Integrating Machine Learning and Symbolic Regression for Transparent Prediction of Ultra-High Performance Concrete Strength

Muhammad Zikria Luqman[1*] & Muhammad Ahmad Afzal[1]

[1] Affiliation 1; Department of Civil Engineering GIK Institute, Topi 23640, KP, Pakistan
* Correspondence: gcv2466@giki.edu.pk

Abstract

Ultra-High-Performance Concrete (UHPC) is characterized in modern high performance building materials as being extremely heavy-duty safe and optimized mix design. However, the interdependence of its ingredients is very nonlinear, which makes the accurate prediction of compressive strength a complicated task. The current research presents a combination of Machine Learning (ML) algorithms with a Symbolic Regression as a means to predict the compressive strength of UHPC based on 810 samples of a reliable open dataset. To increase generalizability, 5-fold stratified cross-validation was used to train 7 ML algorithms including K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Neural Network and Linear Regression. To tune each model hyperparameters and assess both performance based on MAE, RMSE and R2 measure, Orange Data Mining software was used. To improve the explanation, the four models that had the highest accuracy, which is Gradient Boosting, AdaBoost, Neural Network, and Random Forest, were subsequently understood by means of Multi Expression Programming X (MEPX), to come up with explanatory equations. These equations as well as the equation of the original dataset have been tested through mathematical simplicity and practical engineering application. The results revealed that interpretability and predictive soundness could be presented in symbolic models, presenting a viable choice of engineering use. The study thus combines information-based intelligence and clear decision-making over the optimization of UHPC mix and makes it easier, cheaper, and more understandable to design materials in structural engineering disciplines.

Keywords: Ultra-High-Performance Concrete, Machine Learning, Symbolic Regression, MEPX, Compressive Strength Prediction, Interpretable Models.

1. Introduction

The Ultra-High-Performance Concrete (UHPC) is one of the paradigmatic innovations in the field of concrete-materials which is enabled by the mechanical performance never achieved before, its increased durability and the long-life of the structure due to the enhanced resilience that allows maintaining the structural integrity in the case of a failure [1]. UHPC uses a propriety mixture design to regularly attain a compressive strength greater than 150 MPa with values varying to as high as 500 MPa; at the same time it also has better tensile strength and fracture toughness due to the internal dense matrix and

https://doi.org/10.3390/xxxxx

overall microstructure design. These properties help to be used in harsh scenarios such as long-span bridges, seismic retrofit, high-rise constructions, and sensitive military facilities [2].

Although these are its benefits, the realisation of the high performance of UHPC requires an accurate engineering of the proportions of the components, interactive effect and the curing conditions [3]. The contents that should be orchestrally controlled in terms of best results include cement, silica fume, nano-silica, quartz powder, limestone powder, slag, and high-performance fibers. The resulting mixture design, nevertheless, is principally empirical, and the trials and error still come to play [4]. These types of approach are long and resource consuming in addition to not taking into full consideration the nonlinear, interactive nature of phenomena that define UHPC behaviour [5].

The first one, thus, lies in the complexity of UHPC material system itself. Every one of those variables such as dosage of nano-silica, curing temperature or volume of fibre have effect on hydration kinetics, evolution of microstructure and ultimate compressive strength [6]. Such effects can seldom be additive, and act either synergistically or antagonistically, which does not permit their isolation using standard multiple regression procedure. In-order to achieve accurate prediction of UHPC performance complex modelling tools that can help in recognising multivariate, nonlinear relationships must be deployed [7].

Current studies on civil engineering make machine learning (ML) a powerful alternative to traditional material modeling methodologies [8]. The ensemble-based methods, including Random Forest and Gradient Boosting, and neural-network models, can condense the knowledge in the complex datasets without pre-conceived priori mathematical definitions and reveal the latent patterns and can be effectively used across a broad range of input attributes with a high level of predictive accuracy [9]. However, this method regularly faces a criticism regarding the so-called black-box behavior: they provide little understanding of the mechanism behind inferences made, as well as the relative importance of each of the explanatory variables [10].

In a bid to address this trade-off on the predictive accuracy against interpretability, symbolic regression has emerged [11]. Instead of just curving the data, symbolic regression tries to find analytical formulae that summarise the input-output relation [12]. Genetic programming has been used to develop new tools like Multi Expression Programming X (MEPX) that can be used to develop new mathematical expressions based on prior ML inferences, in order to produce interpretable and closed-form solutions [13]. Symbolic equations that are generated during symbolic regression offer a desirable addition to black-box ML models in engineering practice, where clarity, simplicity, and physical relevance are still core requirements to be adopted [14].

The current research proposes an elaborate data-driven model of modeling the compressive strength of ultra-high-performance concrete (UHPC). The data set involves 810 distinct mixture designs all characterized by 14 input parameters, which are contents of a binder, pozzolanic substances, curing time, water level, temperature, super plasticizer concentration. Compressive strength (MPa) is the target variable, the most important performance indicator in the UHPC applications [15]. The training and validation of machine learning (ML) models are ensured by using cross-validation methods of robust performance [16].

Besides prediction ability, interpretability is also emphasized in the study due to symbolic regression being used on the two models that had the best performance [17]. The objectives are to develop some easy equation, which can be easily calculated by engineers in order to derive the strength of UHPC without going through a lot of calculations by simulation [18]. Such equations are also seen to explicate how the various variables interact in the process of making the determination of the development of strength hence either strengthening or refuting any known material science [19].

Its resultant hybrid framework combining data-driven prediction with physical understanding provides practical mixture optimization tools, underpins performance-based concrete design, and contributes in general to the effort to bring UHPC to normal practice in terms of accessibility, cost, and reliability as a structural material [20].

2. Methodology

2.1 Dataset Description and Preprocessing

The current study uses high quality experimental data on the Ultra-High-Performance Concrete (UHPC) stored on Mendeley Data (doi:10.17632/85r7bh4zsz.1). It includes 810 mixtures of design samples with 14 main characteristics which are: cement, slag, silica fume, limestone powder, quartz powder, fly ash, nano-silica, aggregate, water, fiber, superplasticizer, curing temperature and age and its compressive strength is the output [15]. These characteristics expose the complex dependency of material composition, admixtures and curing conditions on the performance of UHPC. The dataset was tested on completeness before model building, and it was found that there is no missing value. Appropriate scaling routines were utilised: scaling to similar magnitude values as an input variable was applied to neural network-based model, but raw values were applied to tree-based model, which was more resistant to scaling. These preprocessing processes were necessary to guarantee the integrity of the data set and make it viable in machine learning enabled determination of compressive strength.

Table 1: Specifications Table [15]

Field	**Description**
Subject	Civil and Structural Engineering
Specific subject area	Mix design and compressive strength of Ultra-High-Performance Concrete
Data format	Raw and analysed
Parameters for data collection	Cementitious materials, admixtures, aggregates, curing temperature and age
Description of data collection	810 laboratory-tested UHPC mixes with varied proportions and curing regimes
Data accessibility	Repository Name: Mendeley Data DOI: 10.17632/85r7bh4zsz.1

2.2 Machine Learning Model Development and Cross-Validation

Here, seven machine learning algorithms were optimized and tested with regards to their individual capability to represent features of target datasets. These models were applied in the Orange Data Mining, which is a visual programming environment that is easy and convenient to experiment and create prototypes. Selection criteria prioritized algorithm diversity, spanning ensemble boosting techniques (AdaBoost and Gradient Boosting), instance-based learning (k-Nearest Neighbors), linear approximations (Linear Regression), and neural architectures (Multilayer Perceptron). Each model was set to be used with an apt hyperparameter configuration; e.g. the AdaBoost model was set to 50 estimators, learning rate was 1.0, and linear loss. Gradient Boosting (XGBoost) used 100 trees, a learning rate of 0.3, and a depth limit of 6. Random Forest used 40 trees and splits attributes at every node. The employed Neural Network incorporated 75 hidden neurons with the provision of tanh function activation and optimized using the L-BFGS-B solver. Decision Tree and k-Nearest Neighbors had default parameters left but minor parameters were changed. Prior to training all models, a 5-fold stratified cross-validation structure was employed in order to ensure a statistically sound way of making reasonable comparisons between models, as well as minimising overfitting, with the target variable (continuous) being evenly distributed among the folds through quantile based stratification. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2) were computed for each folded dataset and averaged across all folds to assess generalization performance. Such a consistent methodology allowed every algorithm to be tested in similar albeit strict conditions, thus enabling the performance to be compared objectively prior to the further symbolic regression modeling.

2.3 Symbolic Regression Using MEPX

In the study research, the idea of enhancing intuitive interpretation and reducing the gap between the data-driven forecasting and in-house engineering knowledge based on Multi Expression Programming X (MEPX) symbolic regression was pursued. In contrast to the traditional regression methods, symbolic regression does not require the prior knowledge of a certain model structure; rather, it runs an iterative process of building mathematical formulae that optimally explain the relation between input features and target response. The project generated symbolic representations relating to the four best models of machine-learning by using predictions of contemporary models on the earlier-used feature inputs to the real-world input data. The main purpose was to humanize, internal logic of such high-performing models through the form of equations. MEPX was set up to operate on continuous inputs and address a symbolic regression issue where Mean Absolute Error was employed as the key measure. The evolutionary engine used two subpopulations of 200 individuals and had 400 generations per run, a maximum code length of 30, crossover probability of 0.9 (uniform crossover) and the mutation probability of 0.01. Heterogeneous set of mathematical operators, such as elementary arithmetic, power, square root, logarithmic, exponential, and trigonometric functions was provided with a view to encouragement of expressive formulae. Auto-generated constants were limited to the set 0-1 and evolution would take place only within this domain. Internal validation set was applied to validate training procedures and the whole process was repeated 20 times in independent runs, with random seed initialization value of 0, to ensure reproducibility. Such an approach to methodology produced symbolic expressions that reflect the predictive behavior of complex machine-learning models in a very close manner, thereby providing a practical value in engineering within environments where transparency and interpretability are required.

2.4 Comparative Evaluation and Interpretability Assessment

The last stage of the investigation implied a combined analysis of predictive-performance assessment and interpretability of machine-learning models and symbolic-regression output. The performance of each of the machine-learning models was evaluated using the 5-fold stratified cross-validation and evaluated using the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and coefficient of determination (R2) average over the folds to make the performance reliable. These four models, which had the most significant mean performances, were then used in symbolic-regression analysis with Multi Expression Programming X (MEPX). Further, symbolic regression was directly applied to the original data where the actual values of the target variable were used and the final output is five symbolic expressions, one based on the original data and four based on the chosen machine learning model outputs. These expressions were compared not just in terms of predictive accuracy using the same statistical measures, but were also compared in terms of mathematical complexity (number of operators, depth, and structure) and in terms of their engineering interpretability, which involved the question of whether the equations fulfilled physically meaningful expressions and were dimensionally consistent, and thus were amenable to practical engineering use. This dual-level assessment scheme provided a glimpse into the predictive power, as well as the explicative clarity of the models thus providing the needed equilibrium between statistical performance and explanatory clarity of structural engineering design processes.

2.5 Final Workflow and Implementation Framework

The current research applies a systematic approach, which combines the data preprocessing, creation of the model, assessment, and obtaining the symbolic expressions. The preprocessing of data started by acquiring the UHPC data that was then cleaned, validated, and scaled selectively to make the data convenient to subsequent learning algorithms. The modelling was carried out in Orange Data Mining using visual pipelines to train, cross-validate and test results on the same set of criteria. The quantitative measures in terms of performance were used to determine the four best models that possess maximum predictive accuracy. Parallel to that, we carried out symbolic regression with MEPX on the original data as well as the predictions produced by each of these four selected models. Individual MEPX configurations were optimised with standard evolutionary tuning settings (crossover likelihood, mutation probability and populace size) to produce resilience and repeatability. The ensuing symbolic models gave meaningful mathematical formulae representing the nonlinear patterns behind. A second evaluation offered a measure of the physical pertinence, and feasibility of realization of this symbolic equations. Altogether, the hybrid pipeline, where statistical learning and symbolic modeling are integrated, provides an extremely exact predictive system and a comprehensible, portrayable set of tools, which could be used to optimize UHPC mix design in actual practice of engineering.

3. Results and Discussions

3.1 Summary Statistics

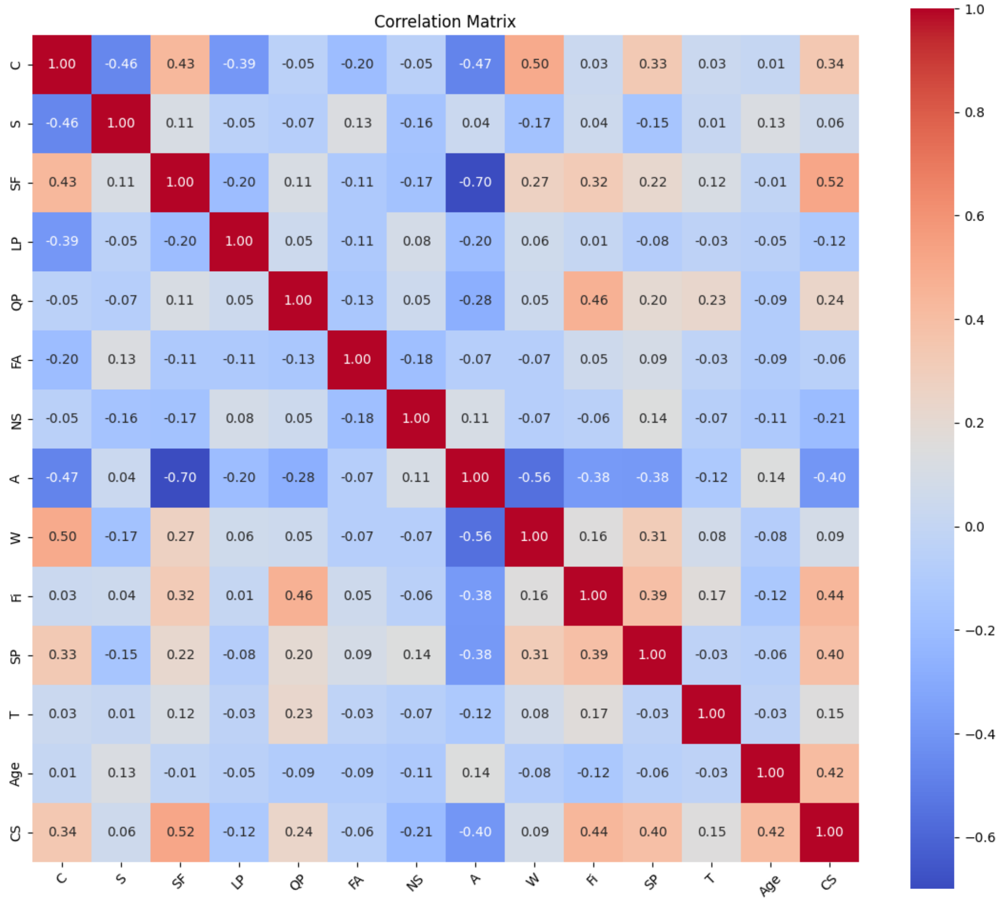
Table 2 giving the summary statistics renders important observations related to the composition of UHPC mixtures. It has a relatively high mean of 737.91 kg/m 3 with a large variation that ranges to 270 to 1251.2 kg /m 3 in effect, portraying highly variable mix designs. The skewness value (or (minus) 0.23) is near to zero, hence it shows a symmatric distribution which is friendly towards regression modeling. The median value (144 kg/m 3) is larger than the mean value (136.99 kg/m 3), indicating that the distribution of silica fume (SF) is skewed on the left-hand side; this distribution indicates that the mixtures that exclude SF entirely are not present. The nano-silica (NS) has a low value of the mean (3.64 kg/m 3), and large skewness (2.53) that underlines the low yet potentially significant application. Fibers (Fi) provide a mean of 56.04 kg/m 3, and a median of zero, which means that half of the mixtures will not have any fibers. The compressive strength (CS) is 123.13 MPa on average with low skewness (0.002), which means that the target variable has a normal distribution and is subject to regression analysis.

Table 2: Summary Statistics

Feature	Mean	Standard Deviation	Standard Error	Sample Variance	Skewness	Kurtosis	Range	Min	Max
C	737.9	173.5	6.1	30087.4	-0.2	-0.1	981.2	270.0	1251.2
S	25.2	74.4	2.6	5530.2	3.0	8.2	375.0	0.0	375.0
SF	137.0	104.1	3.7	10846.1	0.3	-0.6	433.7	0.0	433.7
LP	41.9	133.1	4.7	17724.0	4.8	28.3	1058.2	0.0	1058.2
QP	33.3	79.7	2.8	6347.9	2.3	4.2	397.0	0.0	397.0
FA	26.3	67.5	2.4	4551.1	2.5	5.3	356.0	0.0	356.0
NS	3.6	7.8	0.3	60.5	2.5	6.7	47.5	0.0	47.5
A	1150.1	312.2	11.0	97438.9	0.2	-0.2	1584.2	407.8	1992.0
W	179.9	25.6	0.9	653.7	0.6	1.7	182.6	90.0	272.6
Fi	56.0	75.2	2.6	5659.6	0.8	-1.0	234.0	0.0	234.0
SP	30.0	14.0	0.5	195.8	-0.2	-1.1	55.9	1.1	57.0
T	23.9	16.2	0.6	262.8	9.1	91.7	190.0	20.0	210.0
Age	37.1	53.1	1.9	2821.3	3.8	18.6	364.0	1.0	365.0
CS	123.1	40.2	1.4	1619.2	0.0	-0.6	192.0	28.5	220.5

3.2 Correlation Matrix

The graphical representation of the correlation matrix between the variables under consideration and the compressive strength (CS) are shown in Figure 1. Cement (C), silica fume (SF), and fibers (Fi) have a very positive correlation with CS and this tendency is in line with the most popular assumption that the increase in concrete strength is the predictable consequence of increasing binder and fiber percentages. However, the water (W), on the other hand, has a negative correlation, and this has actually been observed over the years that excess water has a tendency of weakening the concrete matrix. The other input variable, nano-silica (NS), has only weak correlation with CS; the same might be explained by context-dependent effects, or the necessity of having some interaction terms to better describe the effect of NS. Also, the matrix indicates some areas of multicollinearity, in particular, between cement and silica fume, which is to be considered in the course of a feature selection in order to reduce the threat of overfitting.



	C	S	SF	LP	QP	FA	NS	A	W	Fi	SP	T	Age	CS
C	1.00	-0.46	0.43	-0.39	-0.05	-0.20	-0.05	-0.47	0.50	0.03	0.33	0.03	0.01	0.34
S	-0.46	1.00	0.11	-0.05	-0.07	0.13	-0.16	0.04	-0.17	0.04	-0.15	0.01	0.13	0.06
SF	0.43	0.11	1.00	-0.20	0.11	-0.11	-0.17	-0.70	0.27	0.32	0.22	0.12	-0.01	0.52
LP	-0.39	-0.05	-0.20	1.00	0.05	-0.11	0.08	-0.20	0.06	0.01	-0.08	-0.03	-0.05	-0.12
QP	-0.05	-0.07	0.11	0.05	1.00	-0.13	0.05	-0.28	0.05	0.46	0.20	0.23	-0.09	0.24
FA	-0.20	0.13	-0.11	-0.11	-0.13	1.00	-0.18	-0.07	-0.07	0.05	0.09	-0.03	-0.09	-0.06
NS	-0.05	-0.16	-0.17	0.08	0.05	-0.18	1.00	0.11	-0.07	-0.06	0.14	-0.07	-0.11	-0.21
A	-0.47	0.04	-0.70	-0.20	-0.28	-0.07	0.11	1.00	-0.56	-0.38	-0.38	-0.12	0.14	-0.40
W	0.50	-0.17	0.27	0.06	0.05	-0.07	-0.07	-0.56	1.00	0.16	0.31	0.08	-0.08	0.09
Fi	0.03	0.04	0.32	0.01	0.46	0.05	-0.06	-0.38	0.16	1.00	0.39	0.17	-0.12	0.44
SP	0.33	-0.15	0.22	-0.08	0.20	0.09	0.14	-0.38	0.31	0.39	1.00	-0.03	-0.06	0.40
T	0.03	0.01	0.12	-0.03	0.23	-0.03	-0.07	-0.12	0.08	0.17	-0.03	1.00	-0.03	0.15
Age	0.01	0.13	-0.01	-0.05	-0.09	-0.09	-0.11	0.14	-0.08	-0.12	-0.06	-0.03	1.00	0.42
CS	0.34	0.06	0.52	-0.12	0.24	-0.06	-0.21	-0.40	0.09	0.44	0.40	0.15	0.42	1.00

Figure 1: Correlation Matrix

3.3 Distribution Plots

3.3.1 Analysis of Cement (C) Distribution Plots

A histogram of cement concentration of the tested UHPC mixes is shown in Figure 2. When visualized, a bimodal pattern emerges, with peaks of density around 600 and 900 kg/m 3, and two major cement-contents strategies can be observed, a moderate range and a high range. The two-peak nature is conformed by both the histogram and the Kernel Density Estimation (KDE) graph. The respective boxplot is symmetrically dispersed, with the median around 770 kg/m 3 and there is no outlier. The violin plot also requires mapping the concentration of the density at the two points of peak; besides the horizontal span that is observed in the scatter plot reveals a steady use of cement throughout the samples. The QQ plot indicates that there is a gradual loss of normalcy at the extreme ends and it is therefore possible that some statistical modeling methods have errors in calculation, in case the cement concentration variable is transformed. Drawn together, the plots notably stress the necessity to take into consideration the bimodality of cement concentration when exploring the connection between the latter and the compressive strength too since the modeling might demand distinguishing between the two patterns of mix design.

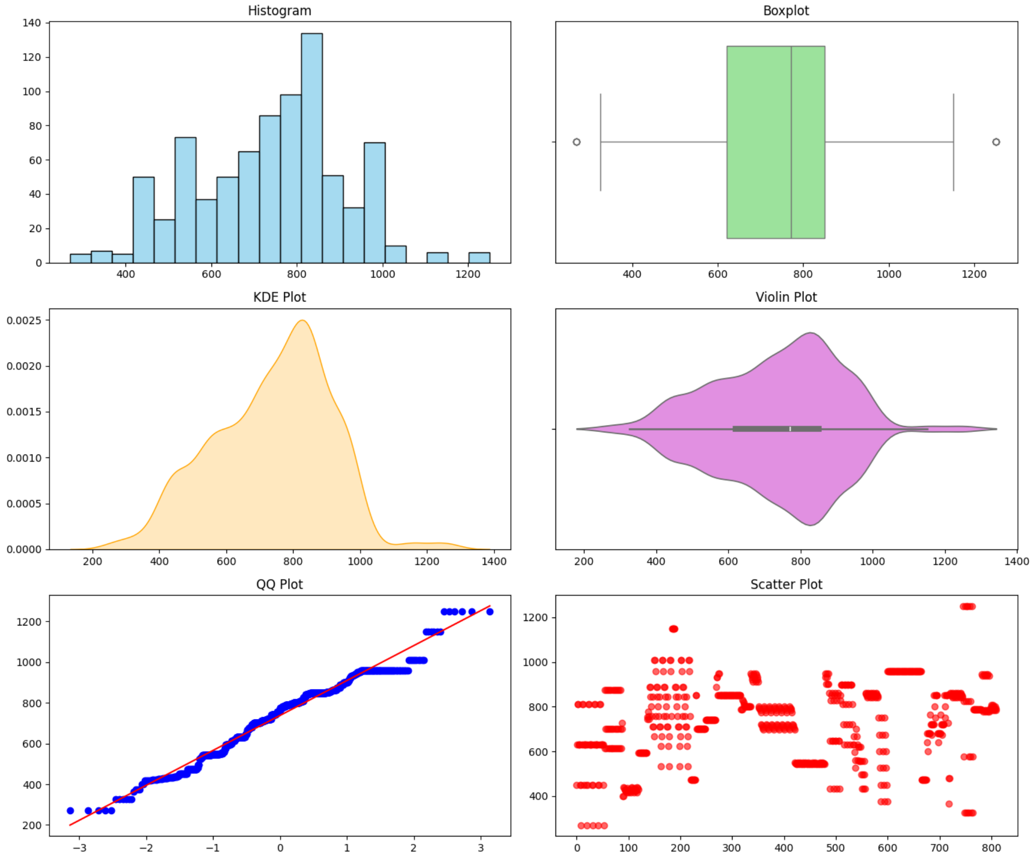


Figure 2: Cement (C) Distribution Plots

3.3.2 Distribution Plots for Slag (S)

As is shown in Figure 3, the distribution plots of S in the Ultra-High-Performance Concrete (UHPC) dataset summarize its statistical character with relative brevity. Stringent drop is visible on the right side with the dominance of zeros and within a relatively small number of bars the value of 375 is attained. The given trend and the related skewness of 3.0176 point to the right-skewed nature of the distribution, as the mean, median, and mode equal 0.0; there is no slag at most records. The KDE plot confirms this arrangement as a narrow peak centred on zero that falls sharply and that could be indicative of a distribution that is skewed to the right and that is leptokurtic (kurtosis 8.2266). Violin plot provides a thick cluster at 0 and a low whip making the violin plot to differ with those that have a standard deviation of 74.3655, and the outliers above the upper whisker, which indicate the boxplot. The Q-Q plot shows evident distortions to normality, the Q-Q plot slants positively at the larger quantiles, as a result of the heavy tail. As the scatter plot shows, there is a zero-centric cluster to which the points are scattered and reach up to 375 indicating the erratic, zero-rich character of the distribution. All these plots point to the fact that slag is used sporadically and it is difficult to predict its contribution to the strength modeling of UHPC.

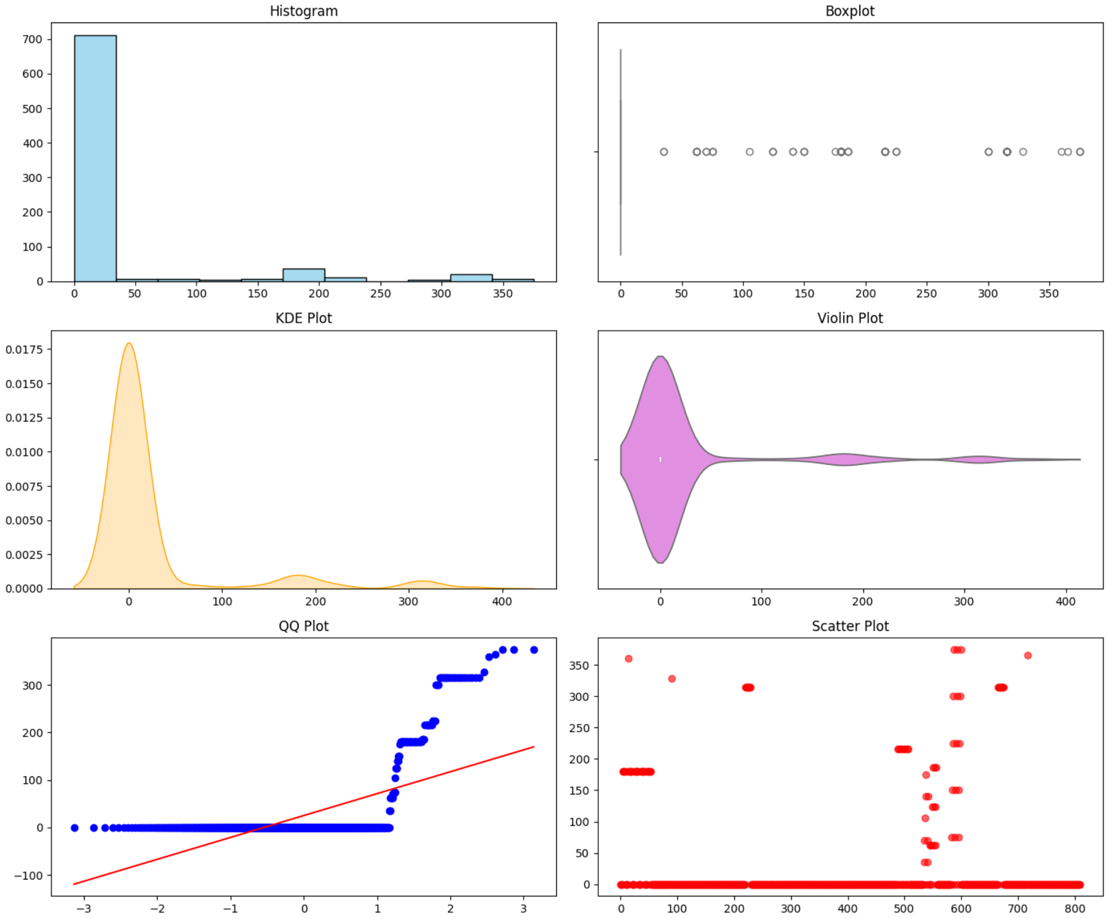


Figure 3: Distribution Plots for Slag (S)

3.3.3 Distribution Plots for Silica Fume (SF)

In Figure 4, the silica fume distribution is quite slightly right-skewed, as the skewness lies in 0.259. There is quite a balanced distribution as the median of 144.0 is not very far off as compared to the sample mean 136.987272 kg/m 3. However, a value of 0.0 kg/m 3 in a mode means that there is a high percentage of the observations that do not include any silica fume. When coupled with the median equal to 144.0 kg/m 3 it would seem that there was a possible two-peaked distribution, one centred at zero and another about 144.0 kg/m 3. The negative kurtosis of -0.5967 also portrays that it is flatter in nature and has the possibility of more than one mode, whereas the standard deviation of 104.1446kg/m 3 maintains moderate dispersion within a maximum range of 0.0 to 433.7kg/m 3. Two clear clusters can be identified one at zero silica fume and one near 144.0 kg/m 3 which implies two practices; inclusion or no inclusion of silica fume and good inclusion.

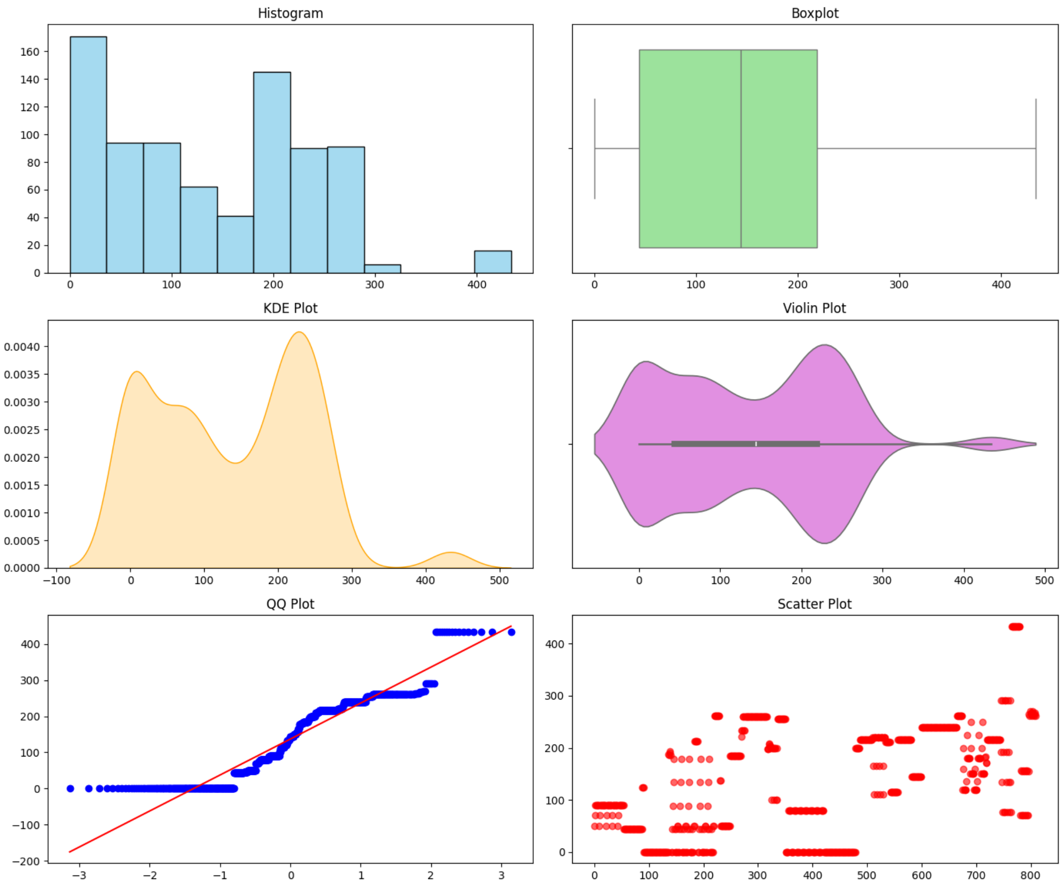


Figure 4: Distribution Plots for Silica Fume (SF)

3.3.4 Distribution Plots for Limestone Powder (LP)

Analysis of Figure 5 indicates that limestone powder is extreme in distribution all to the right, which was evident by skewness coefficient of 4.7579. Both the median and the mode value are at 0.0 kg/m 3, which means that the largest number of observations do not have this component in them, whereas the mean value at 41.9295 kg/m 3 is a result of an occurrence of a low number of samples with abnormally high values, namely, a maximum of 1058.2 kg/m 3. The value of kurtosis is far too high (28.3356), indicating a very sharp peak at the mark of zero and a long, thin right tail; in addition, the standard deviation of 133.1315 kg/m 3 depicts significant freedom in the distribution of the non-zero values. Such distribution would have a very sharp peak at zero, but quickly diminish, with a thin, long tail in the higher parts of the concentration, which further demonstrates that even though the presence of limestone powder can, and often does occur in UHPC, it is still a very low possibility with a very low probability of occurring.

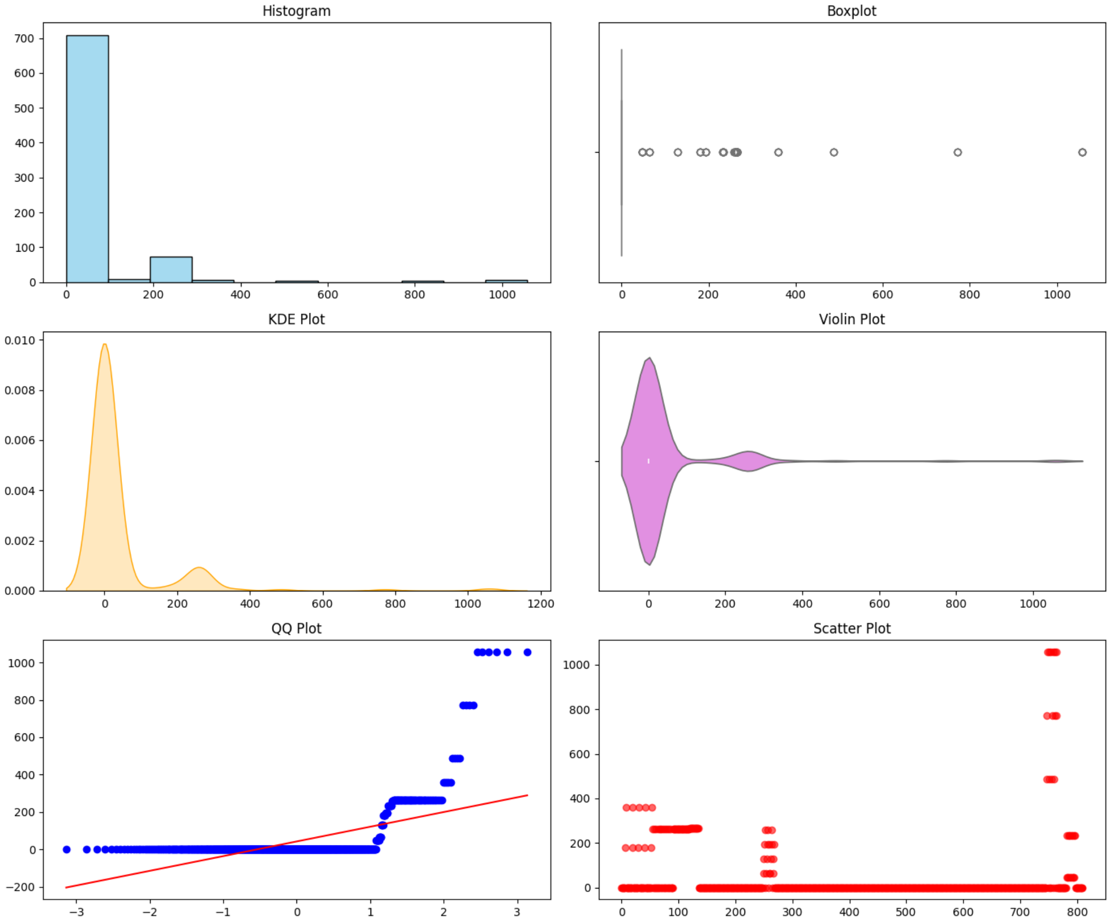


Figure 5: Distribution Plots for Limestone Powder (LP)

3.3.5 Distribution Plots for Quartz Powder (QP)

The rightward deviation is observable in Figure 6 and quartz powder shows 2.2829 of skewness. The median as well as the mode is 0.0 kg/m 3 indicating that most of the samples do not contain quartz powder. On the other hand, the average of 33.271 kg/m 3 portrays occasional use though not above 397.0 kg/m 3. The positive value of the kurtosis (4.2442) points to the peaked distribution with an unusually long right tail and the 79.6739 kg/m 3 standard deviation moderately represents the dispersion between non-zero values. Therefore, the customary plot of the distribution would most probably have a stark high at zero, slowly dropping off into lower points, and extending out into a tail to greater concentrations indicating the selective but in some cases, substantial addition of quartz powder to UHPC.

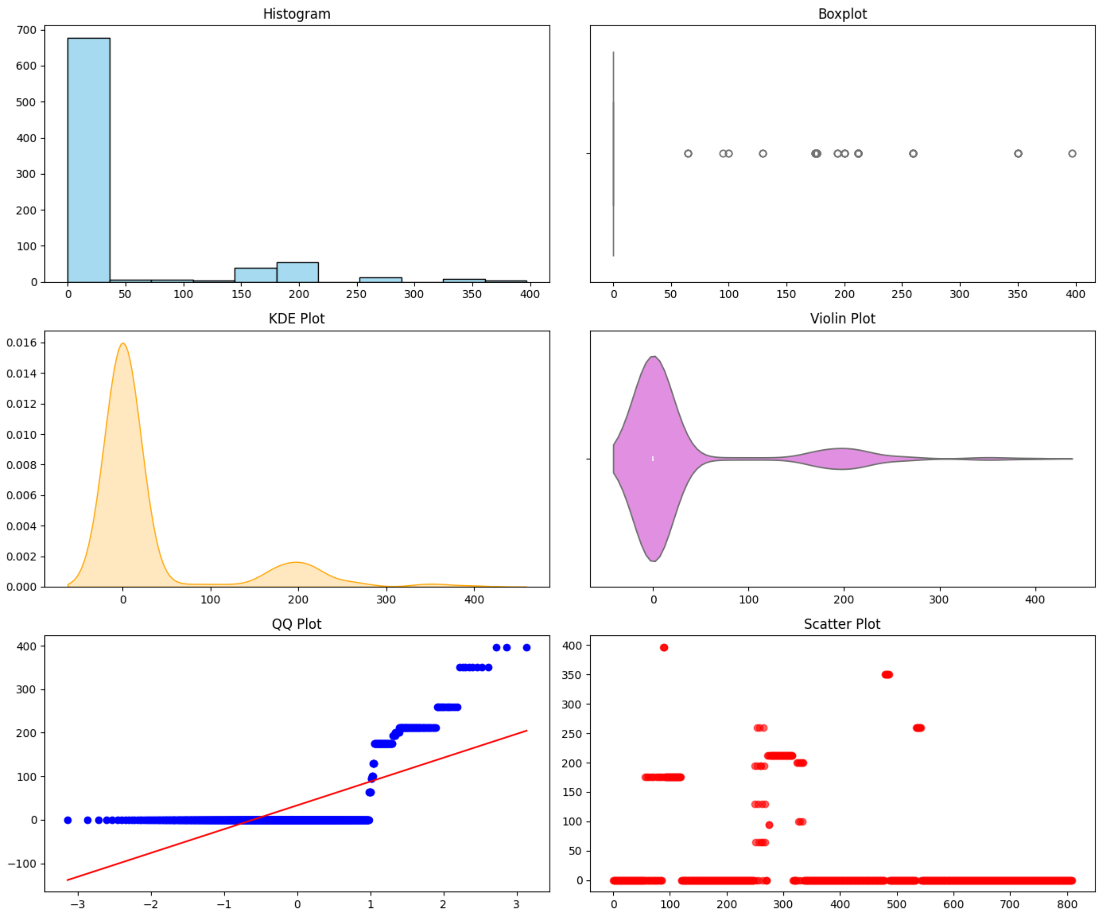


Figure 6: Distribution Plots for Quartz Powder (QP)

3.3.6 Distribution Plots for Fly Ash (FA)

It is observed that the fly ash is right skewed as shown in figure 7; it is also noting that skewness is 2.4922. Its median and mode are 0.0 kg/m3 and this means that in the majority of samples it is absent, whereas the mean of 26.2649 and a standard deviation of 67.4617 kg/m 3 means that in some cases the level of fly ash can be as high as 356.0 kg/m 3. A value of kurtosis of 5.3377 means that it is extremely raised at 0, and the right side of the graph is long and probably covers outliers. The distribution plot would probably indicate one thick spike centered at zero, a steep decline, and an elongated tail at the right end, which represents significantly rare yet considerable application of fly ash in some UHPC products.

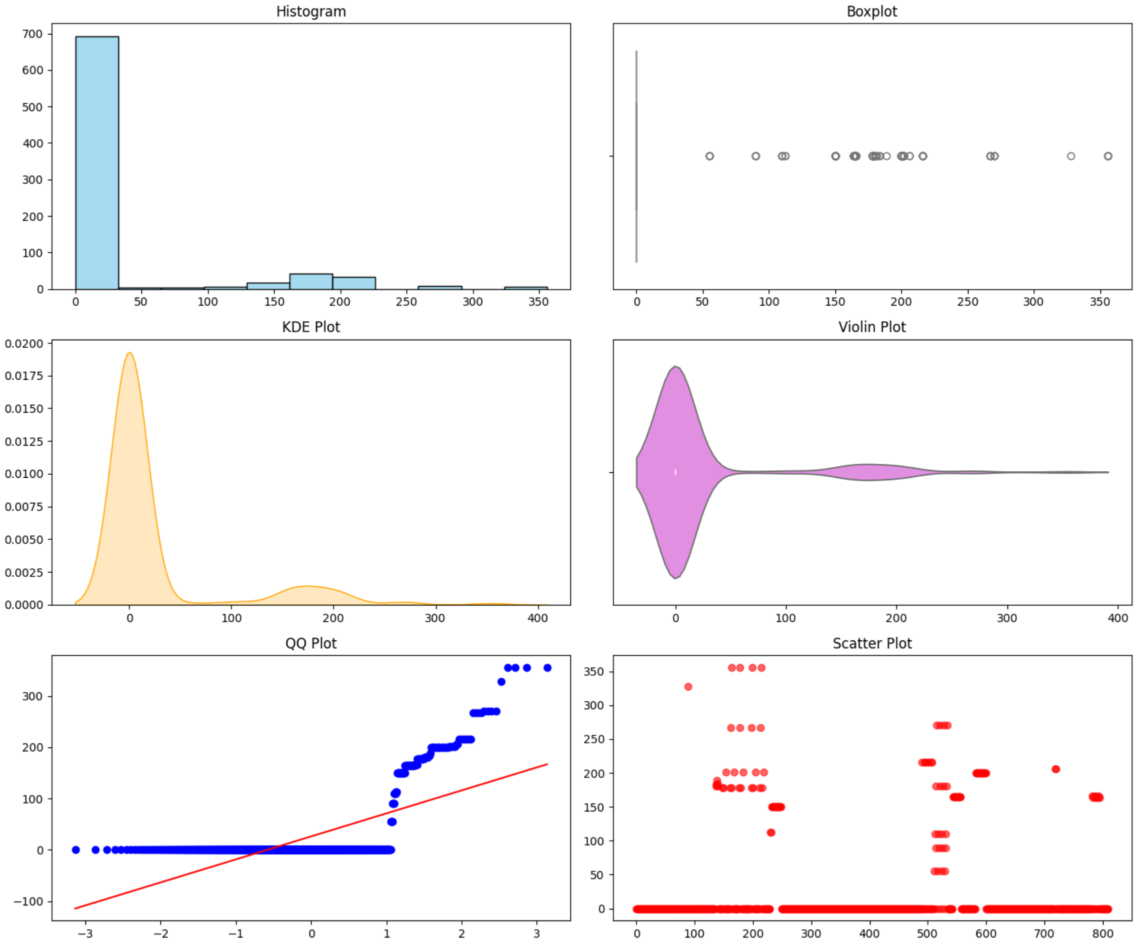


Figure 7: Distribution Plots for Fly Ash (FA)

3.3.7 Distribution Plots for Nano-Silica (NS)

As indicated in figure 8, nano-silica is right skewed and its skewness is 2.5324. The small mean value of 3.6386kg/m 3 and the median and mode of 0.0kg/m 3 show that it was not present in many samples and when it is present it is in small portions with a maximum range of 47.5kg/m 3. The kurtosis value of 6.7083 indicates that the peak around zero is sharp and having a long right tail whereas the standard deviation of 7.776 kg/m 3 means that the range non-zero values is not very large. The distribution plot would have a significant spike centered at zero, and a sharp drop-off and a light tail at higher values, the distribution plot depicting the fact that nano-silica does improve UHPC properties, infrequently.

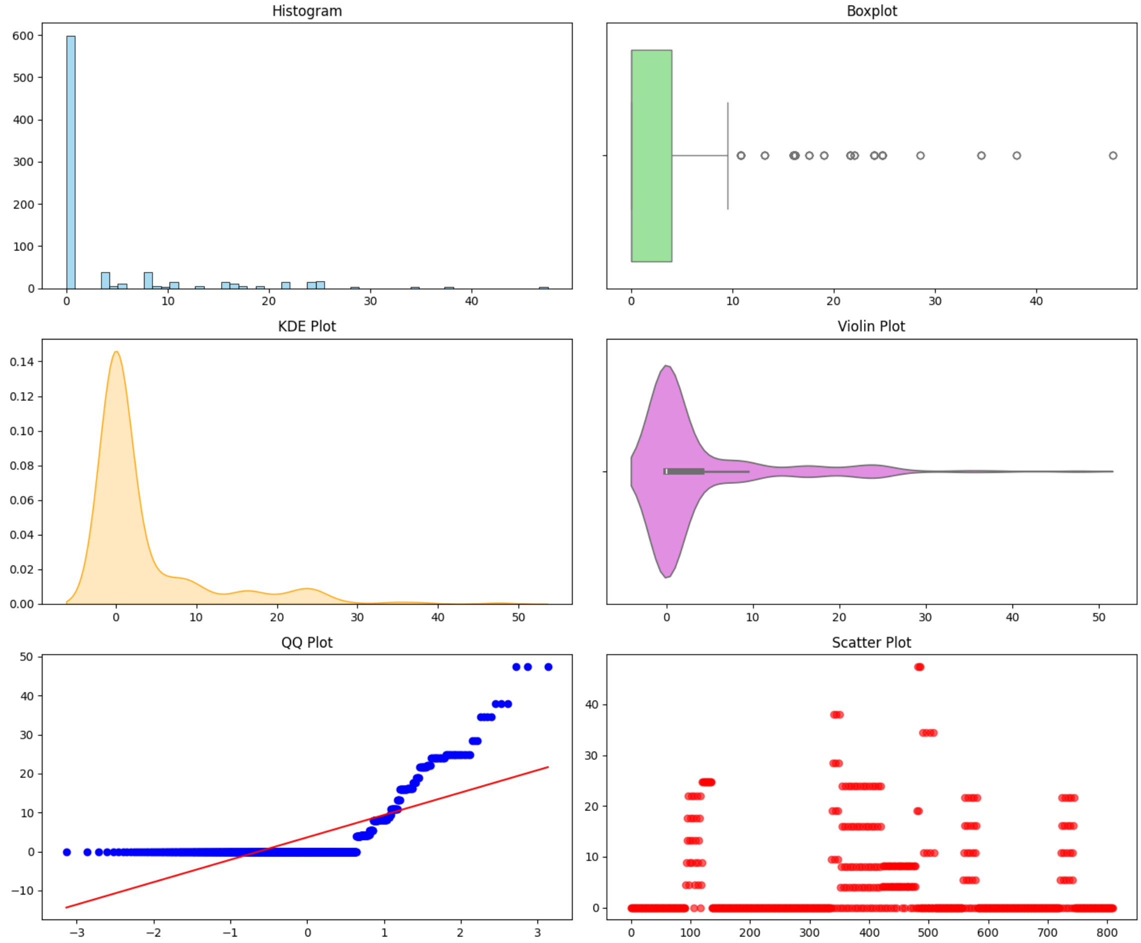


Figure 8: Distribution Plots for Nano-Silica (NS)

3.3.8 Distribution Plots for Aggregate (A)

Figure 9 shows that the aggregate content has a moderately right-skewed distribution and the value of skewness stands at 0.2436. However, the mean 1150.11 kg/m 3 has a slight right-tail compared to the median 1116.0 kg/m 3. The concentration is higher as the mode is 1231.0 kg/m 2. Besides, the negativity of kurtosis, -0.2107 shows that there is a flattened distribution that is not normal with its width range of 407.8 to 1992.0 kg/m3 and its standard deviation of 312.152 kg/m3. In turn, the distribution plot will be moderately symmetrical with a slight right skew and a fairly smooth-looking appearance, thus indicating the uniform use of aggregate in the UHPC mixes with a slight component of variation.

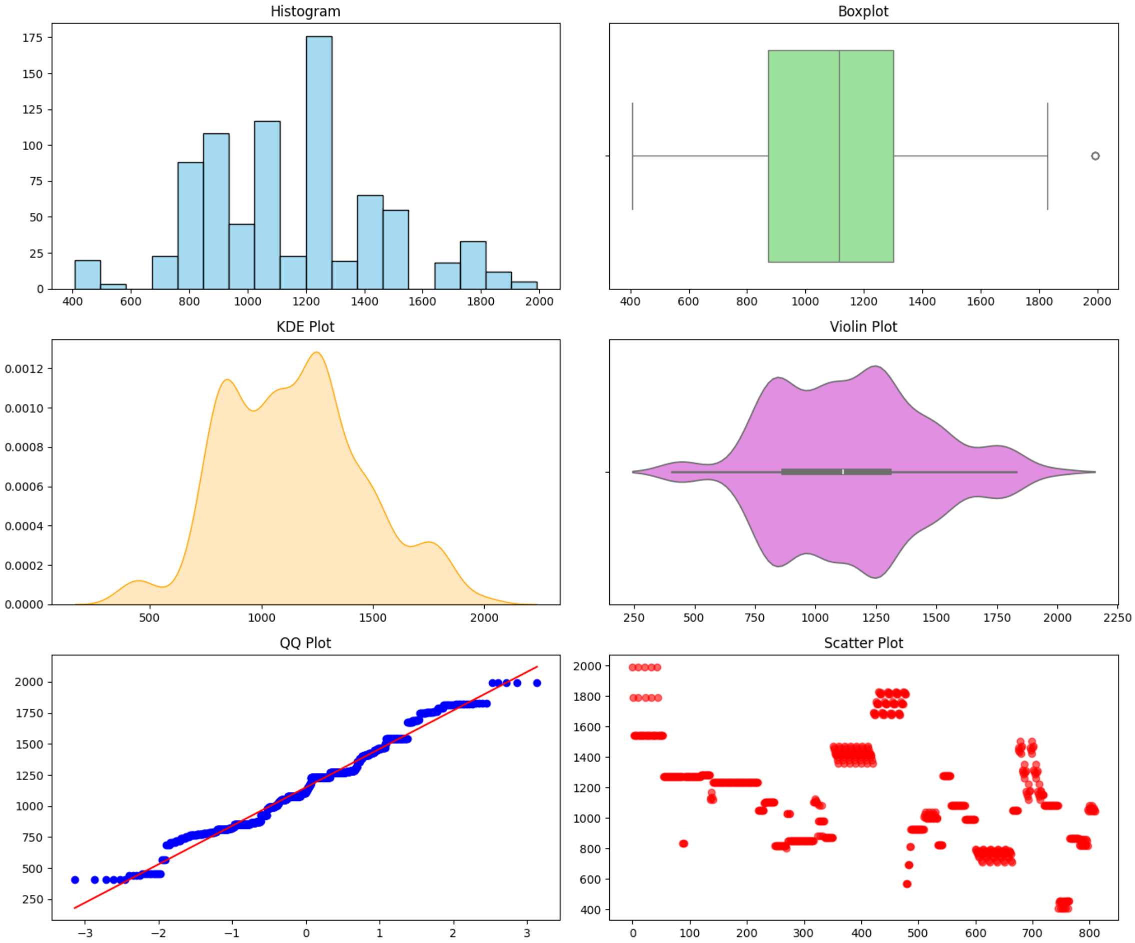


Figure 9: Distribution Plots for Aggregate (A)

3.3.9 Distribution Plots for Water (W)

By viewing figure 10 it is evident that the distribution of water content is a right skewed distribution with skewness of 0.6261. The value of 179.8911 kg/m 3 has a mean value that is slightly higher than the median value of 177.0 kg/m 3, which implies a tail towards the higher concentrations of water, but the mode is 160.0 kg/m 3, which implies maximum concentration in lower concentrations of water. Positive Kurtosis (K) of 1.7112 indicates that the distribution is peakier than the normal distribution and the standard deviation (standard deviation) of 25.5682 kg/m 3 indicates moderate variation between 90.0 and 272.6 kg/m 3. The distribution plot would quite possibly show a peak between 160.0 to 177.0 kg/m 3 with a middling right tail and this would therefore reflect the effect of water variation on workability and strength of UHPC.

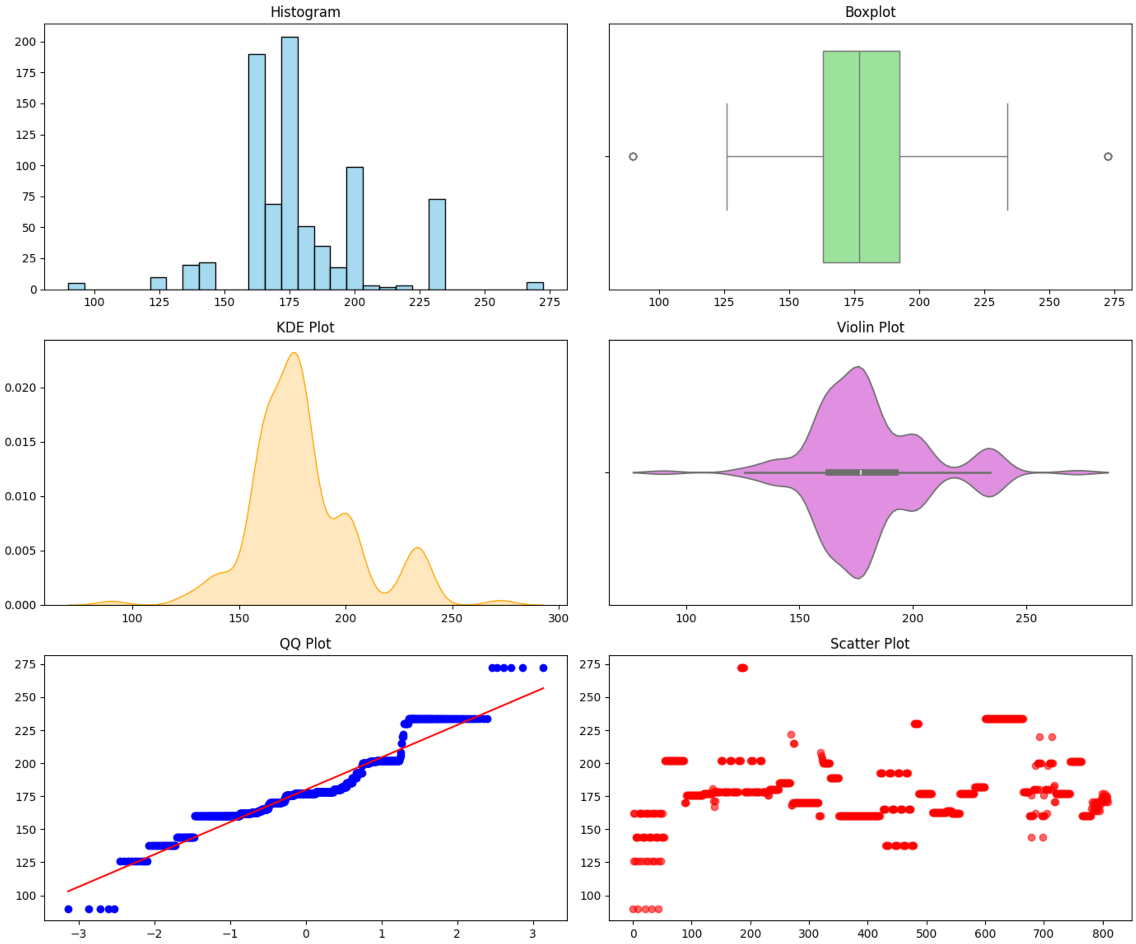


Figure 10: Distribution Plots for Water (W)

3.3.10 Distribution Plots for Fiber (Fi)

By looking at the Figure 11, it is clear that fiber content is right-shifted skewed, as its skew statistic is 0.8172. The medians and mode, which are all on the same point, 0.0 kg/m 3, means that many of the samples have no fiber in them, but the mean, which is 56.0444 kg/m 3, means that significant numbers are present when they are, with individual observations totalling up to 234.0 kg/m 3. The lower value of the kurtosis (0.9811) indicates a nearer to flat distribution, with values that are broadly spread out among the non-zero observations, and the corresponding standard deviation of 75.2306 kg/m 3 confirms the spread. As is likely to be seen in the distribution plot, this would tend to be very sharply peaked at zero, with a wider, lower outlined peak towards larger values, reflecting the wide range of quantities of fibers usually added to UHPC.

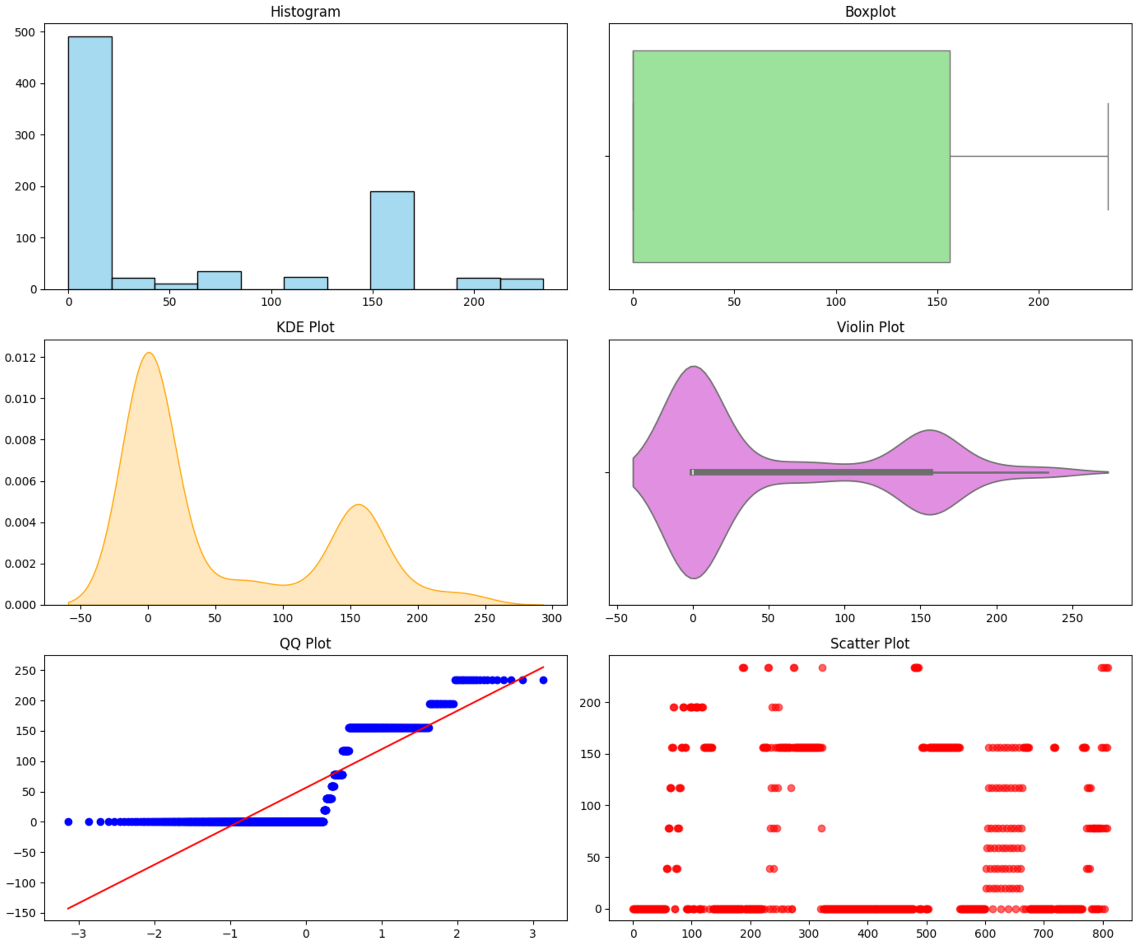


Figure 11: Distribution Plots for Fiber (Fi)

3.3.11 Distribution Plots for Superplasticizer (SP)

Distribution of superplasticizer content is close to symmetrical with a slight left tail as can be explained by the skewness of -0.176, as shown in Figure 12. The mode (45.0 kg/m3) suggests the presence of the concentration in the higher values, and a comparison between the mean (30.0309 kg/m3) and the median (30.2 kg/m3) shows that they are similar to each other. With a kurtosis of -1.0941 the distribution is flattened and has a less dense tail region whereas the standard deviation of 13.9935 kg/m 3 indicates moderate variations with a range of 1.1 to 57. 0 kg/m 3. As a result, the distribution plot would mostly be symmetrical with just a lean to the left and a broad and flat design with a sharp peak at 45.0 kg/m 3 due to the consistent application of superplasticizers to ensure the workability of UHPC.

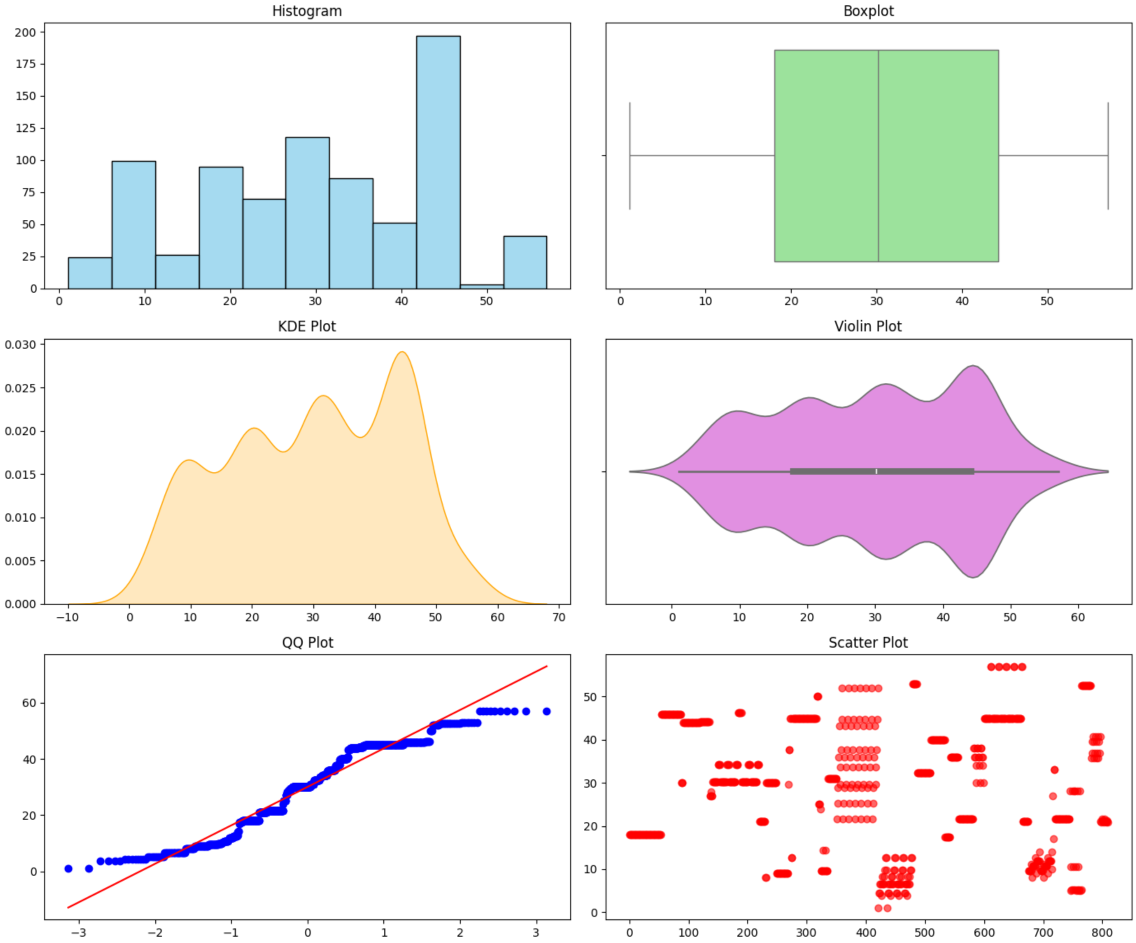
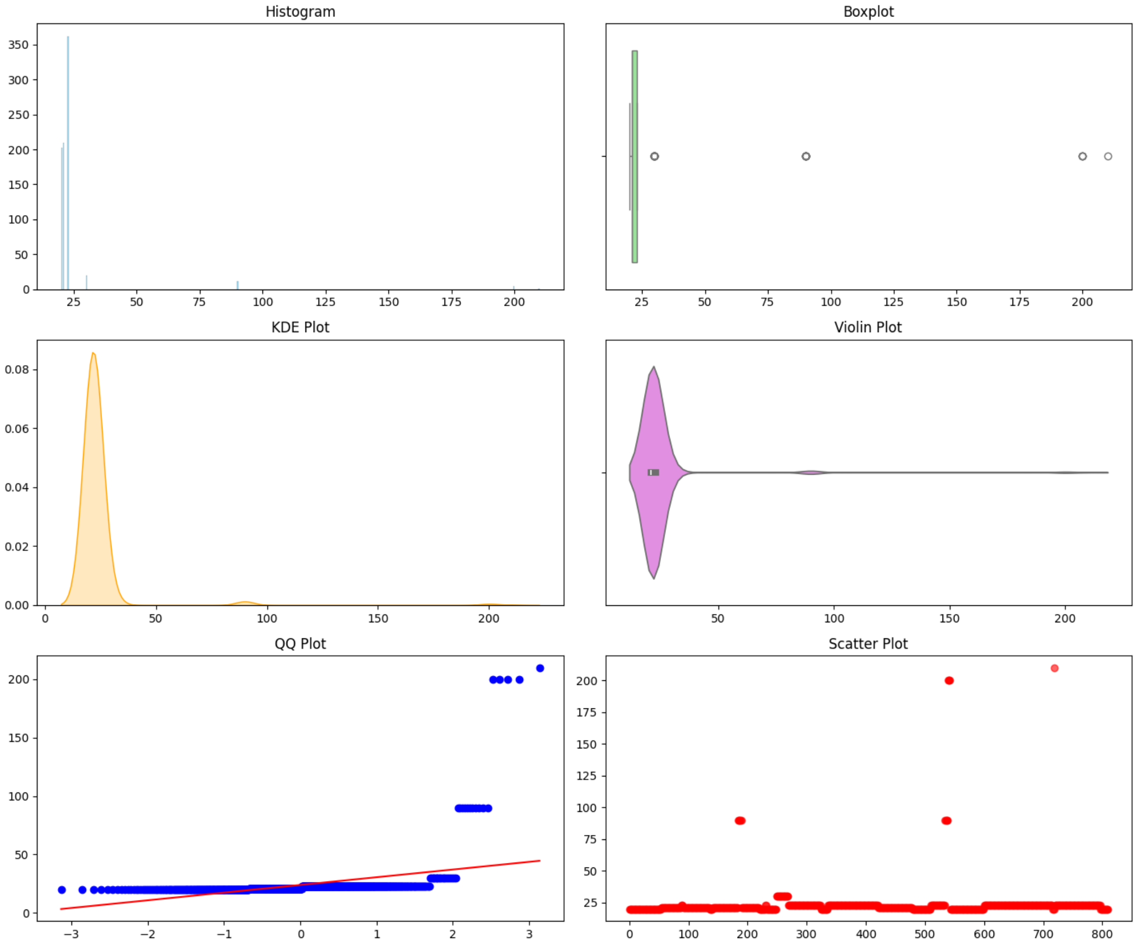


Figure 12: Distribution Plots for Superplasticizer (SP)

3.3.12 Distribution Plots for Temperature (T)

Figure 13 reveals that temperature has a highly right-skewed distribution with the skewness of 9.1441. The median of 21.0 o C and the mode of 23.0 o C indicate that the largest number of observations taken were recorded at fairly low temperatures, and the mean of 23.921 o C is high due to the existence of a long right tail to a high reading of 210.0 o C. The kurtosis of 91.7097 indicates the very concentrated peak about the point of 21.023.0 C, with a relatively sparse tail and a clear outlier, which is confirmed by the standard deviation of 16.2115 C. The matched distribution plot would thus have a peak that is quite tall and narrow at the low end and a long tail of the distribution to the high temperatures that might indicate specialized curing conditions in some of the experiments.



385

Figure 13: Distribution Plots for Temperature (T) 386

3.3.13 Distribution Plots for Curing Age (Age) 387

As seen in figure 14, there is a strong skew (3.8115) in the distribution of curing age as captured in the strong right tail. The modes of 28.0 days and median of 28.0 days points to a normal curing having its values, but the mean value of 37.0938 days is inflated by the right tail, which carries upto the maximal value of 365.0 days. The large Kurtosis of 18.6114 indicates the existence of a sharp peak at the point of 28.0 days with long tails and the relatively large standard deviation of 53.1159 supports the existence of a large variability of long curing times. This distribution will most likely portray a peak at 28.0 days and then sharply decrease with a long tail on the longer side, which is an observable tendency as per the daily testing needs along with the need to extend curing. 388–396

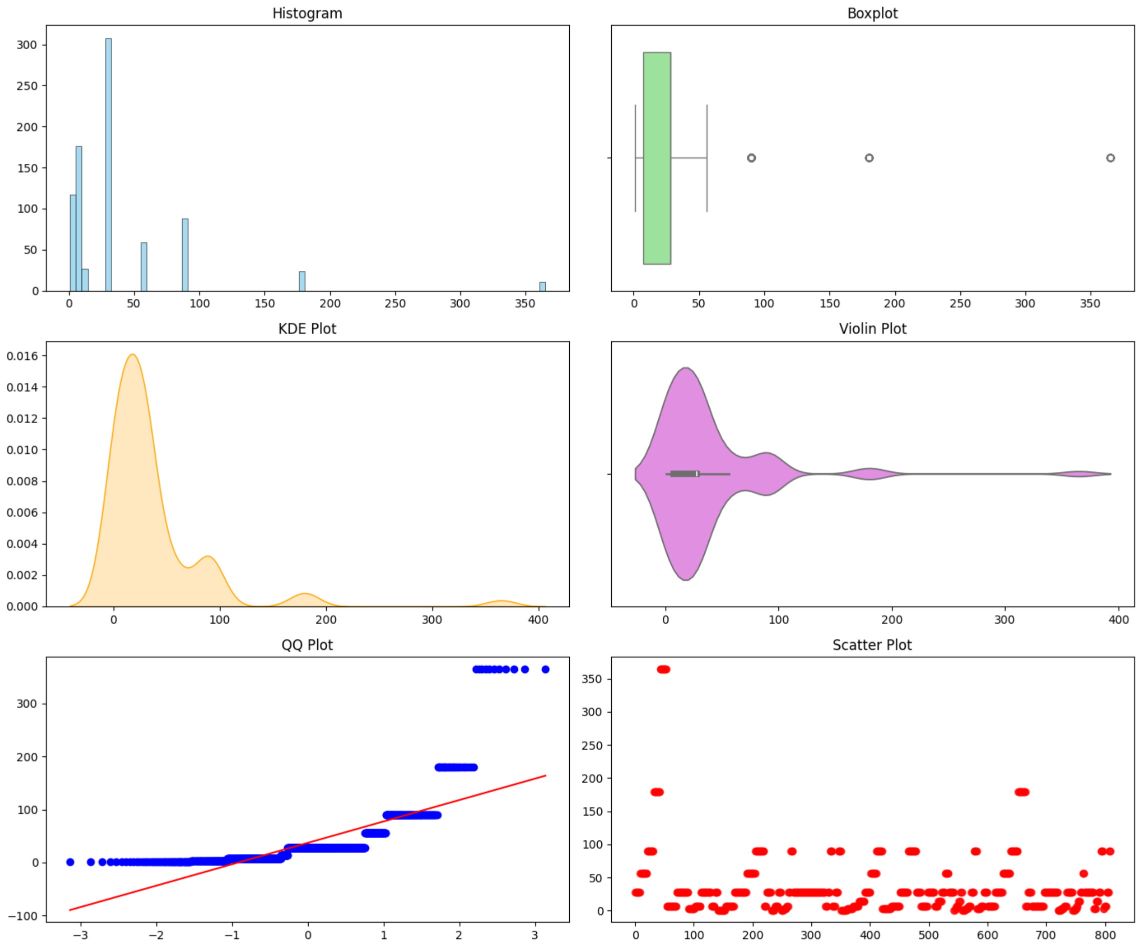


Figure 14: Distribution Plots for Curing Age (Age)

3.3.14 Distribution Plots for Compressive Strength (CS)

Using figure 15, it can be seen that the values of compressive strength are circumscribed by almost symmetric distribution since the skewness is 0.0024. The arithmetic mean of 123.1315 MPa and the median of 122.3 MPa are basically the same and the mode of 120.9 MPa is in a narrow gap which reinforces symmetry. The negative kurtosis (-0.5517) has stated that the distribution is slightly flattened as compared to normal distribution, with the range between 28.51 and 220.5 MPa and standard deviation of 40.2387 MPa. The related distribution plot would thus have a bell-shaped shape which would average at approximately 122123 MPa and be characterized with a harmonious as well as slightly concave shape which would make it quite exemplary to the predictive models used in the works conducted on UHPC.

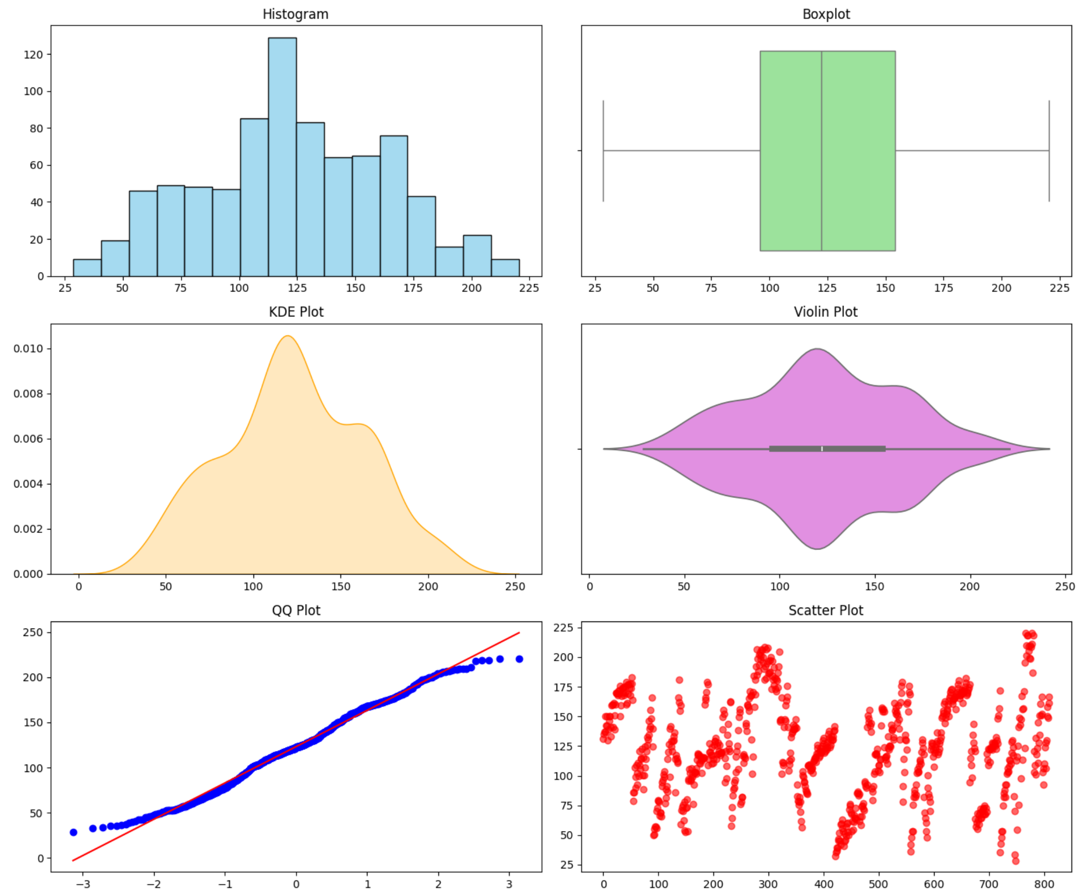


Figure 15: Distribution Plots for Compressive Strength (CS)

3.4 SHAP Summary Plot

Figure 16 below contains the SHAP (SHapley Additive exPlanations) summary that provides an effective insight into the predictions made by the machine learning by organizing the features according to their effect on CS and their contribution to the output on the value span that is associated to the features. However, the plot itself is not actually provided, but generally it illustrates properties in vertical order of their importance and horizontal distributive SHAP values (both positive and negative) along with color gradings of features proportions or feature magnitude (low to high feature values). By analyzing the results of UHPC properties and the dataset, such features as cement (C), silica fume (SF), and curing age (Age) might be considered the most important contributors since their roles in strength development are already determined. Large C (up to 1251.2) or SF (up to 433.7) values with high SHAP values would indicate that they have negative effects, raising the level of CS predicted, while water (W) could have negative SHAP values lowering the value of the CS, when it passes the optimal level. Zeros proliferated features (S, LP, QP, FA, NS, Fi) might be having variable effects: when they happened, they would boost: adding a NS feature increased the score by up to 47.5, but these features are infrequent so it cannot make much difference in the abstract. The dispersion and orientation of the SHAP values would demonstrate which of the consequences are linear or nonlinear and where the effects are interactions, and, thus, this graph would be critical in figuring out which factors contribute the most to UHPC strength estimations.

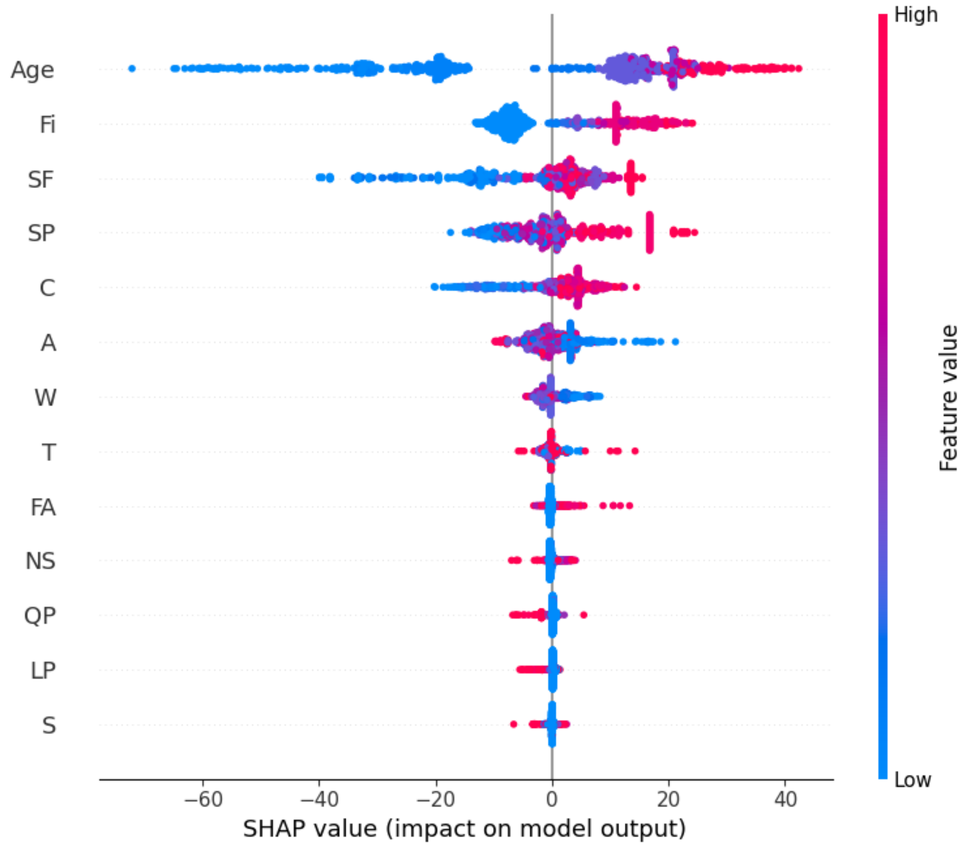


Figure 16: SHAP Summary Plot

3.5 SHAP Dependence Plot

The SHAP dependence plot goes into a close up on the relationship between a single feature and CS, and plots its values against those of the SHAP values, and in many cases color is used to indicate that an interaction with some other feature has been encountered. Lacking the particular characteristic as in the case of curing age (Age) whose range is wide (1.0 to 365.0) and its mean is large (37.0938) as displayed in figure 17. The plot may either indicate SHAP values raised as the Age proceeds, showing that the longer the curing, the stronger the concrete is, as it is with the principles of concrete curing. There might be a flattening of the steep increase in SHAP values past the 28 days interval (median: 28.0) due to declining returns. colouring by 'Temperature (T) may indicate that Age has a positive synergistic effect with T up to 210.0. Alternatively, a plot of a feature such as water (W) would give decreasing values of SHAP, with patterns of scattering as W combines with SP or C. The plot is a more subtle look at the behavior of each individual feature in the model in addition to the other information provided in the SHAP summary plot.

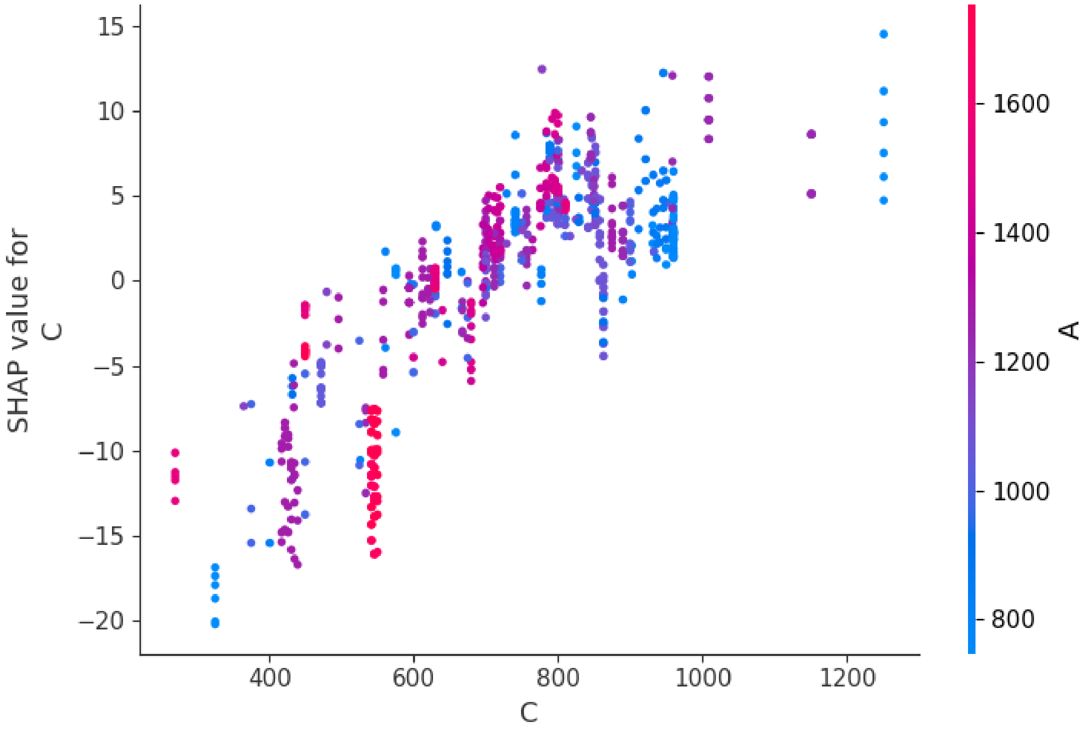


Figure 17: SHAP Dependence Plot

3.6 Parametric Sweeps

3.6.1 Parametric Sweeps for Cement (C)

Figure 18 represents parametric sweep (C) in Ultra-High-Performance Concrete (UHPC) database, which perhaps tests the effect of different cement content to the compressive strength (CS) since its mean is 737.9146 kg/m 3 and a range of 270.0 to 1251.2 kg/m 3. Having a standard deviation of 173.4572 kg/m 3 with a slightly left skew (-0.2288), the sweep would first trend positively with increment in cement content when initial cement strength would progressively increase with the density of the matrix being improved near an optimum level around the median (770.5 kg/m 3) or mode (960.0 kg/m 3). Along and beyond this line, the curve could be flat or slightly decreasing because of diminished returns or too much cement, which caused the workability of the concrete to be a problem, as exemplified by the fact that the sample variance is quite big at 30087.1408. In this sweep the importance of cement would be emphasized to allow the mix design to focus on the measurement of strength and utility.

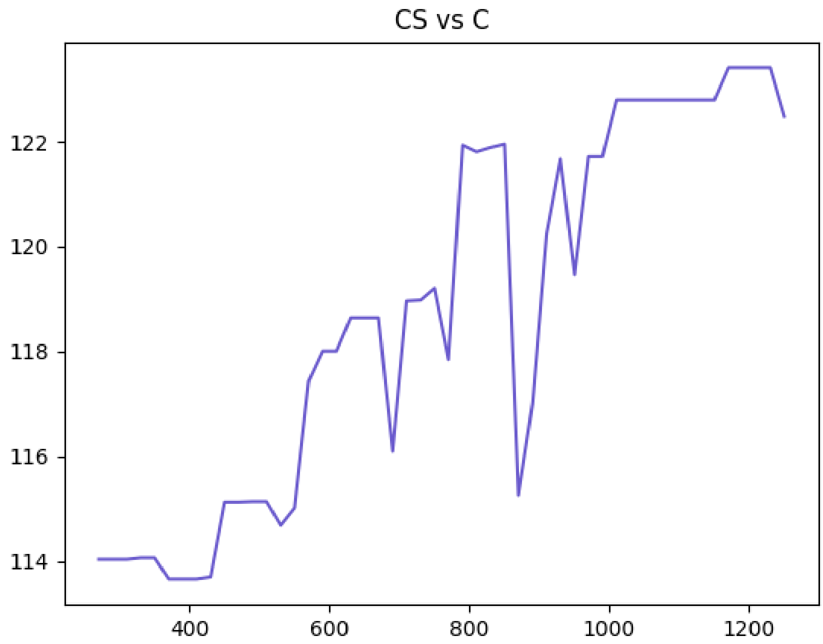


Figure 18: Parametric Sweeps for Cement (C)

3.6.2 Parametric Sweeps for Slag (S)

Figure 19 below parametric sweep for slag (S) explores how it affects CS with the mean of 25.1946 kg/m, and the range it testes with the range that is between 0.0 and 375.0 kg /m has the median and mode of 0.0 which means the absence of slag in most of the samples. The skew of 3.0176 and kurtosis of 8.2266 coupled with a standard deviation of 74.3655 kg/m 3 indicate that the sweep test would result in little or no effect close to zero and with a tendency of increased CS as slag content increases possibly due to pozzolanic effects increasing microstructure. But effect may decays or fluctuates at increased values (as much as 375.0 kg/m 3) in cases of non-consistent application, which is in line with the sporadic right-motivated distribution and difficult to determine the exact contribution with certainty.

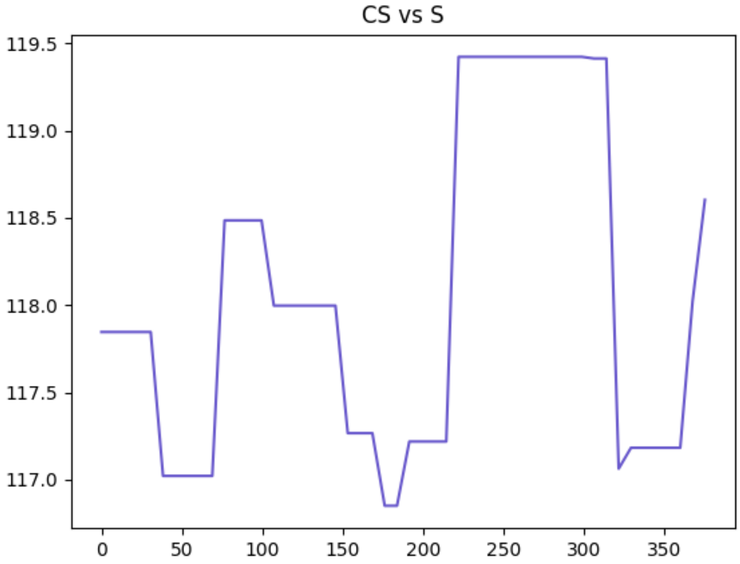


Figure 19: Parametric Sweeps for Slag (S)

3.6.3 Parametric Sweeps for Silica Fume (SF)

Parametric sweep of silica fume (SF) is provided in the following figure 20 that addresses its effect on CS, with a mean of 136.9872 kg/m 3, a median of 144.0 kg/m 3, a range between 0.0 to 433.7 kg/m 3, even though its mode was 0.0. With the negative kurtosis (-0.5967), due to the mild skewness of 0.259, it can be seen that there could be little change at zero, hence a steady increase in the CS with an increase in silica fume leveraging its abilities in pozzolanic effect and micro filling effects. The curve may reach a maximum at the median or even higher values, and stagnate, which shows the bimodal pattern (zero vs. 144.0kg/m 3) and means that silica fume made a great contribution when it is added to enhance strength.

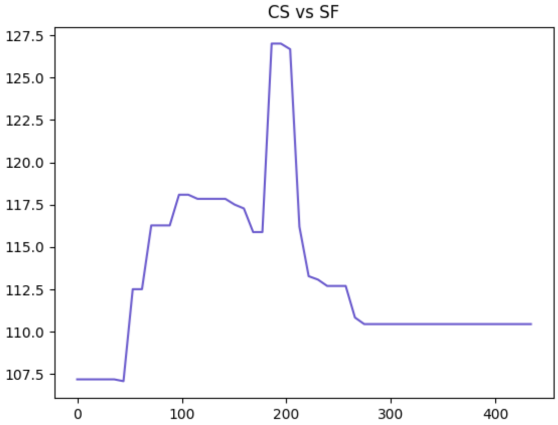


Figure 20: Parametric Sweeps for Silica Fume (SF)

3.6.4 Parametric Sweeps for Limestone Powder (LP)

It is to be seen that the next figure, 21, is parameter sweep of limestone powder (LP), which measures its effect on CS and has an average of 41.9295 kg/m3 and a range of 0.0 to 1058.2 kg/m3; the median and mode value of 0.0 are indicating that the limestone powder is rather quite rare. The high index of skewness (4.7579) and kurtosis (28.3356) setting levels of standard deviation to 133.1315 kg/m 3 show that this response has a flat shape at zero, increased strength may be recorded at low-medium values across to filler effects, but decreasing returns or irregularity may be at higher ones (up to 1058.2 kg/m 3). The sweep would tend to display a threshold effect, benefits were restricted, unless delved on cautiously into proportion, the sparse, heavily tailed distribution.

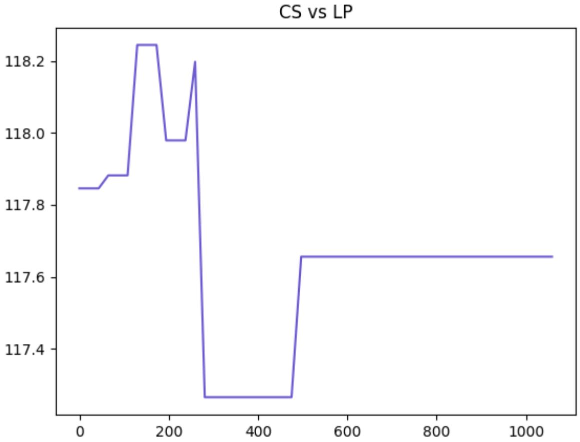


Figure 21: Parametric Sweeps for Limestone Powder (LP)

3.6.5 Parametric Sweeps for Quartz Powder (QP)

In figure 22, parametric sweep of quartz powder (QP), the average 33.271 kg/m 3 varied between an upper limit of 397.0 kg/m 3 and lower limit 0.0 with a median and mode of 0.0 which implies that it is not habitually used. With the standard deviation of 79.6739 kg/m 3 and skewness of 2.2829 and kurtosis of 4.2442, the curve indicates a slight effect at zero, then tendency rise at low levels of CS because the element is a fine aggregate, and it might settle at a higher value as represented (up to 397.0 kg/m 3). The effect of the sweep would be nonlinear, and inclusion of the sweep in sparse form, that is, a non-proportional addition of the sweat, will make its optimization in UHPC mixes hard.

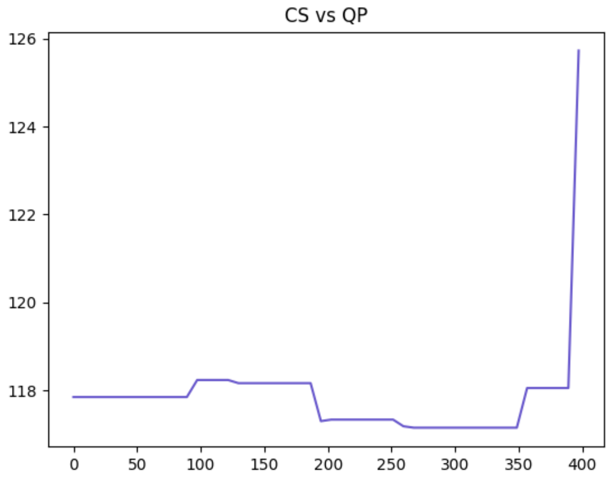


Figure 22: Parametric Sweeps for Quartz Powder (QP)

3.6.6 Parametric Sweeps for Fly Ash (FA)

The next figure 23, parametric sweep of fly ash (FA) determines its effects on CS with a mean of 26.2649 kg/m h and a range of 0.0 to 356.0 kg/m h with the median and mode of 0.0 indicating that it is not extensively utilized. The skewness value of 2.4922 and kurtosis value of 5.3377, the standard deviation of 67.4617 kg/m 3, depicts that most calculations were close to zero with the inspection that CS may show slow rise at moderate values but diminishes when CS is intensified to greater height so far due to pozzolanic effect. The sweep would indicate that there perhaps is a cut-off point that further fly ash does not contribute much, as indicated by the right-accurate and leptokurtic distribution of fly ash.

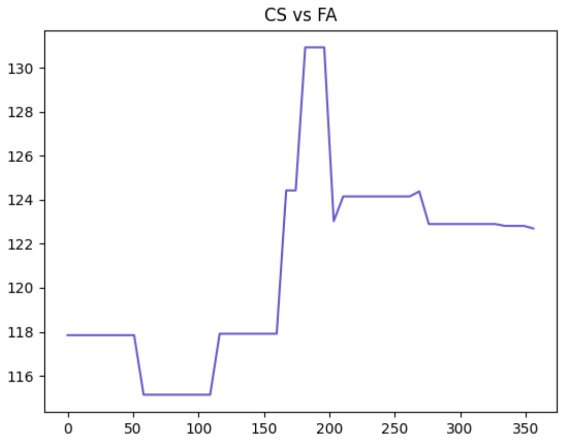


Figure 23: Parametric Sweeps for Fly Ash (FA)

3.3.7 Parametric Sweeps for Nano-Silica (NS)

The average value of nano-silica (NS) is 3.6386 kg/m 3, whereas the range is 0.0 and 47.5 kg/m 3 as the following figure 24 shows parametric sweep for nano-silica (NS) and the rarity of CS is observed at its median and mode of 0.0. With skewness of 2.5324 and kurtosis of 6.7083, and a 7.776 kg/m 3 standard deviation, an initial flat trend up to zero, followed by a steep rise of CS in the lower-level due to its nano-scale reinforcement of the matrix, may level off above a CS of 47.5 kg/m 3. This would make the sweep emphasize its strong yet narrow impact, which is indicative of the sparse and heavily tailed impact.

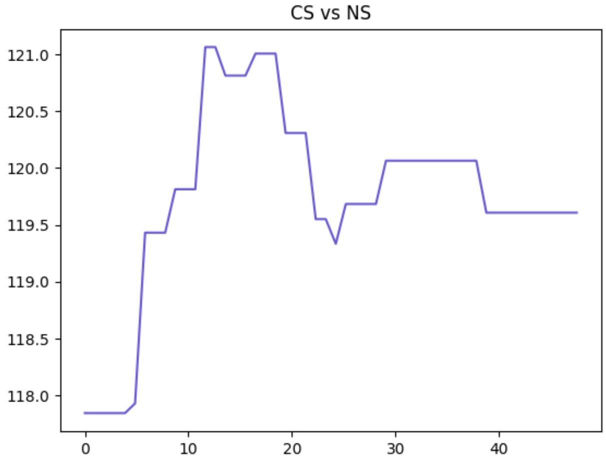


Figure 24: Parametric Sweeps for Nano-Silica (NS)

3.6.8 Parametric Sweeps for Aggregate (A)

The parametric sweep of aggregate (A) to evaluate its utilization in CS with a mean of 1150.11 kg/m3 and a cover of 407.8 to 1992.0 kg/m3 is illustrated in the following figure 25 with the median at 1116.0 kg/m3 and the mode at 1231.0 kg/m3 showing an even spread. The standard deviation of 312.152 kg/m 3 indicates that the mild skewness (0.2436) and the negative kurtosis (-0.2107) indicate that the level of CS is increasing steadily to an optimum level in the vicinity of mean and then deteriorating due to excessive weight or poor workability. The sweep would give an optimum aggregate content that would be balanced as far as strength and proportions of mixes are concerned.

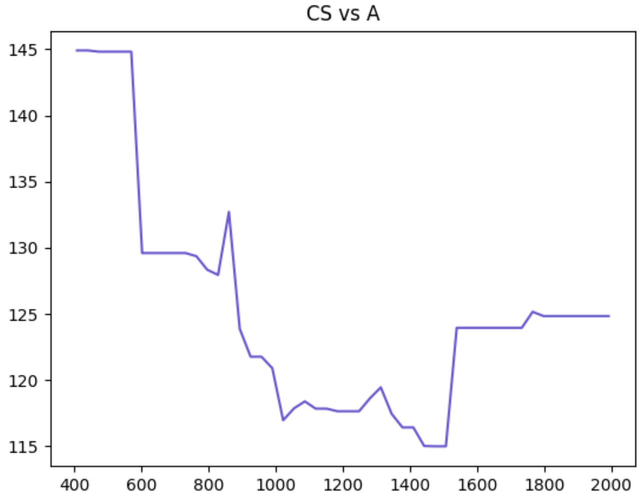


Figure 25: Parametric Sweeps for Aggregate (A)

3.6.9 Parametric Sweeps for Water (W)

Figure 26 presented below studies the effect of water (W) on CS, with a mean value of 179.8911 kg/m 3) between 90.0 and 272.6 kg/m 3) with a median of 177.0 kg/m 3) and mode of 160.0 kg/m 3) indicating a moderate use. There is too much skewness of 0.6261 and kurtosis of 1.7112 with a standard deviation of 25.5682 kg/m 3, which implies that the CS peaks at lesser water contents and decreases with the increase in water, then slightly increasing beyond 272.6 kg/m 3. The sweep would put some emphasis on the inversely proportional relationship between water content and strength, which is important in the design of UHPC.

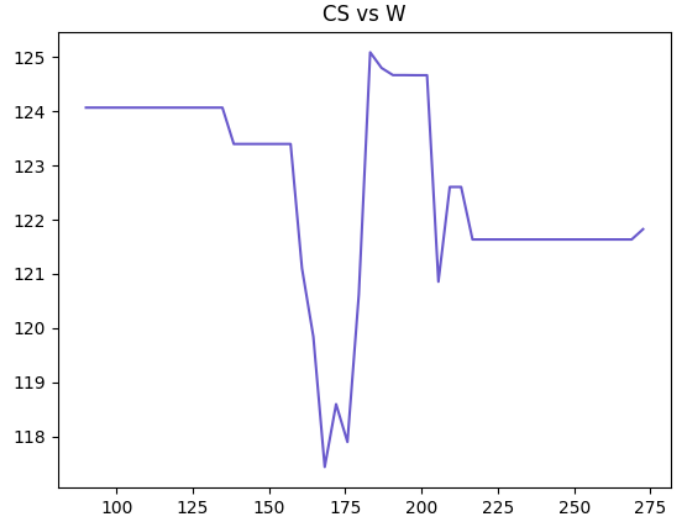


Figure 26: Parametric Sweeps for Water (W)

3.6.10 Parametric Sweeps for Fiber (Fi)

The above figure 27 reveals parametric sweep on fiber (Fi) and its effect on CS where mean indicates CS to be 56.0444 kg/m 3 with range of 0.0 to 234.0 kg/m 3, where 0.0 is the median and mode showing optional fiber usage. Kurtosis is negative (k = -0.9811) and the skewness is positive corresponding to 0.8172, which indicate that the response curves (CS and ductility) will be flat at zero whereas it rises progressively at moderate rates, but tends to remain constant or diverges at higher courses. This vigorish would serve as evidence of the use of fiber in increasing toughness with its advantageous aspect lying in the occasional availability of the ingredient.

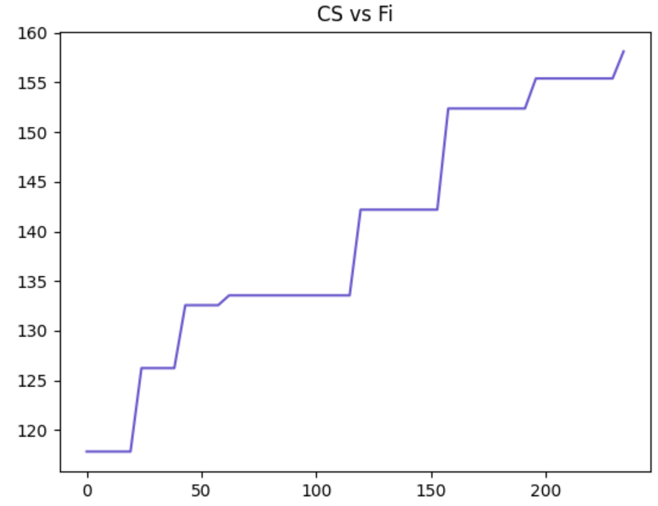


Figure 27: Parametric Sweeps for Fiber (Fi)

3.6.11 Parametric Sweeps for Superplasticizer (SP)

The parametric sweep used to determine the influence of superplasticizer (SP) on CS is given in the following figure 28, with an average of 30.0309 kg/m [3], a range between 1.1- 57.0 kg/m [3] and the median 30.2 kg/m [3] and mode 45.0 kg/m [3] appear to be used consistently. The negative kurtosis (-1.0941) is associated with minor left skew (-0.176) indicating the state of the CS is stable with only slight increments on workability chances possible, as opposed to the direct increase of strength, and stabilizes after 57.0 kg/m3. The sweep would emphasis its by-who-little part in the optimization of UHPC mixes.

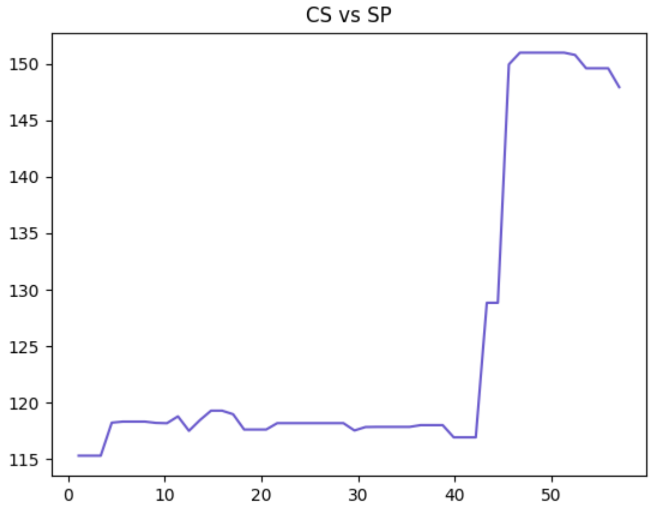


Figure 28: Parametric Sweeps for Superplasticizer (SP)

3.6.12 Parametric Sweeps for Temperature (T)

The parametric sweep of temperature (T) is presented in the figure 29 below, which analyses how it can affect CS; with the mean of 23.9210 C and a range of 20.0-210.0 C, the median (21.0 C) and the mode (23.0 C) show what would be the usual conditions. The very high skews (9.1441) and kurtosis (91.7097) with standard deviation of 16.2115 degree C indicate a fast rate of CS rising at the low temperatures and hits maximum at about 21-23 degree C then declines or shows variations at the higher values based on thermal effects. The sweep would also unveil the crucial place of temperature, and there would also be optimum ranges to cure the curing.

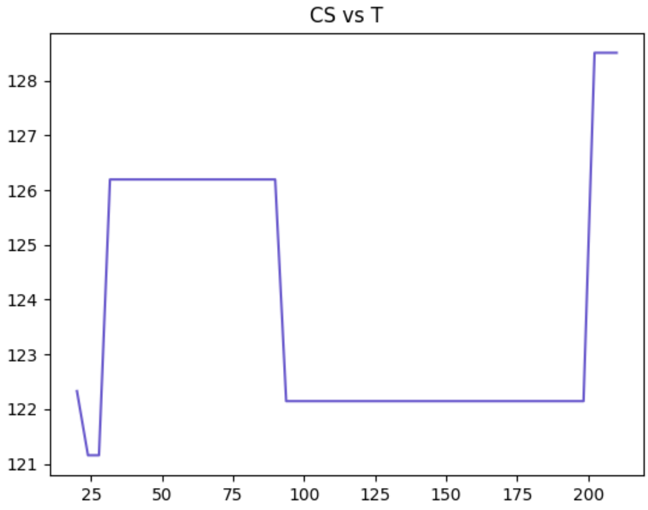


Figure 29: Parametric Sweeps for Temperature (T)

3.6.13 Parametric Sweeps for Curing Age (Age)

The next figure 30 indicates that parametric sweep of curing age (Age) evaluates its effect on CS with a mean of 37.0938 days and a range of 1.0 days to 365.0 days whereas the median 28.0 days and mode 28.0 days correspond to the standard practice. Skewness of 3.8115 and kurtosis of 18.6114 with a standard deviation of 53.1159 days show a sharp rise CS till 28 days and later the rise will be slower or will become stagnant with the decrease of returns with increasing days after 365.0 days. The sweep would verify that ageing, a significant effect that is time-dependent with UHPC strength, gets cured.

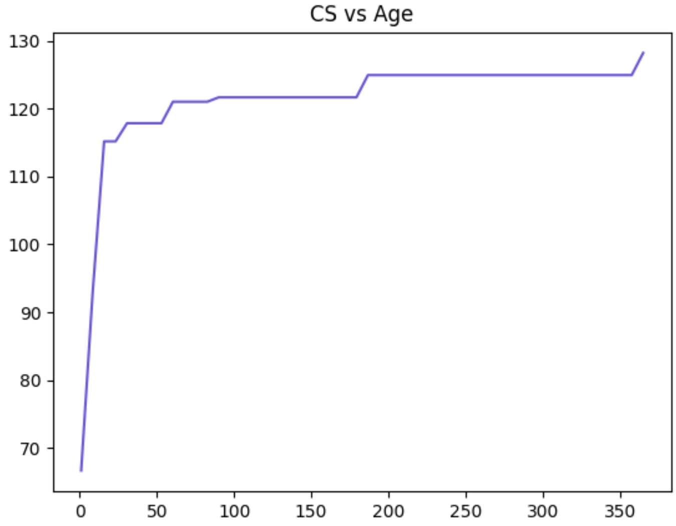


Figure 30: Parametric Sweeps for Curing Age (Age)

3.7 Model Evaluation and Selection

Stratified 10-fold cross-validation was used to evaluate the performances of different machine learning models since it is robust and fair in prediction compressive strength of Ultra-High-Performance Concrete (UHPC). The best model among the tested algorithms is the Gradient Boosting Regressor (GBR) with the R 2 0.970 result, and the best metrics of error (MAE: 6.98 MPa, RMSE: 8.08 MPa, MAPE: 4.86%). The sequential minimum residual capacity of GBR confers it with suitability when used in reflecting the multivariate nonlinear problems used by the 14 UHPC input variables. There were AdaBoost and Random Forest that closely followed GBR with R 2 of 0.961 and 0.960 respectively. These ensemble models had outstanding predictions ability in the fact that they combined many weak learners and represented feature interactions properly. Another model, the Neural Network (Multilayer Perceptron) also did quite well with R 2 of 0.947, demonstrating that it can learn lower-level patterns in data, but it had a somewhat larger RMSE and MAE than the ensemble models.

3.8 Rationale for Top Four Model Selection

The four of most accurate models that were chosen and culminated on winning were Gradient Boosting, AdaBoost, Random Forest, and Neural Network due to their good predictive and generalization accuracy as presented in the table 3 below. Ensemble (GBR, AdaBoost, RF) was found to be the most beneficial when it comes to minimizing the bias and variance, which is an ultimate consideration when dealing with materials such as UHPC that no longer exhibit similar behavior to linear and lower-dimensional systems. Compared to simpler models like kNN and Linear Regression models, these four advanced models had low cross-validation error and a high R 2 which is indication of good model stability with the entire dataset. In addition to this, these models were also preferred in succeeding symbolic regression because of their consistency in terms of substantial feature patterns, which is needed in producing interpretable equations that would possibly help in practical application in the field of engineering. All in all, the models chosen also give a strong base in terms of not only precise projection but also open derivation of equations when working on predicting UHPC compressive strength.

Table 3: Comparative Performance Metrics of Machine Learning Models

Model	R² (Test)	MAE	RMSE	MAPE (%)	CVRMSE	Comments
Gradient Boosting	0.970	6.98	8.08	4.86	5.67	Best performance overall — lowest errors, highest accuracy.
AdaBoost	0.961	7.93	7.93	5.94	6.44	Robust ensemble model, slightly less accurate than GB.
Random Forest	0.960	8.08	8.08	5.93	6.57	Strong performance, great generalization.
Neural Network	0.947	9.25	9.25	6.55	7.51	Good accuracy but slightly higher errors.
Decision Tree	0.887	13.49	13.49	10.59	10.96	Interpretable, but weaker generalization and prone to overfitting.
k-Nearest Neighbors	0.790	18.45	18.45	13.46	14.98	Simple but underperforms due to sensitivity to high dimensionality.
Linear Regression	0.685	22.57	22.57	17.97	18.33	Weakest performance — cannot model nonlinearities in UHPC data.

3.9 Symbolic Regression Interpretation for Original Dataset

The symbolic regression model derived from the original dataset using MEPX offers a mathematically interpretable expression that captures the complex interplay between various concrete mix parameters and compressive strength (CS). The equation integrates key variables such as water content (W), silica fume (SF), age (Age), fiber content (Fi), and superplasticizer (SP) through a combination of logarithmic, exponential, and multiplicative operations which is the following:

$$\boldsymbol{CS = (ln^2(W) - log10(W)) \cdot ln(Fi + (\frac{SP}{log10(W)}) + Age.ln(SF + ln(w)))}$$

Interaction between the silica fume and the water is linked with the concrete strength but in a non-linear and synergetic association. To capture the changing effect of pozzolanic action over time, the subsequent product of the term mean strength after ageing with age, which is logarithmic to age, is used. The fact that the effect of fineness index and specific surface point are additive linear means that the two of them support the improvement of the concrete performance. In general, the proposed formula offers clarity and a facile to interpret framework that is in line with the ideas of an engineer and offers a transparent approach to forecast the concrete strength without involving the black-box models.

3.10 Symbolic Regression Interpretation Based on Gradient Boosting Model

The symbolic regression expression generated from the Gradient Boosting model captures a complex yet meaningful relationship between key UHPC mixture components and compressive strength is as follows:

$$\boldsymbol{Compressive\ Strength} = \mathbf{tan}\ (\boldsymbol{Age}) + [\boldsymbol{SP} + \mathbf{2} \cdot (((\sqrt{\mathbf{log}(\boldsymbol{C})} + \mathbf{log}(\boldsymbol{C}) + (\mathbf{log}\ (\mathbf{log}\ (\boldsymbol{C}) \cdot \boldsymbol{Age}) \cdot \sqrt{\mathbf{log}(\boldsymbol{C})} + \mathbf{log}\ (\boldsymbol{C})) - \mathbf{log}\ (\boldsymbol{C})) - \mathbf{log}\ (\boldsymbol{C})) + \frac{\boldsymbol{SF}}{\boldsymbol{SP}}$$

The last model structure is a composite regression where the cement content C, the curing age A, silica fume SF and superplasticizer SP are presented in addition to mathematical operators of advanced degree including logarithm, square root, tangent and multiple linear combination of interactions. Cement concentration is presented in logarithmic form and in square root form, implying an increasing-return outcome at big doses, whilst curing age influences prediction of strength in a linearly scaled logarithmic product form and a non linear tangential ratio form, hence implementing a tangent dynamic of hydration with time.

The level of superplasticizer normalizes the silica fume thus indicating their joint component of optimal particle packing and work ability. Generally, the equation has worked positively with the non-linear, synergistic reaction of the chemical composition as well as the curing conditions. Despite its mathematical complexity, the model is interpretable, allowing an engineer to come up with a transparent, practical approximation of the ultrahigh-performance concrete compressive strength, strengthening the faith in performance-based mix design and optimization.

3.11 Symbolic Regression Interpretation Based on AdaBoost Model

The symbolic regression model derived from the AdaBoost algorithm provides an interpretable mathematical expression for estimating the compressive strength of Ultra-High-Performance Concrete (UHPC) based on key input variables. The equation incorporates superplasticizer (SP), temperature (T), age of curing (Age), and silica fume (SF), and is expressed as the following:

$$\boldsymbol{CS} = \frac{\boldsymbol{Silica\ Fume}}{\sqrt{\boldsymbol{Superplasticizer}}} + \left(\mathbf{2} \times \boldsymbol{Superplasticizer} + \sqrt{\mathbf{2} \times \boldsymbol{Temperature} \times \boldsymbol{Age}}\right)$$

This model lays emphasis on both chemical and environmental factors which have a synergetic effect on strength development. The use of Superplasticizer has a dual effect in that it directly increases workability in addition to indirectly offsetting the impact of silica fume as they exhibit an inverse proportional relationship to one another. The variable temperature and curing age is a term that expresses their synergistic interaction in cement hydration and microstructure refinement with time. By unveiling this nonlinear but physically-based relationship, the symbolic model provides an estimatable formula, in a human-readable form, which engineers can use in a readily usable form to estimate quickly and to optimize the mix process without need for complex simulation and black-box modeling tools.

3.12 Symbolic Regression Interpretation Based on Neural Network Model

The symbolic regression equation derived from the Neural Network model offers a nonlinear yet interpretable representation of the compressive strength of Ultra-High-Performance Concrete (UHPC) based on four influential parameters: cement content (C), fiber content (Fi), superplasticizer (SP), and curing age (A). The equation is expressed as:

$$\boldsymbol{CS = [(\sqrt{C + Fi}.\,ln(SP + A) - (\frac{SP}{A}) + \frac{\sqrt{A} + Fi}{SP}] + [\mathbf{log10}((\sqrt{A} + Fi).\,A)]^{\,2}}$$

In this formulation, the values indicate an increase of compressive strength with an increase in the curing age and fiber addition particularly when combined with proper dosage of cement and adequate super plasticizer. This equation contains both additive and multiplicative constructive effects and penalizing inverse and logarithmic effects, which means that there is a compromise between the efficiency of the binder, the maturity of the hydration, and the reinforcement impact. The non-linear presences of nested terms of square root and logarithms reflect the non-linearities that are layered in terms of reasoning of the neural networks on successes in modeling material behavior. Although quite complicated, that equation still can be used as a simplified version to be appreciated by UHPC designers providing a mathematical approximation consisting of the studied relationships of high-capacity model neural network.

3.13 Symbolic Regression Interpretation Based on Random Forest Model

The symbolic regression expression generated from the Random Forest model provides a compact yet non-linear analytical representation of compressive strength in Ultra-High-Performance Concrete (UHPC) using five critical features: cement (C), age (A), quartz powder (QP), superplasticizer (SP), and nano-silica (NS). The equation is given by:

$$\boldsymbol{CS = [log(C.A)^2 + tan(QP) - tan(tan(QP)) - (tan(tan(QP)) - SP)] - NS}$$

This model treats the multiplicative factor of concrete cement and curing age as a logarithmic term such that the strength gain could be stressed as a result of long hydration time. The fact that the way quartz powder influences workability and packing density appears to be complex and nonlinear is revealed by the nested tangent transformations performed on it. Superplasticizer (SP) and nano-silica (NS) is added as moderating variable and finishing subtractive variable, respectively, which can be interpreted as over-refining or interfering at very high dosages. Even though a non-parametric and rule-based model like Random Forest is already satisfying this need by the form of non-readable, precise output, here this symbolic output comes to present more usable output and present the internal logic of the model as a clear equation.

3.14 Symbolic Regression Performance Across Models

The interpretability and performance of symbolic regression were compared to assess the original dataset and the results of the best four machine learning models, namely AdaBoost, Gradient Boosting, Neural Network, and Random Forest, through the use of MEPX. Every symbolic representation produced by MEPX was an attempt to estimate the desired compressive strength of a very simplistic mathematical expression in terms of the input features. The comparisons of the models obtained were based on the ranking of Mean Absolute Error (MAE). The symbolic regression that was developed using the original dataset had a MAE of 18.70, whereas the same was the case of the expressions provided using the AdaBoost, Gradient Boosting, Neural Network, and Random Forest prediction, giving 19.85, 16.96, 17.02, and 17.23 respectively. It is interesting to note that the symbolic model trained on Gradient Boosting predictions had lowest error, which is why it best described the data patterns in the interpretable format. These results hint at the potential utility of symbolic regression as a means of modifying the utility of black-box ML systems into an interim version between them and more easily interpretable equation-based engineering systems.

4. Discussion

The article conducted an effective study on various machine learning algorithms in respect to predicting the compressive strength (CS) of the Ultra-High-Performance Concrete (UHPC) in relation to a very large number of input variables. Across all tested models, Gradient Boosting displayed the best performance with R 2 of 0.970, the smallest MAE of 6.98 and the lowest RMSE of 8.08, which entails that it is able to efficiently model non-linear patterns and reduce the error in predictions. AdaBoost and Random Forest also performed well with R 2 values being 0.961 and 0.960 respectively and RMSE values 7.82 and 8.08 respectively. That property of many weak learners was the power used to affect the stability and generalization of such ensemble models. The Neural Network (Multilayer Perceptron) performed quite well with an R 2 of 0.947, however, and an RMSE of 9.25 placed it marginally behind the ensemble models. On the contrary, Decision Tree, k-Nearest Neighbors, and Linear Regression models performed orders of magnitude less successfully. The Decision Tree had a low R especially compared to KNN and LR which had R of 0.790 and 0.685 at large with the same having high RMSE and MAE as it failed to capture the high-order nonlinearities and interactions on UHPC datasets.

In addition to predictive accuracy, interpretable was handled in terms of symbolic regression with MEPX. Directly generated symbolic model using the original dataset yielded a greater RMSE= 18.7 leading to the fact that symbolic model faces difficulty of generating raw, non-iteratively optimizable complex relationships as with ensemble methods. Nevertheless, it is advantageous to provide a lucid, human-writable equation that reflects the literally physical interaction of the variables, which is of vital essence to the realms of engineering. Interesting to note, when MEPX was applied to approximate the top 3 models (Gradient Boosting, AdaBoost, Random Forest, Neural Network), it generated symbolic expressions that were in matching error performance to that of the models, with 16.96, 17.02, 17.23 and 19.85 in error metrics of Gradient Boosting, Neural Network, Random Forest, and AdaBoost, respectively. This demonstrates that strong symbolic regression can be the intermediate technology between black-box inference and engineering elucidation in that the meaningful approximations are kept but the causal structure of data-driven elucidation becomes explicit. Thus, the hybrid approach that integrates machine learning and symbolic modeling, not only advances the predictivity but also advances the trust and portability in the real-world engineering context requiring transparency.

5. Conclusion and Recommendations

In the present study, a hybrid forecasting technique has been developed in an attempt to predict the compressive strength of Ultra-High Performance Concrete (UHPC) using machine learning (ML) models and symbolic regression. Seven determinations of ML were used such as: Gradient boosting, AdaBoost, Random forest, Neural network, Decision tree, k-nearest neighbors and linear regression were researched on a mix of 810 experimental samples. Every model was evaluated with a five-fold stratified cross validation method in order to have a general assessment of the models and provide results that were not dependent on any particular validation set. The finest outcome was provided by the Gradient Boosting algorithm with R 2 of 0.970 and the smallest RMSE of 8.08, and close positions were made by AdaBoost, Random Forest, and Neural Network. Compared with the individual models, the qualities of the ensemble models commanded the advantage of a better model in capturing non-linear and complex interactions between UHPC input parameters and the result of compressive strength, bestowing an approving source of information on the data-driven design of constructions material engineers.

The objective of symbolic regression was to make the models easier to interpret and engineer usable, which was applied to Multi Expression Programming X (MEPX). This method was used in respect to the initial dataset as well as in respect to the forecasts of the four best ML models. The interpretable mathematical equations that were deduced by the symbolic models showed how major input variables, including silica fume, level of superplasticizer, amount of cement, and curing age affected UHPC compressive strength. In the original dataset based symbolic regression equation, the results obtained were an MAE of 18.70 MPa and that of the Symbolic model obtained based on Gradient Boosting also decreased the MAE to 16.96 MPa. The MAEs obtained with the similar symbolic models based on AdaBoost, Neural Network, and Random Forest were 19.85, 17.02, and 17.23 MPa respectively which also confirms the ability of MEPX to approximate the ML predictions in more fungible mathematical form. Such symbolic equations can be used effectively by structural designers who would need to make quick estimations on compressive strength without having to use a lot of computational resources or even having to conduct a massive amount of tests.

The results of the present research emphasis on the importance of using an ML with both high accuracy and transparency represented in symbols to predict concrete materials. The suggested workflow will provide engineers with efficient, but comprehensible tools, which will mitigate the gap between AI models and using it practically in the structural design. In order to prepare this investigation repeatable in the future, appreciating the recognition of extending the list of mix design variants, and curing conditions, as well as mechanical or durability properties obscure UHPC, the authors would like to recommend the expansion of the database on the interest of more mix design varieties, and curing conditions, as well as mechanical or durability properties. In addition, hybrid approaches incorporating domain knowledge into the symbolic forms or regularized ML pipelines would be worth investigation to have a greater trade-off dynamism between accuracy and interpretability. Last but not least, the implementation of such predictive framework into building information modeling (BIM) systems or quality control procedures at construction sites might play a major role in the enhancement of productivity, savings in tests, and the achievement of higher confidence in high-performance concrete structures.

References

1. Mengesha, G. (2025). ULTRA-HIGH-PERFORMANCE CONCRETE (UHPC/UHPFRC) FOR CIVIL STRUCTURES: A COMPREHENSIVE REVIEW OF MATERIAL INNOVATIONS, STRUCTURAL APPLICATIONS, AND FUTURE ENGINEERING PERSPECTIVES. STRUCTURAL APPLICATIONS, AND FUTURE ENGINEERING PERSPECTIVES (May 14, 2025).
2. Du, J., Meng, W., Khayat, K. H., Bao, Y., Guo, P., Lyu, Z., ... & Wang, H. (2021). New development of ultra-high-performance concrete (UHPC). *Composites Part B: Engineering*, *224*, 109220.
3. Alsalman, A., Dang, C. N., Martí-Vargas, J. R., & Hale, W. M. (2020). Mixture-proportioning of economical UHPC mixtures. *Journal of Building Engineering*, *27*, 100970.
4. Singniao, P., Sappakittipakorn, M., & Sukontasukkul, P. (2020, July). Effect of silica fume and limestone powder on mechanical properties of ultra-high performance concrete. In *IOP Conference Series: materials Science and Engineering* (Vol. 897, No. 1, p. 012009). IOP Publishing.
5. Redžić, N., Grgić, N., & Baloević, G. (2025). A Review on the Behavior of Ultra-High-Performance Concrete (UHPC) Under Long-Term Loads. *Buildings*, *15*(4), 571.
6. Li, W., Huang, Z., Cao, F., Sun, Z., & Shah, S. P. (2015). Effects of nano-silica and nano-limestone on flowability and mechanical properties of ultra-high-performance concrete matrix. *Construction and Building Materials*, *95*, 366-374.
7. Qian, Y., Yang, J., Yang, W., Alateah, A. H., Alsubeai, A., Alfares, A. M., & Sufian, M. (2024). Prediction of ultra-high-performance concrete (UHPC) properties using gene expression programming (GEP). *Buildings*, *14*(9), 2675.

8. El-Abbasy, A. A. A. (2025). Artificial intelligence-driven predictive modeling in civil engineering: a comprehensive review. *Journal of Umm Al-Qura University for Engineering and Architecture*, 1-24.
9. Kavzoglu, T., & Teke, A. (2022). Predictive performances of ensemble machine learning algorithms in landslide susceptibility mapping using random forest, extreme gradient boosting (XGBoost) and natural gradient boosting (NGBoost). *Arabian Journal for Science and Engineering*, *47*(6), 7367-7385.
10. Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, *77*, 29-52.
11. Zhang, Y., Li, Y., Li, Y., Zhao, L., & Yang, Y. (2025). Interpretable Machine Learning Models and Symbolic Regressions Reveal Transfer of Per-and Polyfluoroalkyl Substances (PFASs) in Plants: A New Small-Data Machine Learning Method to Augment Data and Obtain Predictive Equations. *Toxics*, *13*(7), 579.
12. Luo, J., & Yu, C. L. (2023). *The Application of Symbolic Regression on Identifying Implied Volatility Surface. Mathematics 11, 9 (2023)*.
13. Ali, M., Chen, L., Qureshi, Q. B. A. I. L., Alsekait, D. M., Khan, A., Arif, K., ... & Khan, M. (2024). Genetic programming-based algorithms application in modeling the compressive strength of steel fiber-reinforced concrete exposed to elevated temperatures. *Composites Part C: Open Access*, *15*, 100529.
14. Makke, N., & Chawla, S. (2025). A Perspective on Symbolic Machine Learning in Physical Sciences. *arXiv preprint arXiv:2502.17993*.
15. Kashem, A., Karim, R., Malo, S. C., & Das, P. (2023). Ultra-high-performance concrete (UHPC) [Data set]. Mendeley Data, 1, 10.17632/85r7bh4zsz.1.
16. Sweet, L. B., Müller, C., Anand, M., & Zscheischler, J. (2023). Cross-validation strategy impacts the performance and interpretation of machine learning models. *Artificial Intelligence for the Earth Systems*, *2*(4), e230026.
17. Shmuel, A., Glickman, O., & Lazebnik, T. (2024). Symbolic regression as a feature engineering method for machine and deep learning regression tasks. *Machine Learning: Science and Technology*, *5*(2), 025065.
18. Alsalman, A., Kareem, R., Dang, C. N., Martí-Vargas, J. R., & Hale, W. M. (2022, January). Prediction of modulus of elasticity of UHPC using maximum likelihood estimation method. In *Structures* (Vol. 35, pp. 1308-1320). Elsevier.
19. Tarka, P. (2018). An overview of structural equation modeling: its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & quantity*, *52*(1), 313-354.
20. Kashem, A., Karim, R., Malo, S. C., Das, P., Datta, S. D., & Alharthai, M. (2024). Hybrid data-driven approaches to predicting the compressive strength of ultra-high-performance concrete using SHAP and PDP analyses. *Case Studies in Construction Materials*, *20*, e02991.

Funding: This research received no external funding.

Data Availability Statement: The data is available on reasonable request to the corresponding author.

Acknowledgments: The authors would like to acknowledge the Department of Civil Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology (GIKI), for providing administrative and tech nical support throughout the course of this research.

Conflicts of Interest: The authors declares no conflicts of interest