Fusion Journal of Engineering and Sciences

Article

Formulation of Penetration Resistance, Softening Point and Viscosity of Plastic Modified Bitumen using Genetic Expression Programming.

Maaz Khan[1]*, Usama Asif[2], Touqeer Ali Rind[1], Muhammad Farooque Panhyar[3]

1 Department of Civil Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, KPK, Pakistan. Email: gcv2463@giki.edu.pk; touqeer.ali@giki.edu.pk

2 Department of Civil Engineering and Environmental Engineering, University of Maryland, College Park, US. Email: uasif@umd.edu

3 Department of Engineering, Institute of Business Administration University, Sukkur, Pakistan. Email: *muhammad-farooq.panhyar@gmail.com*

***** Correspondence Author: *gcv2463@giki.edu.pk*

Abstract

Use of plastic waste in bitumen is a sustainable and green technique which can improve asphalt properties and also solve the global issue of plastic waste. The focus of the present study was on the new application of Genetic Expression Programming (GenEP) to create predictive models for some of the most important traditional properties of plastic-modified bitumen, penetration, softening point, and viscosity. Overall, a large database was created by reading literature which offered eleven input parameters (plastics and blending conditions) and then applying the GenEP technique to generate explicit mathematical equations for each of the properties. Then the models were validated using R-squared (R^2) value, Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE) and all the models were highly accurate with good generalized predictions on training, validation, and test datasets. To better understand the model predictions and also to find a value for each of the discrepancies of each input variable, the analysis included a Shapley Additive Explanations (SHAP) analysis. Information gained using the SHAP analysis also led to some interesting conclusions about the relative importance of some of the input parameters as well as increased transparency and support for close examination of the models. The study confirms GenEP with SHAP analysis as a suitable modelling predictive method for asphalt research and supports ongoing use of plastic waste in road construction as part of a sustainable infrastructure development process.

Keywords: Genetic expression programming (GenEP), plastic waste, asphalt mix, plastic waste asphalt (PWA).

1. Introduction

Asphalt is a popular material used for road construction and is endowed with a lot of advantages [1]. Its durability and longevity are the biggest advantages. Asphalt has the capacity to withstand heavy traffic and climatic conditions owing to its hardness and strength. Apart from being hard, asphalt can also produce a smooth surface of the road for cars in order to facilitate a comfortable journey and reduce noise pollution [2, 3]. Asphalt is also simple to repair and keep up too that makes asphalt an economical choice for road construction work. Asphalt is a ductile material too that expands and contracts along with the temperature so does not crack or form potholes [4, 5]. Therefore, asphalt roads need less maintenance and repair in the long term. Another benefit of asphalt is its environmentally friendly characteristics. Asphalt is recyclable and reusable, a decrease in new material usage, and saving natural resources. Asphalt is also convenient to use, a quick and efficient method of road construction work. Its quick and simple process of installation minimizes obstructions and delays [6]. Overall, asphalt's strength, comfort, affordability, and environmental advantages are an extremely desirable option for road construction.

Utilization of various forms of waste in asphalt has gained popularity in recent years to minimize the waste being deposited in landfills, lower the cost of overall asphalt production, and ensure more environmentally friendly construction processes [7]. Among the various waste products employed in asphalt, some of them include recycled asphalt pavement (RAP), recycled asphalt shingles (RAS), and industrial by-products such as fly ash, slag, and blast furnace slag. These industrial by-products are generated in cement, steel, and other industrial product manufacturing processes, and they can be recycled and used for asphalt production [8, 9]. Utilization of industrial by-products in asphalt is advantageous in several ways. It minimizes the quantity of material going to landfills since these by-products would be use as a waste otherwise. It also saves energy on the manufacture of asphalt as such by-products tend to be less expensive than virgin materials. Besides this, the utilization of industry by-products in asphalt can enhance the overall quality of the pavement since they are able to enhance the stiffness and fatigue strength of asphalt. While there are some limitations of the utilization of by-products in asphalt, including maintaining the quality and homogeneity of the by-products, the advantages exceed the disadvantages in most situations [10]. Consequently, the application of waste in asphalt will most likely increase in popularity over the next few years.

Plastic waste has been a major environmental issue in recent times, with billions of tons of plastic being taken to landfills or into the ocean annually [9]. Nevertheless, scientists have been trying to find means of recycling plastic waste in a manner that does not destroy the environment as much. It is possible to use plastic waste in bitumen [11]. There are several potential benefits to using plastic trash in bitumen [12]. One of the main benefits is that it can reduce the amount of plastic trash that ends up in landfills or the ocean [13]. This is because the plastic trash is being reused and integrated into another product, rather than being discarded [14]. Additionally, the use of plastic waste in bitumen can potentially reduce the overall carbon intensity in the process of asphalt making because it can reduce the use of traditional petroleum-based bitumen [15]. It was established that the application of plastic waste in bitumen is likely to lower greenhouse gas emissions when compared to conventional asphalt production processes [16]. Nevertheless, the extent of such benefits depends on the plastic type used, the mixing conditions, and the asphalt mix. Therefore, the creation of a model to predict the potential effects of wet or dry-modified plastic asphalt blends would be beneficial, as it would provide assurance in the plastic addition's benefits.

The use of machine learning techniques (support vector machine, random forest, and boosted regression tree) to predict the properties of asphalt has gained attention in recent years due to the potential for improved efficiency and accuracy in asphalt design and production. Machine learning models [17-20]; soft computing techniques such as neural networks (ANN), gene expression programming (GEP), multivariate adaptive regression spline (MARS), adaptive neuro-fuzzy inference system (ANFIS), and random forest (R.F.) [21-24] has been used by various researcher to detect and predict [25-29] the results based on existing literature. Normally, fuzzy algorithm [30-33]; ANFIS hybrid model [34, 35]; error minimization technique [36]; hybrid ANN [37] ; gene expression programming [38] Bayesian machine learning models [23, 39] and deep learning methods [40, 41] are used for different purposes. As Asphalt is a complex material, and its properties are influenced by a wide range of factors, including the type and proportions of the raw materials used, the production process, and the environmental conditions in which it is used. Consequently, the accurate prediction of asphalt properties can be a difficult task. One such method that has been attempted for predicting the properties of asphalt through machine learning is the creation of artificial neural networks (ANNs). ANNs have been employed to predict a range of properties such as rutting resistance, fatigue resistance, and tensile strength of asphalt mixtures. The outcome of the study indicated that the ANNs could predict the rutting resistance of the asphalt mixtures with less than 3% error. Support vector machines (SVMs) are another machine learning method that has been employed to predict the characteristics of asphalt. SVMs have been applied in predicting the rutting resistance, fatigue resistance, and tensile strength of asphalt mixtures using the type and amount of the raw materials utilized. Application of machine learning methods has the potential to enhance the efficiency and quality of asphalt production and design through enabling the forecasting of the asphalt mixture properties based on the proportion and type of the raw materials employed. Additional research is required to achieve complete exploitation of the capabilities of these methods in this application.

However, as per authors knowledge, no research has been conducted to predict the basic properties conventional properties like penetration, softening point and viscosity of plastic modified bitumen. Hence, this study is conducted to develop equations for these basic properties of plastic modified bitumen using novel machine learning technique known as Genetic expression programming (GenEP). In section 2, a brief introduction and methodology of GenEP is explained.

In section 3, detailed methodology about data collection and modeling parameters are given. In section 4, results are presented with SHAP analysis in section 5 while conclusions are given in section 6 of this article.

2. Genetic Expression Programming

Genetic programming (GP) is a form of artificial intelligence that uses evolutionary algorithms to generate computer programs that perform a desired task. It was first introduced by John Koza in the 1980s to automatically generate computer programs that can solve problems without being explicitly programmed to do so. GP works by representing computer programs as trees of functions and terminal nodes, which are then evolved through a process of selection, crossover (recombination), and mutation. The objective of GP is to identify the optimal program that can solve a problem using a fitness function to compare the programs and choose those which execute best to be reproduced. GenEP is one type of GP that employs a genome to symbolize a program, with the genome being composed of genes representing functions, terminals, and the program structure. GenEP employs a transcription, translation, and mutation process to create programs based on the genome. The primary benefit of GenEP is that it can produce more complex programs than GP is able to since it can pack more information in the genome. But GenEP has greater computational complexity because it has additional transcription and translation steps. GenEP was presented by Ferreira in 2002 and has been used to solve a range of problems, such as function approximation, pattern recognition, and control systems.

1. The GenEP procedure can be divided into the following steps:

Define the problem: In the initial step of GenEP, the problem to be solved by the generated programs needs to be defined. This involves declaring the inputs, outputs, and constraints or restrictions on the programs.

2. Identify the function set and terminal set: The function set is an enumeration of functions that can be applied in the programs, e.g., arithmetic operations and logical operators. The terminal set is an enumeration of terminal nodes, which are either constants or variables employed as inputs to the programs.

3. Initialize the population: The second step is to generate an initial population of genomes, which will serve as the basis for the evolution process. This may be achieved through random initialization, in which the genes of the genome are randomly sampled from the function and terminal sets or through a heuristic approach for generating more competent genomes.

4. Assess the genomes' fitness: The fitness of a genome is how well it can solve the problem that has been defined. This is commonly achieved by converting the genome into a program, executing the program on a set of test cases, and determining a fitness score based on the program's output accuracy.

5. Select genomes for reproduction: The second step is to choose the genomes which are to be used for reproduction. This is usually carried out by employing a selection algorithm like tournament selection in which a random set of genomes is picked, and the best one is chosen for reproduction, or roulette wheel selection in which genomes are chosen to proportionate to their relative fitness.

6. Carry out crossover and mutation: Crossover, or recombination, is the operation of mixing the genes of two genomes to produce a new genome. This is carried out by choosing a crossover position at random and exchanging the genes on either side of the position between the two genomes. Mutation is the operation of randomly changing genes in a genome. This is accomplished by substituting a gene with a member of the function or terminal set, or by changing the value of a terminal.

Repeat steps 4-6: The process of evaluating the fitness of the genomes, selecting genomes for reproduction, and performing crossover and mutation is repeated for a specified number of generations. The goal is to find the best program that can solve the defined problem using a fitness function to evaluate the programs and selecting the ones that perform the best for reproduction.

Test the best genome: Once the evolutionary process has completed, the fittest genome is transcribed into a program and tested on a set of unseen test cases to evaluate its performance.

Conventional machine learning techniques like Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Random Forests (RF) have been most applied in asphalt modeling, these work generally as black-box models and fail to provide explicit mathematical equations. GenEP, on the other hand, derives interpretable closed-form equations that are beneficial for engineering design and decision-making purposes. In addition, its evolutionary character renders it easily adaptable to intricate, nonlinear relationships characteristic of plastic-modified bitumen systems. The innovation in this research is not only the utilization of GenEP, but also its application in predicting base properties (penetration, softening point, and viscosity) of plastic-modified bitumen that, to the best of our knowledge, has not been studied before.

3. Methodology

3.1 Database Generation

To use machine learning algorithms to generate the proper prediction models, data collection is the first step. Collection and generation of reliable data for any model is the most tedious task in machine learning. In this study, a huge database is collected from vast literature review. Out of 14 variables, 11 were taken as input and 3 were taken as output. Input variables included in this study are mixing time (MixTi), speed (Revolutions Per Minute (RMP)), and temperature at which plastics were mixed with asphalt (MixT), size of added plastic (PlSi), percentage of added polyethylene terephthalate (PET), polyethylene (PE), polypropylene (PP), polystyrene (PS), polyvinyl chloride (PVC), styrene-butadiene-styrene (SBS), and crumb rubber (CR). While change in penetration (CiP), change in softening point (CiSP) and change in viscosity (CiV) were taken as output variables. A total of 267, 265, and 280 data points are collected for CiP, CiSP, and CiV. It is important to mention at this point that input variables were selected after vast literature review. As most of the research is carried out with addition of some other additives with plastics or different types of plastic, different kinds of plastics were considered and combined in this research to make the beneficiary circle large. As the performance of all machine learning models is affected by distribution of data inside the database, the distribution of inputs and outputs were checked by descriptive statistics (Table 1-3) and histograms (Figure 1). The dispersion of the input variables is not homogeneous, and the frequencies of the factors are proportionally greater, as shown in Figure 1. It should be considered that variables with high frequencies can lead to a more accurate model. It is advised that the suggested formulations be used to this given range to achieve accurate estimates of mechanical characteristics.

Table 1 Statistical analysis for CiP

	PE	PET	PP	PS	PVC	CR	MixT	Speed (RPM)	MixTi	PlSi	CiP
Mean	0.03	0	0	0	0.01	0.01	172.17	2905	1.32	4.93	-0.43
Standard Error	0.00	0	0	0	0.00	0.00	0.46	141	0.04	0.23	0.01
Median	0.03	0	0	0	0.00	0.00	173.72	2956	1.25	4.62	-0.42
Mode	0.00	0	0	0	0.00	0.00	180.00	4000	1.00	2.50	-0.64
Standard Deviation	0.03	0.02	0.01	0.01	0.03	0.03	7.49	2319	0.62	3.70	0.20
Sample Variance	0.00	0	0	0	0.00	0.00	56.11	5380832	0.38	13.66	0.04
Kurtosis	4.01	65.22	21.39	92.15	40.3	7.16	-0.04	4.9	4.73	1.57	-0.47
Skewness	1.61	7.62	4.44	9.59	5.95	2.82	-0.52	1.5	1.94	1.29	-0.26
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	150.00	60	0.08	0.00	-0.94
Maximum	0.20	0.20	0.07	0.07	0.30	0.15	190.00	13000	4.00	15.00	-0.03
Count	267	267	267	267	267	267	267	267	267	267	267

Table 2 Statistical Analysis for CiSP

	PE	PET	PP	PS	PVC	CR	MixT	RPM	MixTi	PlSi	*CiSP*
Mean	0.03	0.00	0.00	0.00	0.01	0.01	173.12	3177.00	1.42	4.96	0.38
Standard Error	0.00	0.00	0.00	0.00	0.00	0.00	0.46	136.35	0.04	0.23	0.02
Median	0.02	0.00	0.00	0.00	0.00	0.00	174.65	3117.68	1.17	4.65	0.25
Mode	0.00	0.00	0.00	0.00	0.00	0.00	180.00	4000.00	1.00	2.50	0.02
Standard Deviation	0.03	0.02	0.01	0.01	0.03	0.03	7.34	2198.61	0.71	3.74	0.40
Sample Variance	0.00	0.00	0.00	0.00	0.00	0.00	53.92	4833892.86	0.51	13.96	0.16
Kurtosis	4.81	63.43	20.72	89.65	39.20	6.86	0.58	6.00	2.35	1.46	8.43
Skewness	1.69	7.52	4.37	9.46	5.86	2.77	-0.87	1.69	1.46	1.27	2.40
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	150.00	60.00	0.08	0.00	-0.06
Maximum	0.20	0.20	0.07	0.07	0.30	0.15	190.00	13000.00	4.01	15.00	2.67

Table 3 Statistical analysis for CiV

	PE	PET	PP	PS	PVC	CR	MixT	RPM	MixTi	PlSi	CiV
Mean	0.03	0.00	0.00	0.00	0.00	0.01	172.71	3229.52	1.49	2.38	3.32
Standard Error	0.00	0.00	0.00	0.00	0.00	0.00	0.61	126.92	0.04	0.09	0.31
Median	0.02	0.00	0.00	0.00	0.00	0.00	180.00	3421.43	1.25	2.19	1.88
Mode	0.00	0.00	0.00	0.00	0.00	0.00	180.00	5000.00	1.00	0.00	0.00
Standard Deviation	0.03	0.02	0.01	0.01	0.01	0.03	10.15	2123.74	0.61	1.43	5.22
Sample Variance	0.00	0.00	0.00	0.00	0.00	0.00	102.98	4510266.11	0.37	2.05	27.27
Kurtosis	-0.64	46.21	24.35	111.01	15.88	8.29	-0.81	6.15	0.73	-0.73	32.37
Skewness	0.60	6.46	4.72	9.48	4.07	2.98	-0.48	1.50	1.20	0.04	4.79
Minimum	0.00	0.00	0.00	0.00	0.00	0.00	150.00	120.00	0.50	0.00	-0.06
Maximum	0.10	0.15	0.08	0.15	0.05	0.15	190.00	13000.00	3.00	5.15	49.75

Table 4 Variable Description Table

Symbol	Full Form	Unit
PE	Polyethylene Content	Percentage (%)
PET	Polyethylene Terephthalate Content	Percentage (%)
PP	Polypropylene Content	Percentage (%)
PS	Polystyrene Content	Percentage (%)
PVC	Polyvinyl Chloride Content	Percentage (%)
CR	Crumb Rubber Content	Percentage (%)
MixT	Mixing Temperature	Degrees Celsius (°C)
Speed (RPM)	Rotational Speed	Revolutions per minute (RPM)
MixTi	Mixing Time	Minutes (min)
PlSi	Plastic Size	Millimeters (mm)
CiP	Change in Penetration	Decimillimeters (dmm)
CiSP	Change in Softening Point	Degrees Celsius (°C)
CiV	Change in Viscosity	Pascal-seconds (Pa·s)

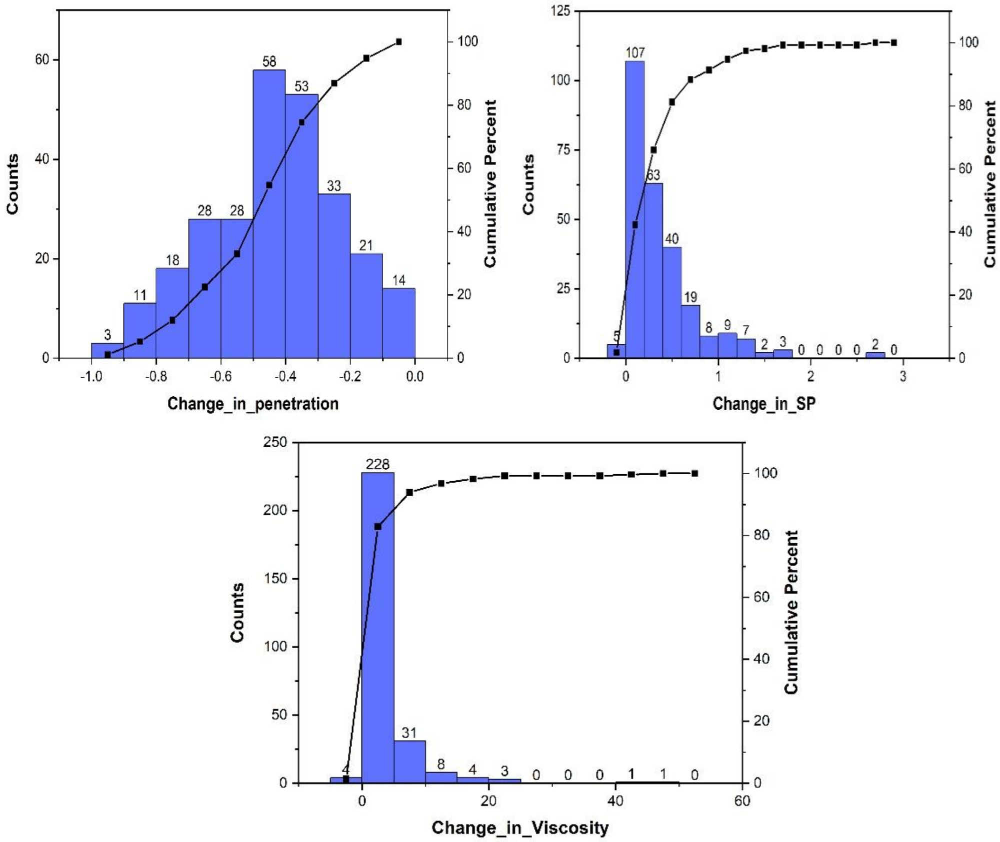

Figure 1 Distribution of Output Variables

The study conducted multiple tests to ensure the accuracy and reliability of the database. Data sets that differed greatly from the overall pattern were excluded from the model building and assessment. The database was separated randomly into three groups: training, validation, and testing. The training data was used to train the model through GenEP, the validation data was used to check the model's ability to apply to new data, and the testing phase involved using the model on previously unseen data.

One of the biggest challenges in implementing artificial intelligence-based techniques is multi-collinearity. This problem happens due to the interdependence of different input parameters. This, in turn, can reduce the effectiveness of the established model. To prevent this problem, it has been recommended that the correlation coefficient (R) among two input variables should be kept below 0.8. To ensure that the model developed is not affected by multi-collinearity, the R value for all likely patterns of input variables is calculated. As shown in Figure 2, it can be inferred that R values for all combinations, whether positive or negative, fall well below the 0.8 threshold. This indicates that there is little risk of multi-collinearity among the variables during the modeling process, and it guarantees that the model developed will be efficient and accurate. One of the best ways to avoid multi-collinearity is by selecting the right set of features or independent variables, by analyzing their correlation among each other or implementing feature selection techniques. The importance of avoiding multi-collinearity cannot be overstated as it can lead to incorrect estimation of regression coefficients, unstable or unreliable results and misinterpretation of results.

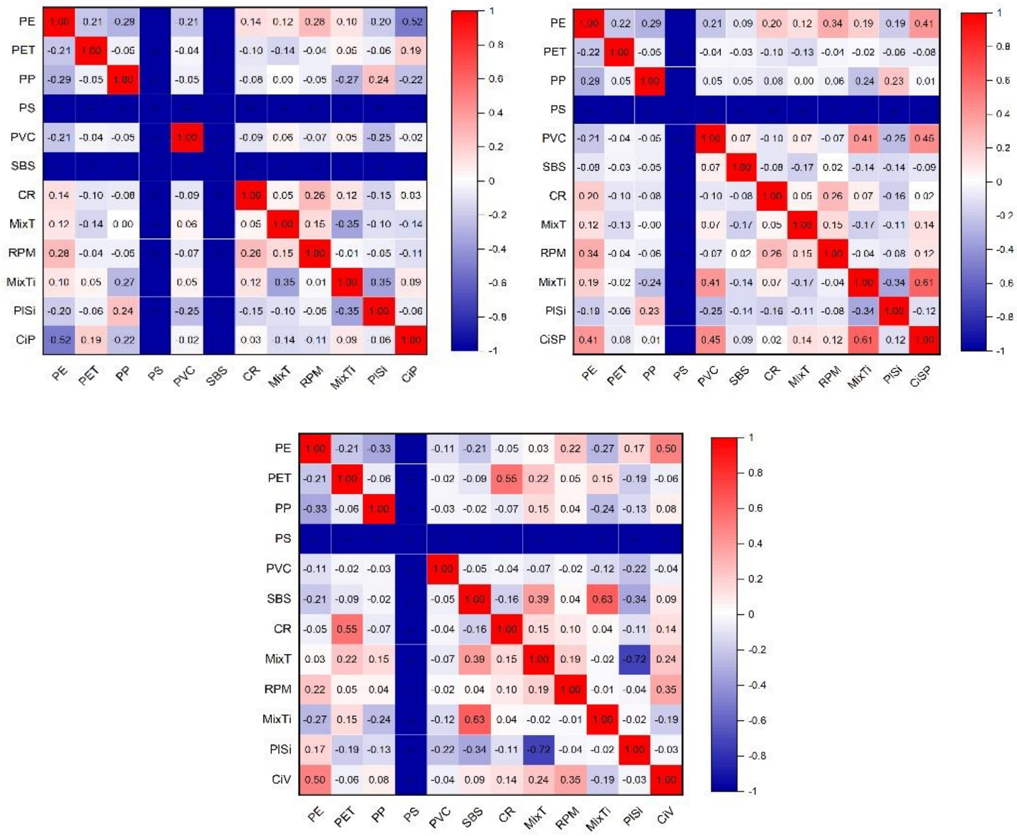

Figure 2: Co-efficient of correlations

3.2 Development of Model

Prior to model creation, the initial step is to identify input factors that might influence plastic modified bitumen attributes. To find the influential variables on the attributes of plastic modified bitumen for the construction of a generalized relationship, every variable in the database was investigated systematically, and the performance of many preliminary runs was assessed. Consequently, the mechanical characteristics of CMWFS are seen as a function of the below dependent variables: Eq. (1).

$$CiP, CiSP, and\ CiV = f(PE, PET, PP, PS, PVC, SBS, CR, MixT, speed(RMP), MixTi, PlSi)$$

Equation 1

The selection of parameters for GenEP is a crucial step in the process of developing accurate and reliable models. To make the best performance, the parameters must be selected suitably. The first parameter to choose is the population size, which determines the number of agents in the population at any given time. The bigger the population size, the greater the likelihood of obtaining a good solution, but it also increases the computational cost. It is usually a good idea to begin with a population size of approximately 50-100 and then scale according to the problem's complexity. The next parameter to choose is the number of generations, which determines how many times genetic operators are applied to the population. The more the number of generations, the greater the possibility of obtaining a good solution, but computational cost is greater. It is usually best to begin with about 50-100 generations and thereafter vary as needed depending on the nature of the problem. The other crucial parameter is the selection method which determines how the individuals should be selected to reproduce. Popular selection methods are tournament selection, roulette wheel selection, and ranking selection. Every choice technique has strengths and weaknesses, and the choice technique should be chosen according to the type of problem. Another important parameter is the crossover rate and mutation rate, which define the probability that the new individual is to be generated by using crossover or mutation. A large crossover rate will imply that more information will be inherited from the two parents and a large mutation rate will imply more population diversity. Both rates are generally selected around 0.8-0.9. Finally, the representation of chromosomes and the kind of genetic operators should be considered. The representation, i.e., tree-based or linear, will depend on the problem. The appropriate kind of genetic operators, i.e., one-point or two-point crossover, should also be determined dependent on the problem. Generally, it is a crucial step in the process of obtaining correct and valid models to select the GenEP parameters. The selected parameters for all 3 output variables utilized in this study are shown in Equation 1.

The correlation coefficient is a typical performance metric (R). Due to R's insensitivity to division and multiplication of target value by a constant, it cannot be used as the primary gauge of the model's predictive accuracy. There are several formulas that can be used to assess the prediction capacity of machine learning algorithms in concrete.

Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the sum of the absolute differences between predictions and actual values, divided by the number of instances.

Root Mean Squared Error (RMSE) is the square root of the MSE, which is more interpretable in terms of the units of the response variable.

Mean Squared Error (MSE) is like MAE, but it squares the differences between predictions and actual values, which gives more weight to larger errors. It is the average of the squared differences between predictions and actual values.

R-Squared (R^2) is a measure of how well the predictions of a model fit the actual data. It ranges from 0 to 1, with a higher value indicating a better fit. It is the proportion of the variance in the response variable that is explained by the model.

The formulae for the above-mentioned properties are given below.

$$MAE = \frac{\sum_{i=1}^{n}(|ex_i - p_i|)}{n} \qquad \text{Equation 2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(ex_i - p_i)^2}{n}} \qquad \text{Equation 3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(ex_i - p_i)^2 \qquad \text{Equation 4}$$

$$R^2 = \left(\frac{\sum_{i=1}^{n}(ex_i - \overline{ac}_i)(p_i - \bar{p}_i)}{\sqrt{\sum_{i=1}^{n}(ex_i - \overline{ac}_i)^2 \sum_{i=1}^{n}(p_i - \bar{p}_i)^2}}\right)^2 \qquad \text{Equation 5}$$

It is also important to keep in mind that high scores on a performance measure do not always guarantee a good model; it's always recommended to validate the model with unseen data, and also to interpret its predictions for a better understanding of the model's behavior.

4. Results and Discussions

4.1 Formulations for Penetration, Softening point and viscosity

The output generated by GenEP for penetration, SP and viscosity is decrypted to obtain mathematical formulas for the relevant property computation using all input parameters. The specific formulas are represented by Equations 6, 11, and 16 correspondingly. Comparing actual and forecasted penetration for all three different datasets (training, validation, and testing) is depicted in Figure 3. In addition, regression line expressions are also displayed on this graph. Ideally, the slope of the regression line should approach 1. Based on the slope values of for all three different datasets, it can be deduced that the constructed model contains a significant correlation between the measured and predicted values. Furthermore, the quantities are relatively similar and near to the ideal fit for all three datasets, showing that the model is well-trained and has a strong generalization capability, i.e., it performs similarly well on unknown data.

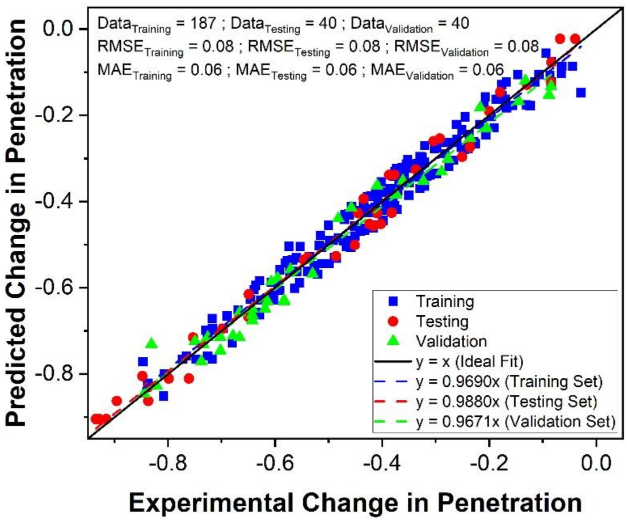

Figure 3: Comparison of CiP

$$CiP = A + B + C + D \qquad \text{Equation 6}$$

Where,

$$A = \frac{\big((PS \times MixT) + 4.1501\big) \times \big((0.3394 \times MixT) - PP\big)}{(10.7121 + RPM) \times (-12.3287 \times PlSi)} \qquad \text{Equation 7}$$

$$B = (-0.2226 - 2PE) \times \left(\left(\frac{PlSi}{2.6903} - PE\right) + 0.7762\right) \qquad \text{Equation 8}$$

$$C = \frac{\Big(\Big(\big(((15.5670 \times CR) + 2.8511) - (2.6309 + PlSi)\big) \times PP\Big) - PS\Big)}{MixTi} \qquad \text{Equation 9}$$

$$D = \Big(\big((MixTi \times PVC \times PlSi) \times (PS - PlSi)\big) - (MixTi - 5.0350) + \left(\frac{PVC}{-8.7548}\right)\Big) \times PVC \qquad \text{Equation 10}$$

Figure 4 depicts a comparison comparable to that made for the SP findings. The proposed model has been adequately trained on the data input to reliably predict the observed SP. The slope of the regression line for three different data sets, is near to the ideal value of 1. This indicates that the issue of model overfitting has been significantly reduced. In addition, the quality and usefulness of such simulation equations are greatly reliant on the quantity of data points included in the modelling. In the compiled database, the number of points were greater than 250, achieving a high degree of precision with few mistakes.

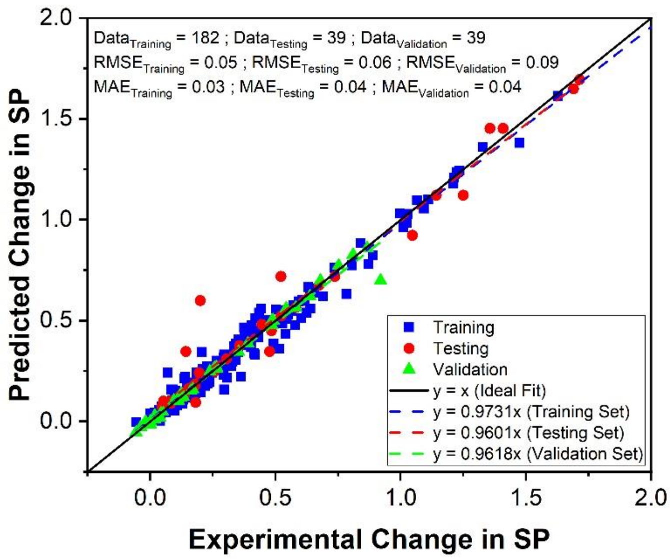

Figure 4: Comparison of CiSP

$$CiSP = A + B + C + D \qquad \text{Equation 11}$$

Where,

$$A = PE + \Big(\Big(\big((CR \times MixTi) + PE\big) + (CR - PVC)\Big) + (PVC \times MixTim^2)\Big) \qquad \text{Eq.12}$$

$$B = \left(\left(\frac{PP + PVC}{MixTim^2}\right) + PE\right) \times \Big(MixTi \times \big((PlSi - 4.8566) + 3.9898\big)\Big) \qquad \text{Equation 12}$$

$$C = \big((PET \times MixTim) \times (MixTim - 3PET)\big) \times MixTim \qquad \text{Equation 13}$$

$$D = \Big(5.9593 \times \big(SBS + ((PVC - 2CR))\big)\Big) + 6.6821 \qquad \text{Equation 14}$$

Graphical portrayal of the difference between predicted results and actual outcomes for viscosity is shown in Figure 5. This graph also includes the equations for regression lines between forecasted values and actual findings. Comparable to the models for CiP and CiSP this model works exceptionally well with test data. Figure 5 demonstrates that the proposed model correctly accounts for the effect of all input factors when predicting the viscosity. As seen by the gradient of the regression lines, the data presented in this figure exhibit a good correlation.

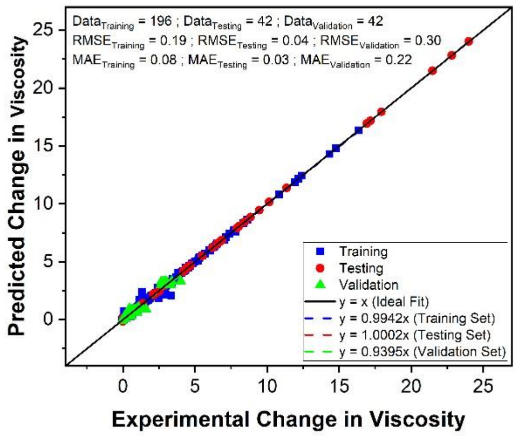

Figure 5: Comparison of CiV

$$CiV = A + B + C + D \quad \text{Equation 15}$$

Where,

$$A = SBS \times \left(\left(\left((PVC + 2.2697) + PVC^2\right) + \left((-1.8302 \times PET) \times 9.9066\right)\right) + PlSi\right) \quad \text{Equation 16}$$

$$B = \frac{(PlSI \times PP) \times (PlSi + MixTi)}{(MixTi - 1.9553) \times -0.1298} + SBS \times PVC \quad \text{Equation 17}$$

$$C = \left(63.4049 + \left(-3.0711 + \left((CR + SBS) \times RPM\right)\right)\right) \times PE \quad \text{Equation 18}$$

$$D = CR \times \left(-9.4528 + \left(561.1334 \times (PE \times -1.1472)\right)\right) \quad \text{Equation 19}$$

4.2 Performance evaluation of GenEP models

The quantity of data utilized in the formulation of a program is critical, as it directly influences the model's accuracy and generalizability. According to Frank and Todeschini (1994), the ratio of observations to input parameters should be greater than 5 to ensure reliable model development. In this study, the dataset was randomly divided into 70% for training, 15% for validation, and 15% for testing. Based on this division, the observation-to-variable ratios during the training phase were approximately 15.25 for penetration, 14.27 for softening point, and 15.56 for viscosity models, which comfortably exceeded the recommended threshold. During the testing phase, the corresponding ratios were 6.73, 6.00, and 6.22, respectively. These values indicate that the dataset size was sufficient to support the development of statistically robust and accurate predictive models using Genetic Expression Programming (GenEP).

As described previously, the performance of the developed models was assessed using standard statistical metrics including the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). Table 5 summarizes these metrics for the training, validation, and testing datasets for each of the CiP, CiSP, and CiV models. The relationship between predicted and experimental values is strong, with R^2 values of **0.980**, **0.985**, and **0.986** for CiP, CiSP, and CiV models during training, **0.976**, **0.987**, and **0.990** for validation, and **0.979**, **0.982**, and **0.990** for testing, respectively. Additionally, the low values of MAE, MSE, and RMSE across all datasets indicate that the models exhibit high predictive accuracy and strong generalization capabilities.

Table 5: Statistical calculations of models

	CiP			**CiSP**			**CiV**		
	Training	**Testing**	**Validation**	**Training**	**Testing**	**Validation**	**Training**	**Testing**	**Validation**
R^2	0.980	0.979	0.976	0.985	0.982	0.987	0.986	0.990	0.990
MAE	0.06	0.06	0.06	0.03	0.04	0.04	0.08	0.03	0.22
MSE	0.01	0.01	0.01	0.00	0.00	0.01	0.03	0.00	0.09
RMSE	0.08	0.08	0.08	0.05	0.06	0.09	0.19	0.04	0.30

To understand the statistics of absolute errors, the database is shown in Figure 6 along with the absolute errors for each data point. According to the picture, the mean error in the predicted values for CiP, CiSP and CiV are 0.028, 0.225 and 0.087, respectively, with a maximum error of no more than 0.119, 1.447 and 1.300. Only 6 out of 267 datapoints for CiP, 9 out of 265 datapoints for CiSP, 13 out of 280 datapoints for CiV or around 2.25%, 3.40%, and 4.64% of the total dataset, had an error higher than 0.05, 1.00, and 0.50, for CiP, CiSP, and CiV, respectively. The density of maximum errors occurrences is noticeably low, which must be noted. Inaccuracies of less than 0.04, 0.46 and 0.16, respectively, were found in 85% of CiP, CiSP, and CiV measurements.

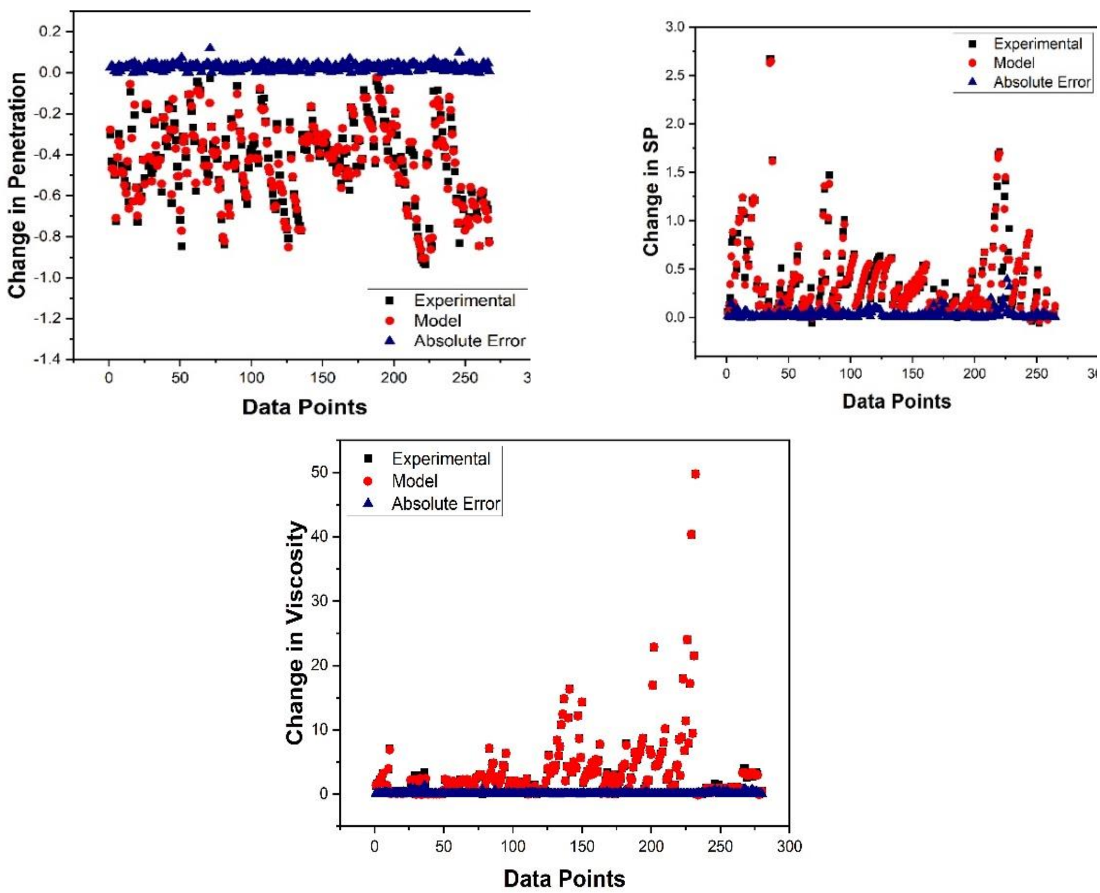

Figure 6: Absolute Errors of Models

Normalization of input variables was not performed in this research before model creation. The rationale behind this choice relied on the internal nature of Genetic Expression Programming (GenEP) itself, an evolutionary algorithm that does not depend on gradient-based optimization techniques compared to most standard machine learning models. GenEP does not evaluate fitness in terms of the absolute input scale of candidate solutions but instead based on their relative performance. Furthermore, the model could deliver high accuracy and low error rates with or without data transformation, as evident in the performance measures (R^2, MAE, RMSE). The input variables were, however, checked for multi-collinearity and for consistency of distribution to verify that scale differences did not skew the model's behavior.

5. Shap Analysis

Section 4's findings indicate that using a machine learning model can accurately predict the outputs. However, in a situation where conditions are not certain, the amount of each component in the mix can vary. Therefore, it is important to study how changes in individual input factors affect the model's output. Hence, the Shapley additive explanation (SHAP) method is used to assess the importance of every input parameter to their respective output. Shapley analysis is a method for determining the importance of individual factors in a complex system, named after Lloyd Shapley, a Nobel laureate in economics who developed the concept in the 1950s. The general principle of Shapley analysis is to attribute a "value" to each factor in a system according to its contribution to the result by averaging the marginal contribution of each factor to every possible coalition of factors. Shapley analysis has some benefits like it considers the interaction between factors, not merely the contribution of each factor individually, and is completely distributive, i.e., the overall worth of the system is shared by all the factors. This is significant because the total outcome of a system frequently results from the interactions among many factors, as opposed to the contribution of one factor. Through the incorporation of interactions among factors, Shapley analysis gives a truer representation of the significance of each factor within the system. Secondly, it is completely distributive, i.e., the system's total value is evenly distributed throughout all factors, as opposed to a subset of factors. This ensures that all factors are given fair credit for their contributions to the outcome, rather than just a select few. Hence, in this study, Shapley analysis is used to assess the contribution of different factors towards each output.

Figure 7 displays the overall significance factors for all input parameters with their respective outputs, and it is vital to note that the overall significance factors indicate the mean of the actual Shapley values for every attribute in the full sample set.

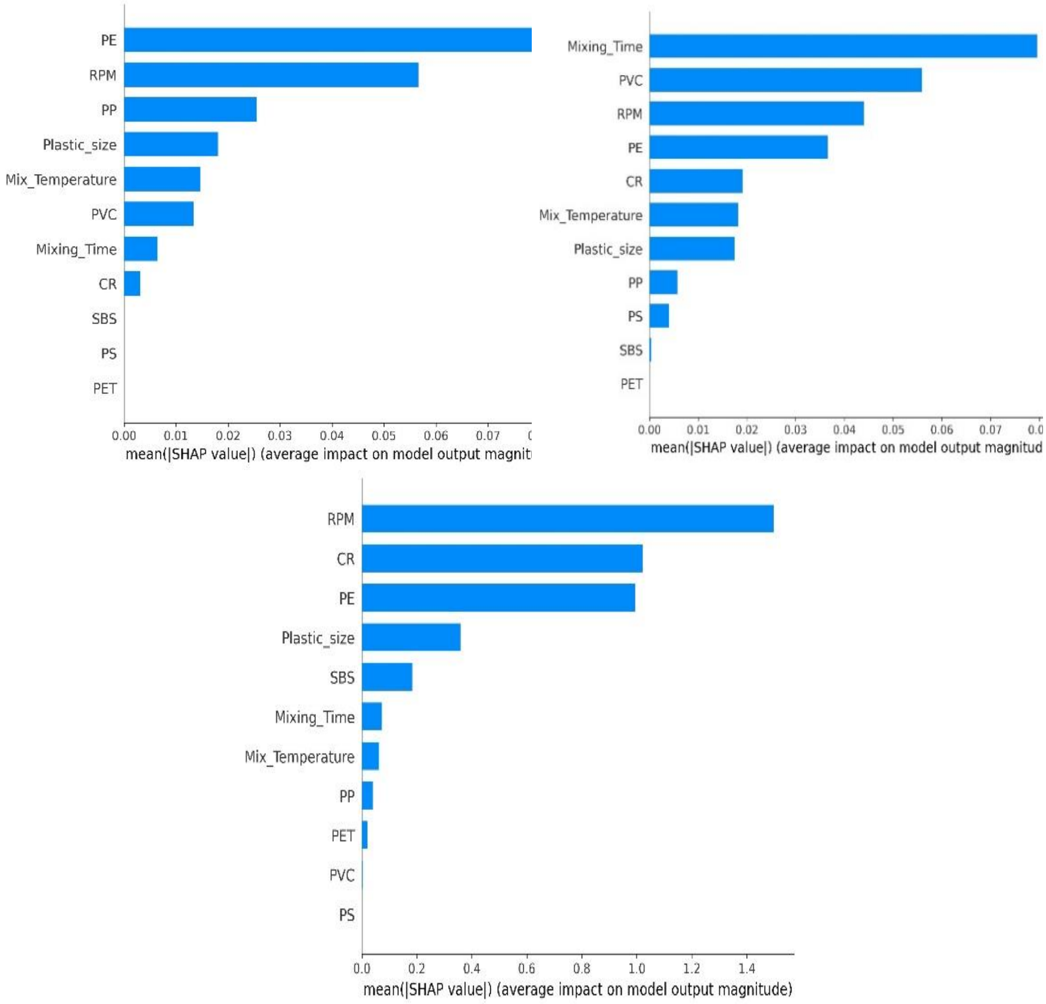

Figure 7: Overall significance factors

Each point in Figure 8 indicates a Shapely number for the input parameters and a single input. The penetration rises with the addition of PE, PP, PlSi, MixT, PVC, MixTi and CR, as shown in Figure 8. In contrast, speed (RPM) has a high influence on the output penetration and increasing these three components reduces the penetration. The same conclusions can be drawn for the remaining two outputs from Figure 8.

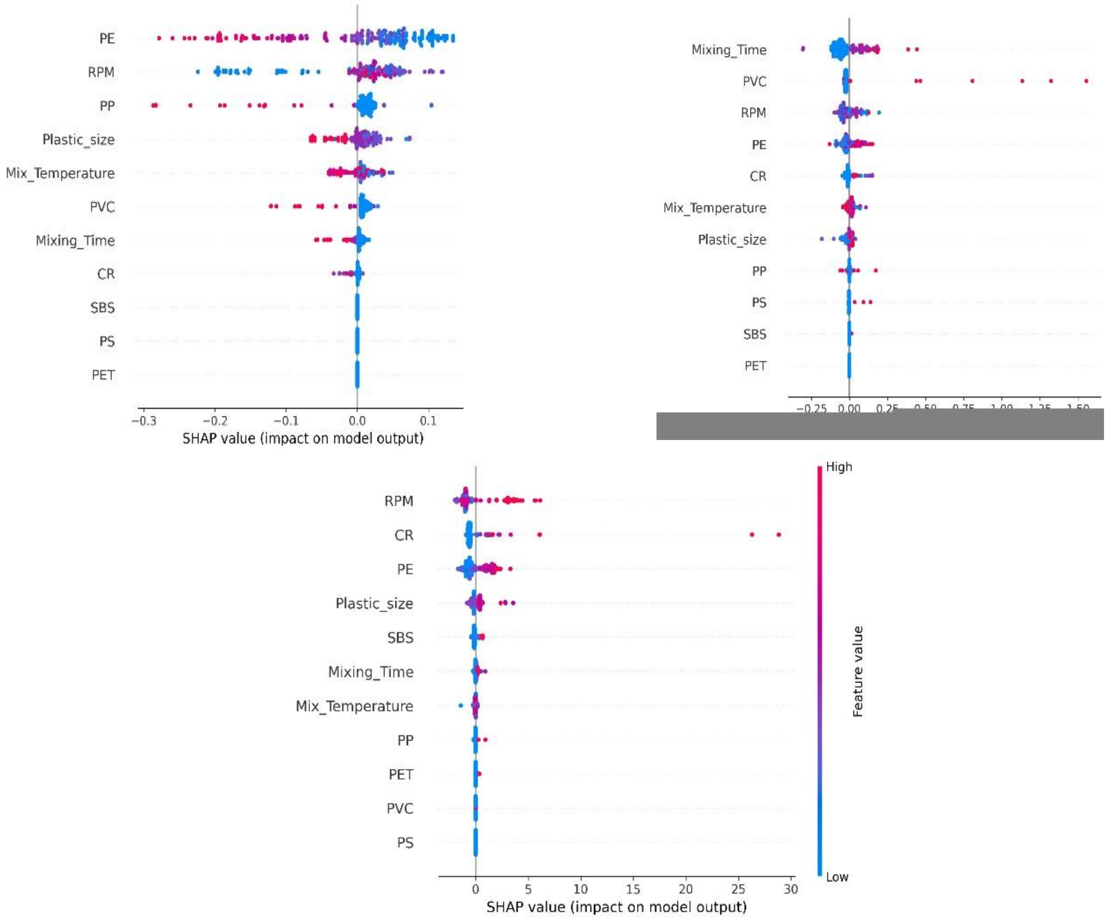

Figure 8: SHAP Values

6. Conclusions

In the present investigation, Genetic Expression Programming (GenEP) has been effectively utilized to create models for predicting the change in penetration (CiP), softening point (CiSP), and viscosity (CiV) of plasticized bitumen. The predictive accuracy of the developed models was found to be high, as confirmed by statistical performance metrics (R^2, MAE, MSE, RMSE) on training, validation, and test datasets. In addition, SHapley Additive exPlanations (SHAP) were utilized to explain the effect of every input variable, which increased the reliability and transparency of the results.The empirical models developed in this research provide a useful tool for predesigning and optimizing plastic waste asphalt (PWA) formulations and possibly eliminating the need for trial-and-error laboratory testing. The results can guide engineers and practitioners toward choosing suitable plastic modifiers and mixing conditions for realizing targeted bitumen properties and leading toward more sustainable pavement practice.

Subsequent research would involve the verification of these models under field conditions, with consideration for variability in material quality, environmental exposure, and the effects of aging. The incorporation of these predictive tools within infrastructure design systems or pavement management systems would also be beneficial for increasing the efficiency and sustainability of road construction operations.Policy-wise, the findings of this study underpin wider efforts promoting the recycling of plastic waste in civil engineering infrastructure. Government departments and industry partners can use such models to develop performance-based plastic-modified binder standards, thus pushing circular economy agendas and combating environmental pollution through the reuse of materials responsibly.

Author Contributions: The authors' contributions are as follows: Maaz Khan, Usama Asif and Touqeer Ali Rind were responsible for the conceptualization, technical implementation, methodology, and drafting of the paper. Muhammad Farooque Panhyar contributed to data collection, compilation, and validation.

Funding: This research received no external funding.

Acknowledgments: The authors would like to express their gratitude to the management of Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, KPK, Pakistan, for their unwavering support and assistance throughout the duration of this study.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding this study and affirm that the work is original, without any form of plagiarism. All sources of information have been properly cited and acknowledged.

References

[1] Debbarma, K., B. Debnath, and P.P. Sarkar, A comprehensive review on the usage of nanomaterials in asphalt mixes. Construction and Building Materials, 2022. 361: p. 129634.
[2] Xu, X., et al., Potential use of recycled concrete aggregate (RCA) for sustainable asphalt pavements of the future: A state-of-the-art review. Journal of Cleaner Production, 2022: p. 130893.
[3] Ali, A., et al. Chemical stabilisation of sandstone road aggregate layers: A literature review. in AIP Conference Proceedings. 2023. AIP Publishing LLC.
[4] Loureiro, C.D., et al., Steel slag and recycled concrete aggregates: replacing quarries to supply sustainable materials for the asphalt paving industry. Sustainability, 2022. 14(9): p. 5022.
[5] Yao, L., et al., Environmental and economic assessment of collective recycling waste plastic and reclaimed asphalt pavement into pavement construction: A case study in Hong Kong. Journal of Cleaner Production, 2022. 336: p. 130405.
[6] Nizamuddin, S., Y.J. Boom, and F. Giustozzi, Sustainable polymers from recycled waste plastics and their virgin counterparts as bitumen modifiers: A comprehensive review. Polymers, 2021. 13(19): p. 3242.
[7] Heydari, S., et al., The use of plastic waste in asphalt: A critical review on asphalt mix design and Marshall properties. Construction and Building Materials, 2021. 309: p. 125185.
[8] Rahman, M.T., A. Mohajerani, and F. Giustozzi, Recycling of waste materials for asphalt concrete and bitumen: A review. Materials, 2020. 13(7): p. 1495.
[9] Abuaddous, M., M.M. Taamneh, and S.R. Rabab'ah, The potential use of recycled polyethylene terephthalate (RPET) plastic waste in asphalt binder. International Journal of Pavement Research and Technology, 2021. 14: p. 579-587.
[10] Haider, S., I. Hafeez, and R. Ullah, Sustainable use of waste plastic modifiers to strengthen the adhesion properties of asphalt mixtures. Construction and Building Materials, 2020. 235: p. 117496.
[11] Radeef, H.R., et al., The mechanical response of dry-process polymer wastes modified asphalt under ageing and moisture damage. Case Studies in Construction Materials, 2022. 16: p. e00913.
[12] Radeef, H.R., et al., Fracture resistance of polymeric wastes modified asphalt using R-curve and digital image correlation. Theoretical and Applied Fracture Mechanics, 2023. 123: p. 103691.
[13] Shah, A.M., et al., Structural Performance of Waste Plastic Bottles Modified Asphalt: A Review. Resources, 2023. 12(1): p. 10.
[14] Veropalumbo, R., et al., Exploring the effect on the environment of encapsulated micro-and nano-plastics into asphalt mastics for road pavement. Environmental Research, 2023. 216: p. 114466.
[15] Zhao, C., et al., Preparation of direct injection waste PE/rubber powder composite modified granules and performance for recycling asphalt. Construction and Building Materials, 2023. 367: p. 130124.
[16] Russo, F., R. Veropalumbo, and C. Oreto, Climate change mitigation investigating asphalt pavement solutions made up of plastomeric compounds. Resources, Conservation and Recycling, 2023. 189: p. 106772.
[17] Rafiei-Sardooi, E., et al., Evaluating urban flood risk using hybrid method of TOPSIS and machine learning. International Journal of Disaster Risk Reduction, 2021. 66: p. 102614.
[18] Mosavi, A., et al., Susceptibility mapping of soil water erosion using machine learning models. Water, 2020. 12(7): p. 1995.
[19] Mosavi, A. and A.R. Varkonyi-Koczy, Integration of machine learning and optimization for robot learning, in Recent Global Research and Education: Technological Challenges. 2017, Springer. p. 349-355.
[20] Mosavi, A., et al., Towards an ensemble machine learning model of random subspace based functional tree classifier for snow avalanche susceptibility mapping. IEEE Access, 2020. 8: p. 145968-145983.
[21] Band, S.S., et al., Feasibility of soft computing techniques for estimating the long-term mean monthly wind speed. Energy Reports, 2022. 8: p. 638-648.
[22] Wang, H., et al., Comprehensive review of load forecasting with emphasis on intelligent computing approaches. Energy Reports, 2022. 8: p. 13189-13198.
[23] Zhang, G., et al., Solar radiation estimation in different climates with meteorological variables using Bayesian model averaging and new soft computing models. Energy Reports, 2021. 7: p. 8973-8996.
[24] Tavoosi, J., et al., A machine learning approach for active/reactive power control of grid-connected doubly-fed induction generators. Ain Shams Engineering Journal, 2022. 13(2): p. 101564.
[25] Azareh, A., et al., Detection and prediction of lake degradation using landscape metrics and remote sensing dataset. Environmental Science and Pollution Research, 2021. 28(21): p. 27283-27298.
[26] Bavili, R.E., et al., A New Active Fault Tolerant Control System: Predictive Online Fault Estimation. IEEE Access, 2021. 9: p. 118461-118471.
[27] Mohammadzadeh S, D., et al., Prediction of compression index of fine-grained soils using a gene expression programming model. Infrastructures, 2019. 4(2): p. 26.
[28] Javed, M.F., et al., New prediction model for the ultimate axial capacity of concrete-filled steel tubes: An evolutionary approach. Crystals, 2020. 10(9): p. 741.
[29] Iqtidar, A., et al., Prediction of compressive strength of rice husk ash concrete through different machine learning processes. Crystals, 2021. 11(4): p. 352.
[30] Zandi, P., et al., Agricultural risk management using fuzzy TOPSIS analytical hierarchy process (AHP) and failure mode and effects analysis (FMEA). Agriculture, 2020. 10(11): p. 504.

[31] Mosavi, A., et al., Fuzzy clustering and distributed model for streamflow estimation in ungauged watersheds. Scientific Reports, 2021. 11(1): p. 1-14.
[32] Mosavi, A. and M. Edalatifar. A hybrid neuro-fuzzy algorithm for prediction of reference evapotranspiration. in International conference on global research and education. 2018. Springer.
[33] Nejad, H.D., et al., Fuzzy State-Dependent Riccati Equation (FSDRE) Control of the Reverse Osmosis Desalination System With Photovoltaic Power Supply. IEEE Access, 2022. 10: p. 95585-95603.
[34] Bemani, A., A. Baghban, and A. Mosavi, Estimating CO2-Brine diffusivity using hybrid models of ANFIS and evolutionary algorithms. Engineering Applications of Computational Fluid Mechanics, 2020. 14(1): p. 818-834.
[35] Rezakazemi, M., A. Mosavi, and S. Shirazian, ANFIS pattern for molecular membranes separation optimization. Journal of Molecular Liquids, 2019. 274: p. 470-476.
[36] Moayedi, H. and A. Mosavi, A water cycle-based error minimization technique in predicting the bearing capacity of shallow foundation. Engineering with Computers, 2022. 38(5): p. 3993-4006.
[37] Ehteram, M., et al., Inclusive multiple model using hybrid artificial neural networks for predicting evaporation. Frontiers in Environmental Science, 2022. 9: p. 652.
[38] Shah, H.A., et al., Prediction of compressive and splitting tensile strength of concrete with fly ash by using gene expression programming. Structural Concrete, 2022. 23(4): p. 2435-2449.
[39] Mosavi, A., et al., Groundwater salinity susceptibility mapping using classifier ensemble and Bayesian machine learning models. Ieee Access, 2020. 8: p. 145564-145576.
[40] Dehghan Manshadi, M., et al., Predicting the Parameters of Vortex Bladeless Wind Turbine Using Deep Learning Method of Long Short-Term Memory. Energies, 2021. 14(16): p. 4867.
[41] Mosavi, A., et al., Comprehensive review of deep reinforcement learning methods and applications in economics. Mathematics, 2020. 8(10): p. 1640.