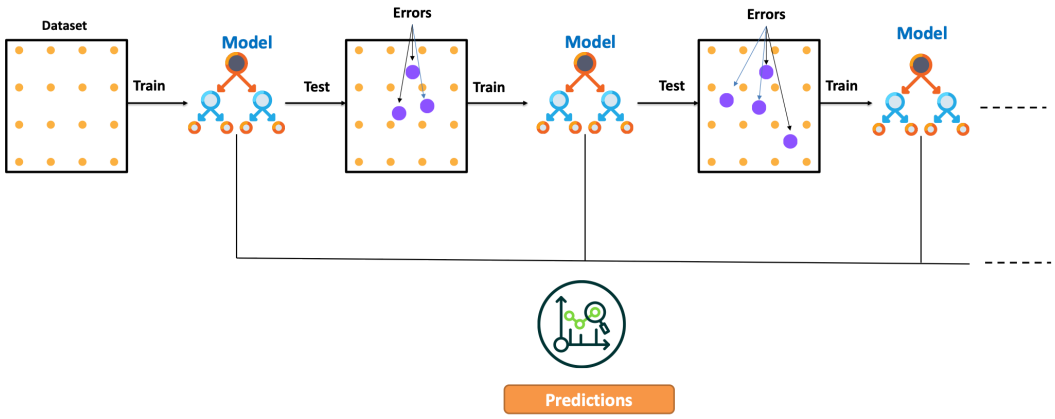

Figure 8: Working mechanism/flowchart of GPR.

In the Figure 9, a complete overview of the whole project is depicted pictorially.

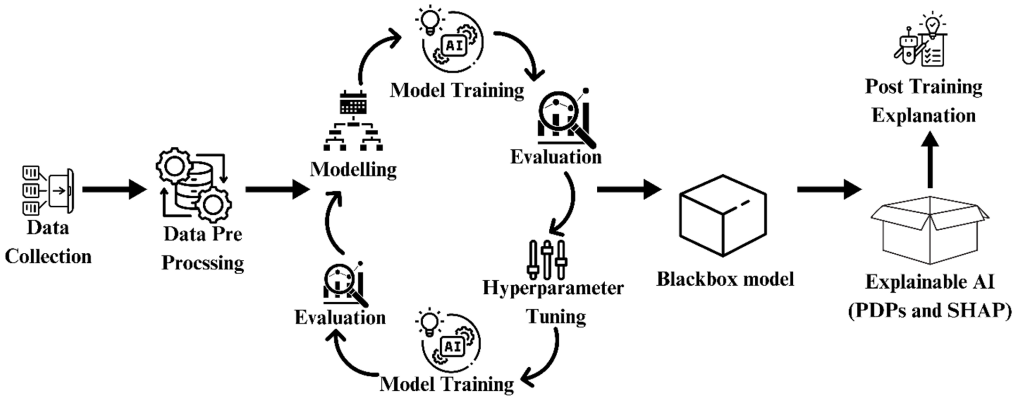

Figure 9: A flowchart explaining the sequence of tasks in the project.

References

1. **Machine Learning-Based Predictive Modeling of Sustainable Lightweight Aggregate Concrete**
Fazal Hussain, Shayan Ali Khan, Rao Arsalan Khushnood, Ameer Hamza, Fazal Rehman
Sustainability (2022-12-30) https://doi.org/gzh6bg
DOI: 10.3390/su15010641

2. **Cleaner Design and Production of Lightweight Aggregates (LWAs) to Use in Agronomic Application**
Carmen Martínez-García, Fernanda Andreola, Isabella Lancellotti, Romina D Farías, Mª Teresa Cotes-Palomino, Luisa Barbieri
Applied Sciences (2021-01-15) https://doi.org/g5vt4j
DOI: 10.3390/app11020800

3. **Concrete strength for fire safety design**
KD Hertz
Magazine of Concrete Research (2005-10) https://doi.org/drrzkq
DOI: 10.1680/macr.2005.57.8.445

4. **Prediction of concrete strength using artificial neural networks**
Seung-Chang Lee
Engineering Structures

5. **The influence of microfillers on enhancement of concrete strength**
A Goldman, A Bentur
Cement and Concrete Research

6. **Concrete strength prediction by means of neural network**
Sergio Lai, Mauro Serra
Construction and Building Materials

7. **Incorporation of Wastes in Lightweight Aggregate of Expanded Clay for Construction Applications**
C Vilarinho
ISWA World Congress 2009 (2009)

8. **Using Mn based on lightweight expanded clay aggregate (LECA) as an original catalyst for the removal of NO2 pollutant in aqueous environment**
Aref Shokri
Surfaces and Interfaces (2020-12) https://doi.org/g8nzzp
DOI: 10.1016/j.surfin.2020.100705

9. **Experimental Study on the Properties of Green Concrete by Replacement of E-Plastic Waste as Aggregate**
Arivalagan S
Procedia Computer Science (2020) https://doi.org/g8nzzn
DOI: 10.1016/j.procs.2020.05.145

10. **Lightweight aggregates for concrete, mortar and grout**
European Committee for Standardization
EN 13055-1 (2016)

11. **Manufacturing of sintered lightweight aggregate using high-carbon fly ash and its effect on the mechanical properties and microstructure of concrete**

Tommy Yiu Lo, Hongzhi Cui, Shazim Ali Memon, Takafumi Noguchi
Journal of Cleaner Production (2016-01) https://doi.org/f77cck
DOI: 10.1016/j.jclepro.2015.07.001

12. **Influence of fly ash and LYTAG lightweight aggregate on concrete**
S Sivakumar, B Kameshwari
International Journal of Applied Engineering Research (2015)

13. **An evaluation of the increased expansion of clay aggregates fired at 1300 °C to maximize lightness for non-structural concrete**
Adalberto Viana Rodrigues, Saulo Roca Bragança
Boletín de la Sociedad Española de Cerámica y Vidrio (2023-01) https://doi.org/g8nzzm
DOI: 10.1016/j.bsecv.2021.11.003

14. **Artificial Lightweight Aggregates Made from Pozzolanic Material: A Review on the Method, Physical and Mechanical Properties, Thermal and Microstructure**
Dickson Ling Chuan Hao, Rafiza Abd Razak, Marwan Kheimi, Zarina Yahya, Mohd Mustafa Al Bakri Abdullah, Dumitru Doru Burduhos Nergis, Hamzah Fansuri, Ratna Ediati, Rosnita Mohamed, Alida Abdullah
Materials (2022-05-31) https://doi.org/grxsg6
DOI: 10.3390/ma15113929 · PMID: 35683229 · PMCID: PMC9181883

15. **Compressive Strength and Durability Properties of Structural Lightweight Concrete with Fine Expanded Glass and/or Clay Aggregates**
Deividas Rumsys, Edmundas Spudulis, Darius Bacinskas, Gintaris Kaklauskas
Materials (2018-11-30) https://doi.org/g8nzzr
DOI: 10.3390/ma11122434 · PMID: 30513643 · PMCID: PMC6317013

16. **Possibilities of determination of thermal conductivity of lightweight concrete with utilization of non stationary hot-wire method**
10th Int. Conf. Slov. Soc. Non-Destructive Test. Appl. Contemp. Non-Destructive Test. Eng.

17. **Influence of mineral additions and different compositional parameters on the shrinkage of structural expanded clay lightweight concrete**
JAlexandre Bogas, Rita Nogueira, Nuno G Almeida
Materials & Design (1980-2015) (2014-04) https://doi.org/gpptkk
DOI: 10.1016/j.matdes.2013.12.013

18. **The Physical and Mechanical Properties of Autoclaved Aerated Concrete (AAC) with Recycled AAC as a Partial Replacement for Sand**
Abdul Rahman Rafiza, Ahmad Fazlizan, Atthakorn Thongtha, Nilofar Asim, Md Saleh Noorashikin
Buildings (2022-01-07) https://doi.org/g6b6hh
DOI: 10.3390/buildings12010060

19. **Optimum Bloating-Activation Zone of Artificial Lightweight Aggregate by Dynamic Parameters**
Young Min Wie, Ki Gang Lee
Materials (2019-01-15) https://doi.org/g8nzzs
DOI: 10.3390/ma12020267 · PMID: 30650611 · PMCID: PMC6356517

20. **Feasibility Study of Expanded Clay Aggregate Lightweight Concrete for Nonstructural Applications**
Shayan Ali Khan, Fazal Hussain, Rao Arsalan Khushnood, Hassan Amjad, Farhan Ahmad
Advances in Civil Engineering (2024-02-29) https://doi.org/g8nzzq

DOI: 10.1155/2024/8263261