**Explainable AI (EXAI) for Sustainable Development: Trends and Applications**

Machine Learning Advancements in Polymer Material Creation: Successful Prediction of Glass Transition Temperature

Akshat Rastogi1\*, Divya Thakur2, Shubh Gupta3, Rupal Mishra4, Sandeep Balabantaray5, and Meet Bikhani6

123456School of computer science and Engineering, VIT Bhopal University, Sehore, India

**Abstract**— Polymers have become a crucial component of contemporary society, playing a critical role in various industries such as packaging, electronics, automotive, and healthcare. Owing to their adaptable nature, polymers can be purposefully designed to showcase particular mechanical, thermal, and chemical features, which make them a preferred material for different usages. Due to their adaptability and versatility, polymers are big molecules that can be utilized for a variety of purposes, from packaging to implants. Machine learning has become a potent tool in polymer research, providing a fresh approach to analyzing and forecasting the behavior of polymers under various circumstances. By leveraging the vast amounts of data generated in experiments, patterns and relationships which are not obvious using conventional approaches and can be found using ML algorithms. The molecular weight, solute concentration, temperature, and any additional solutes all have an impact on how viscous a polymer solution is. As the temperature rises, the molecules become more mobile, increasing the viscosity of the solution and reducing flow resistance. Temperature has an impact on viscosity, as shown by the Arrhenius equation, which relates viscosity to temperature and the activation energy of molecules. To determine if the data’s accuracy has improved or not, we have used simple machine-learning models. The correctness of the data has been compared here using models like Multinomial NB, Gaussian NB, MLPClassifier and DecisionTreeClassifier, with results of up to 95%.

**Keywords**: Monomer, Temperature, Viscosity, Arrhenius equation, Multinomial NB, Gaussian NB, BernoulliNB, MLPClassifier, DecisionTreeClassifier.

1. **INTRODUCTION**

One of the physical properties of numerous types of polymers that are covered by experimental data in this collection is the glass transition temperature (Tg), a significant element that regulates the thermal stability of a polymer material. The objective of this project is to build machine learning models that can realistically calculate the melting point (Tg) of a polymer material based on its own molecular makeup and processing conditions. For the creation of new polymer materials with desired thermal properties, accurate (Tg) prediction is essential. In the past, determining (Tg) required laborious and drawn-out experimental techniques that could be expensive and weren’t always practical for high-throughput screening of several polymer compounds. Consequently, developing precise and efficient prediction models for (Tg).

We will employ supervised learning [1] to train classification models on the Polymer Dataset in order to accomplish this goal. To find the features that are most important for predicting (Tg), we will investigate several feature engineering and selection strategies. The effectiveness of the algorithms are analyzed using a series of metrics, including the accuracy, recall, precision, and f1-score. The findings of this study will give insight on how machine learning algorithms are used to determine polymer (Tg), this will contribute in the construction of more consistent and cost-effective (Tg) measurement methodologies. The proposed models can also be used to screen new polymer materials for their thermal characteristics, which makes it easier to find new materials with the appropriate thermal characteristics for a variety of applications.

The transition temperature of glass (Tg) is an important property of polymers that affects their thermal stability and processing behavior. (Tg) determines the temperature range over which a polymer will use supervised learning to develop regression models on the Polymer Dataset to do this. We will look into several feature engineering and selection strategies for identifying the traits that are most important for prediction of transition temperature (Tg). The performance of the models will be assessed using a number of metrics, including accuracy, recall, precision, and f1-score. The findings of this work will help the growth of more accurate and cost-effective (Tg) measuring techniques as well as knowledge about how machine learning algorithms may be utilized to compute polymer Tg. It is simpler to locate novel materials with the right thermal properties for a range of applications by using the proposed models to screen new polymer materials for their thermal properties. The performance as well as longevity of polymer materials can be affected by changes from a stiff, glassy state to either a softer, rubbery one. Therefore, for designing and synthesizing polymers having desired thermal characteristics, (Tg) measurement must be accurate and effective.

In the past, measuring the glass transition temperature (Tg) required time-consuming and expensive experimental methods like dynamic mechanical analysis (DMA) or differential scanning calorimetry (DSC). However, recent computer modeling and machine learning developments have made it possible to forecast Tg based on processing parameters and molecule composition, eliminating the need for extensive experimental testing.

In recent years, machine learning models have garnered substantial attention for their ability to efficiently and cost-effectively forecast glass transition temperature (Tg) compared to traditional experimental methods. To construct a model for Tg prediction based on chemical composition and processing characteristics, regression models are trained using supervised learning, a prominent machine learning technique, with the Polymer Dataset. The Polymer Dataset comprises an extensive repository of data on the physical and chemical attributes of various polymers, with Tg as the target variable for the regression models.

A pivotal aspect of developing accurate machine learning models for Tg prediction is identifying the most pertinent features or input variables contributing to Tg prediction. This study employs various feature engineering and selection strategies to discern the most influential elements. Feature engineering involves the creation of new features by combining existing ones, while feature selection aims to pinpoint the most informative subset of features. The effectiveness of the models is assessed through diverse metrics, including accuracy, recall, precision, and the f1-score.

Implementing machine learning in polymer research encounters challenges primarily due to the demand for extensive and diverse datasets. While substantial, the Polymer Dataset used in this study applies to only a limited subset of polymers and may not fully represent the broad spectrum of polymer materials. To overcome this limitation, efforts are underway to construct more prominent and diversified datasets, which can be utilized to develop machine learning algorithms with enhanced accuracy and robustness.

Another obstacle lies in the interpretability of machine learning algorithms. Comprehending their relationship to the fundamental physical and chemical processes governing polymer behavior can be challenging despite their high predictive accuracy. To address this issue, endeavors are being made to devise more interpretable machine-learning algorithms capable of shedding light on the molecular-level mechanisms underlying polymer behavior.

The subject of polymer science could undergo a revolution if precise and effective machine learning algorithms for forecasting (Tg) are developed. It is possible to correctly forecast (Tg) regarding the molecular composition as well as processing parameters by merging machine learning approaches with experimental data, which eliminates the need for costly and time-consuming experimental methods. The creation of new polymers having specified thermal properties and the thermal property screening of advanced materials can both benefit from the development of reliable (Tg) prediction models.

1. **PROBLEM DESCRIPTION**

The glass transition temperature of a polymer material is one crucial element in determining its thermal stability (Tg). In this collection of various polymers and associated experimental data, it is a significant physical property. The goal of this study is to develop machine learning algorithm that can accurately predict the transition temperature (Tg) for polymer materials using information about those materials' molecular structure and processing properties as inputs. Creating new polymer materials with particular thermal characteristics depends on accurate Tg prediction. Traditional techniques for measuring Tg were expensive and time-consuming, making them unsuitable for the high-throughput screening of many polymer compounds. Establishing precise and effective prediction models for Tg is therefore highly motivated.

Polymer materials are widely used in various applications, such as packaging, construction, textiles, and electronics. However, their thermal stability is an important factor that affects their performance and lifespan. The glass transition temperature of a polymer material is an important consideration when considering its thermal resistance (Tg). The transitional temperature (Tg) between a polymer's glassy and rubbery states. Several variables, such as manufacturing conditions, chain flexibility, and molecular weight, can affect Tg.

In recent years, polymers' molecular makeup and processing factors have been used to predict the glass transition temperature (Tg) with positive results from machine learning (ML) techniques. ML models help create accurate and effective prediction models for Tg because they are good at analyzing large datasets and identifying complex relationships between input factors and the target variable.

The investigation's findings will be beneficial in understanding how machine learning models may be used to calculate polymer Tg. They will help develop more dependable and affordable Tg testing methods. Furthermore, the suggested models make it easier to screen new polymer materials for thermal characteristics, which would streamline the finding of materials with suitable thermal properties for various applications. This work has the potential to significantly influence the polymer sector by presenting a new instrument for the development of high-performance polymer materials.

1. **LITERATURE REVIEW**

In numerous studies, supervised learning methods have been used to develop machine learning models that estimate the transition temperature (Tg) on the basis of information about polymers.

In [2], the conventional methods for discovering and designing new materials encounter limitations in terms of time and resources, necessitating the development of a new approach to expedite the process. In materials research, machine learning has become an up-and-coming technique that may significantly improve the accuracy and efficiency of material property prediction and the discovery of new materials. The author focuses on the critical algorithms involved, the proven methods for applying machine learning to materials science, and the state of the field's current research. While there have been notable strides in this domain, challenges in applying machine learning to materials science persist and are explored within this study, along with potential solutions and future research directions.

In [3], the author emphasizes how the Materials Genome Initiative may advance methods to informatics that are data-centric, particularly in the development of novel polymer materials. In this article, the author examines the topic of polymer informatics and introduces a brand-new technique for predicting the characteristics of recently produced polymers using machine learning models. The author also details the creation of Polymer Genome, an easily navigable website that contains these prediction algorithms. In addition, a plan has been put out to gradually expand the chemical and property spaces to increase Polymer Genome's adaptability and applicability across a broader range of technological domains.

The author used ellipsometry in [4] to look at how different variables may affect the transition temperature (Tg) on dry polystyrene (PS) brushes. These metrics included thicknesses, molecular masses, and peeling densities. It was discovered that Tg is significantly correlated with the PS brushes' grafting density. Tg rises for high-density brushes as Mn (or brush thickness) declines, but Tg falls for low-density brushes as Mn (or brush thickness) decreases. Surprisingly, the investigation also revealed that Tg is insensitive to Mn or brush thickness in intermediate-density brushes. This is likely because the PS brush shape changes from a mushroom-like structure to a brush configuration.

In [5] the combination of graphene and carbon nanotubes (CNTs) in bitumen composites formed 1D-2D hybrid structures with dramatically better mechanical properties, according to the author’s research. The inclusion of 0.2 weight per- cent graphene and 0.8 weight percent of carbon nanotubes showed the most substantial improvements in stiffness, rutting resistance, and low-temperature cracking resistance. The study emphasizes the need of minimizing graphene sheet stacking and CNT aggregation in order to make carbon Nano-filler-reinforced composites more widely applicable.

In [6], the author acknowledges the intricacies of developing polymers due to their complex chemical composition and structural attributes. However, the study underscores the significant influence of data-driven methodologies in advancing polymer science and engineering. It places a special emphasis on the use of enormous data resources and the implementation of machine learning methods. The author highlights the emergence of polymer informatics as a valuable tool for predicting polymer performance and optimizing polymer development processes. This progress is made possible by continuous advancements in machine learning algorithms and the augmentation of existing databases. The study provides an insightful overview of the developing field of machine learning-assisted polymer computing, which can be very useful to researchers working in the fields of materials science, machine intelligence, and related fields.

For the purpose of forecasting the transition temperature (Tg) of polymers, [7] the author created a technique called Gaussian process regression (GPR) model. This model made use of molecular characteristics, specifically the molecular average hexadecapolar moment and the molecular traceless quadrupole moment. The model is extremely accurate and reliable, demonstrating GPR's capacity to identify and model the connection between quantum chemical parameters as well as (Tg). This model is also simpler and requires less inputs than other approaches, making it useful for a variety of polymers having (Tg) levels above or below room temperature.

In [8], the author did extensive research on oxide glasses, gathering a large quantity of data on glass com- positions and glass transition temperatures. They devised an artificial neural network capable of confidently forecasting oxide glasses' glass transition temperature (Tg). Their method of research is flexible and can be adapted to forecast numerous different glass qualities. This predictive technique will be made publicly available as a web tool, with the potential to inspire new ideas and ANN applications in selection and creation of unique glasses with standout features for cutting-edge applications.

In [9], the author emphasizes how polyimides' mechanical and thermal characteristics—a critical class of engineering plastics—are shaped by the glass transition temperature (Tg). To create novel polyimide polymers with desired properties and uses, predicting Tg for polyimides beforehand is essential. The author investigates three approaches to estimate Tg for polyimides, including machine learning methods and all-atom molecular dynamics simulations. The findings suggest that Tg may be accurately calculated by combining a machine-learned mathematical model with the diffusion coefficient for simple gas molecules within a polyimide. This study is critical to accelerate the creation and identification of new polyimide polymers with particular characteristics and uses.

In reference [10], The author has successfully shown that an architecture with a deep convolutional neural network may retrieve pertinent data about the polymer glass - transition temperature. Given that it only needs to know the chemical composition of the repeating units and thus does not require experimental measurements or calculations as input, this approach is an essential design tool of material scientists and engineers. The author's work offers a potential direction for further research and has significant ramifications for the field of material science.

In [11], According to the author, technological changes and the need for novel polymers necessitate continual study in the design and identification of polymers with specified physical and chemical properties. They emphasize the importance of the glass transition temperature in identifying specific applications and appropriate manufacturing conditions. Polymer complexity makes it challenging to design innovative polymers for specific purposes, and the author emphasizes the benefits of computational tools for learning about the features of developed polymers. In this study, the author investigates the performance of the Graph Attention Network (GAN) and the Convolutional Neural Network (CNN) in predicting the transition temperature of polymers.

In reference [12], The transition temperature (Tg) in thermoset systems may be predicted by the author using only the chemical compositions of resins and hardeners, showing the capability of machine learning (ML). The author gets a good accuracy rate (MAE = 16°C and R2 = 0.86) and effectively predicts Tg for 210 new resin/hardener combinations by training an Machine Learning ensemble model with 94 resin/hardener groupings. These findings accelerate the development of novel thermosets with desired properties and contribute to sustainability by eliminating trial-and-error methods. Currently, the author is enhancing the model's accuracy by incorporating quantum mechanical features of molecules in the dataset. This work underscores the capacity of machine learning in enhancing efficiency and uncovering new relationships in thermoset development.

In [13], virgin and post-consumer recycled (PCR) polymers are in high demand due to their unique characteristics, and the author explores the potential advantages of employing a data-driven approach in polymer design. However, the intricate hierarchical structures of polymers present significant challenges in polymer informatics. To address this challenge, machine learning methods are proposed as cost-effective and scalable solutions for predicting a range of polymer material properties. The author gives a thorough rundown of the essential procedures for integrating machine learning into polymer research, including topics such as polymer design, fingerprinting, representations, open-source databases, and algorithms. In order to enable effective and focused polymer discovery and development, the author emphasizes the significance of continuing research in machine learning applications for PCR polymers in their conclusion.

According to [14], material scientists must detect Glass Transition Temperatures with precision and efficiency. The author recognizes that determining the Glass Transition Temperature is difficult since it depends on several polymer's chemical and physical characteristics. Particularly when dealing with polymers that have a wide variety of glass transition temperatures, empirical approaches might be challenging. This Simplified Molecular-Input Line-Entry System (SMILES) arrangement of polymer molecules and a Long Short-Term Memory model are used by the author in a novel approach to estimate glass transition temperature.

In [15], According to the author, ML and data-enabled methodologies have brought about substantial changes in materials research. As a result of the formation of this new paradigm, the conventional approach towards materials design but also discovery is going to undergo a significant change. Materials informatics, according to the author, is a fully mature science, and ML algorithms are assisting in efficient materials property forecasts, materials design, and discovery.

In order to investigate and forecast the phases and transitions of polymers, the author of [16] underlines the application of AI and machine learning. The article centers on deep learning methodologies and presents dPOLY, an adaptable artificial intelligence instrument to examine molecular dynamics paths. The ability of such dPOLY framework to forecast critical temperatures of phase transitions across a variety of polymer sizes, along with the spiral into globule transition, exemplifies its adaptability. Additionally, this framework may be expanded to include several other phase transitions including dynamic changes in procedures like polymerization and the synthesis of other soft materials. The author highlights how AI and deep learning techniques have the potential to enhance the forecasting and characterization of polymer phases significantly.

In [17], the author acknowledges the limits of traditional analytical and empirical models for forecasting the mechanical properties of concrete mixtures and acknowledges that machine learning (ML) models may be a more efficient alternative. The author appears to promote the use of ML models as a promising strategy to overcoming the limitations of older approaches.

In [18], the author highlights the increasing importance of employing big data, collaborative computing, and machine learning in materials research, which supplements traditional theoretical and experimental methods. The talk focuses on how these strategies might speed up polymer design and find novel structure-property connections, as well as the underlying concepts of data-driven techniques and their limitations in polymer research. While machine learning models may not match the precision of first-principles techniques, they can provide reliable approximations of material properties within the range of their initial training data, making them practical for interpolation tasks. The author also delves into amalgamating first-principles methods with advanced optimization techniques for polymer inverse design, resulting in substantial data generation for discovering new correlations. The author emphasizes the need for additional efforts to address difficulties related to setting up high-throughput instrumentation as well as seamlessly incorporating them with optimization as well as machine learning algorithms, even though experiments with elevated setups but also robot-assisted synthesis & design have shown preliminary success.

In reference [19], the author delved into the study of sucrose inversion kinetics in the presence of various acids, revealing that the reaction rate escalated with rising temperatures. The author postulated that the activation energy of the reaction was linked to the energy needed for reactant molecules to surmount the energy barrier and attain the transition state. Subsequently, the author formulated the Arrhenius equation, which elucidates the connection between a reaction's activation energy (Ea), rate constant (k), the gas constant (R) and temperature (T):

(1)

Here, α symbolizes the pre-exponential factor, signifying the collision frequency between reactant molecules, while "exp" designates the exponential function. This equation has become a standard tool for investigating the temperature-dependent behavior of reactions and processes in chemistry and various other domains, including materials science, biology, and engineering. The research outlined in this paper is considered a seminal contribution to the realm of chemical kinetics and has played a pivotal role in shaping the progress of modern chemistry.

1. **METHODOLOGY**
2. **Data acquisition and pre-processing**

This dataset displays measured data of the relationship between various input parameters and a polymer solution's viscosity. The log of viscosity (in cP), polymer concentration (wt.%), Ca+2 concentration (wt.%), NaCl concentration (wt.%), and log of shear rate (in s-1) are some of the input parameters (in Celsius). The temperature is the output variable.

There are 420 rows in the dataset, and each row represents a different arrangement of the input parameters. The observations were probably made using a rheometer, a tool for measuring the characteristics of fluid flow.

Table 1 - DESCRIPTION OF ATTRIBUTES

|  |  |
| --- | --- |
| **Title** | **Description** |
| log of shear rate (in sˆ-1) | The shear rate is a measurement of how quickly a fluid shears or deforms when a force or tension is applied. |
| Polymer concentration (wt %) | It represents the amount of polymer present in the solution. |
| NaCl concentration (wt %) | It refers to the concentration of sodium chloride (NaCl) in the solution. |
| Ca+2 concentration (wt %) | It speaks of the quantity of calcium ions (Ca+2) present in the solution. |
| Temperature (in Celsius) | It refers to the temperature of the solution during the rheological measurements, expressed in (°C). |
| log(viscosity) in cP | The viscosity of polymer solution, measured in centipoise, is referred to as "log(viscosity) in cP." (cP). |

The dataset could be helpful for understanding how the polymer solution behaves rheologically under various situations and for adjusting the formulation for particular applications. You can investigate the connections between the input factors and the outcome variable and learn more about the behavior of the solution by using a variety of quantitative and data visualization tools.

Viscosity, which measures a fluid's resistance to flowing, is a crucial characteristic of polymer solutions in many industrial applications. The polymer concentration, NaCl concentration, and Ca+2 concentration are all factors that can affect the viscosity of the solution. The rate at which the fluid is sheared is measured by the log of shear rate, which may also have an impact on the solution's viscosity. Temperature is an important output variable because it can affect the behavior of the polymer solution in various ways, such as by influencing the molecular structure of the polymer chains.

The rheometer is a tool that can be used to measure the rheological properties of fluids, such as viscosity, elasticity, and yield stress. The observations in this dataset were likely made using a rheometer, which is a common tool in the field of polymer science and engineering.

You can learn more about the connections between the input and output variables as well as how the mixture behaves rheologically under various circumstances by studying this dataset using data and mathematical visualization tools. For particular applications, such as the creation of coverings, adhesives, and plastics, this knowledge can be used to formulate the solution in the most effective way possible.

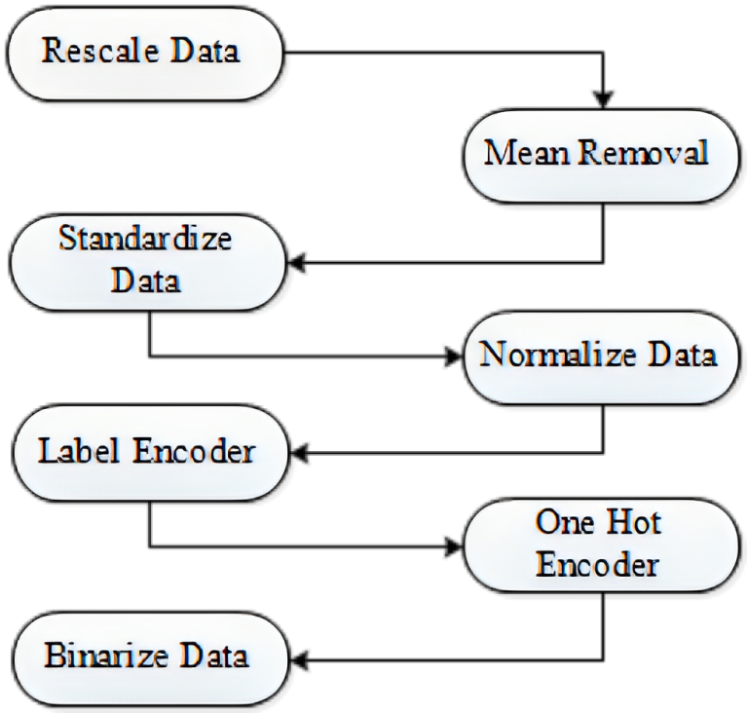


Figure 1 - Data Preprocessing Steps

1. ***Rescaling Data***: A data preprocessing technique called rescaling involves converting a dataset’s feature values to a common scale, often between 0 and 1 or -1 and 1. This method prevents the issue of some features predominating others, which can result in biased outcomes in machine learning models. In rescaling, the feature values are converted to have the same range and to have modified values that are proportional to the original values. Min-Max scaling and Standardization are the rescaling methods that are most frequently utilized. When the data range is known and fixed, min-max scaling is employed; however, standardization is employed when the data range is unknown or not fixed.
2. ***Mean Removal***: Mean removal is a data preprocessing technique used to center the data by removing the mean value of each feature. This technique is used to eliminate the effect of different mean values across features and allows for the comparison of features with different scales.

In mean removal, the mean value of each feature is subtracted from each data point. This results in a new dataset with a zero mean for each feature, which is useful for many machine learning algorithms. Mean removal is particularly important when dealing with datasets with a large number of features or when working with high-dimensional data. This technique can help reduce the computational complexity of machine learning models and improve their performance by a large measure. To get the best results, mean removal should be combined with other preprocessing methods like rescaling or normalization, it is vital to highlight.

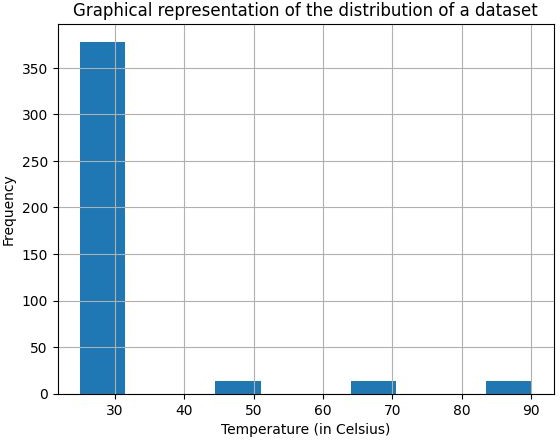
1. ***Standardizing Data***: The obtained features of a dataset are modified using a data preparation method called standardization to have an average of 0 and a sample variance of 1. Using this technique, the data is rescaled to a standard range, which simplifies data analysis and improves the performance of machine learning algorithms. The data is first normalized by subtracting each feature value from its mean, and then divided by its confidence interval. In order to facilitate comparison of traits with different scales, a new sample is built with each characteristic having an average of 0 and sample variance of 1. Classifying data is especially helpful when working with data with various scales and units since it removes the scale influence.
2. ***One hot encoding***: A data preparation method called one hot encoding uses a collection of binary variables—one for each potential category—to represent categorical variables. According to this method, each distinct category is given a separate binary variable, whose value is set to 1 if the category is present and 0 otherwise. The dataset that results has a set of binary variables for each category, making categorical data comparison and analysis simple. When working with categorical variables like color or gender that lack an inherent ordering, one hot encoding is especially helpful. It enables categorical data to be used as input for machine learning algorithms, enhancing their efficiency and precision.
3. ***Normalizing Data***: Scaling numerical data to a common range is done using the data preprocessing technique known as normalization. The efficiency of machine learning algorithms that use distance-based measures or gradient-based optimization [20] methods can be enhanced by using this technique, which is used to make sure that the numerical values of various features in a dataset have a similar scale. The values of each feature in the dataset are scaled to a range between 0 and 1, or between -1 and 1, as part of the normalization process. This is done by dividing by the feature’s range or standard deviation and then subtracting the feature’s mean. This procedure makes sure that each feature has a comparable scale, which can facilitate data comparison and analysis.
4. ***Label Encoding***: Label encoding is a simple and effective way to prepare categorical features for use in machine learning models. It transforms ordinal and nominal converted into information values that are simple for algorithms to process. However, one potential problem with label encoding is that it introduces an arbitrary order to the categories that may not reflect their true relationship. For example, if “red” is assigned a label of 0 and “green” is assigned a label of 2, this suggests that green is somehow “bigger” or “  
   more important” than red, which may not be true. Therefore, it is important to use label encoding with caution and consider whether the order of the labels is meaningful in the context of the problem at hand. Another limitation of label encoding is that it can lead to overfitting, especially in models with a large number of categories. This is because label encoding creates a new feature for each category, which can increase the dimensionality of the data and make it more difficult for the model to generalize to new examples.
5. ***Binarizing Data***: Data preprocessing method called “binarization” is used to translate numerical data into binary values. Setting a threshold value and changing any numbers over it to 1 and any numbers below it to 0 are the steps in this strategy. When working with numerical data that has distinct thresholds, such as the results of medical tests or the pixel intensities of images, binarizing the data is a valuable technique. Setting the threshold too high or too low could lead to data loss or inaccurate results, therefore it’s crucial to select a threshold number that makes sense for the particular situation and dataset.
6. ** Distribution of the dataset represented visually**

Figure 2 - The distribution of a dataset represented graphically

The range of temperature readings in the dataset is shown visually in Fig. 2's histogram plot. By plotting the occurrence of temperature values against their corresponding temperature bins, the plot enables easy identification of the most common temperature values in the dataset, as well as the range of temperatures that occur most frequently. This information can be helpful in understanding the central tendency and spread of the temperature data, which is important in many fields such as environmental science, meteorology, and health studies.

The histogram plot not only accurately depicts the temperature data but also provides important information about the temperature distribution's structure. The histogram, for instance, may indicate that the temperature information is routinely distributed if it is about symmetrical. If the histogram has a long tail on one side, it may indicate that the temperature data is skewed. These characteristics of the temperature distribution can inform further analysis and modeling of the data.

The histogram plot is thus a helpful tool for examining and displaying temperature data. It helps in the finding of patterns and themes that might not be immediately apparent from the raw data by providing a succinct and understandable overview of the data. By presenting the data in a graphical format, the histogram plot allows for quick and intuitive comparisons between different datasets or temperature ranges, enabling more effective analysis and decision-making.

1. **BernoulliNB**

The features of the BernoulliNB algorithm are binary, which means they can either be present or not in the document. The existence of the word "spam" in the email, for instance, could be a characteristic for classifying emails as spam or not.

Following are the steps the algorithm takes to determine how likely it is that a document refers to the particular class (like spam).:

(2)

According to the BernaulliNB algorithm, features can exist without another being present. Therefore, the possibility of the document provided the class can be calculated as the multiplication of the possibilities of each characteristic given the class.:

(3)

Where, P (m | n) is the probability of noticing a feature given a class. Using maximum likelihood estimation, the probabilities of the characteristics given the class may be calculated from the training data. To prevent zero probability, one can alternatively utilize the Laplace smoothing method.

Therefore, when dealing with binary information, the BernoulliNB algorithm is a straightforward yet efficient technique for text classification applications.

1. **GaussianNB**

A variation of the Naive Bayes algorithm used for classification tasks is the Gaussian Naive Bayes (GaussianNB) method. It can handle features with continuous values and makes the assumption that the features are having a Gaussian (normal) distribution.

The algorithm determines the probability of the data belonging to a given class as follows:

(4)

Where:

* P (x | y) shows how likely it is that, given the document's characteristics, it belongs to the class.

|

* P(x) is the prior probability for the class.
* The document's probability, P (y | x), is determined by the class.
* P(y) is the document’s evidentiary probability

So, GaussianNB algorithm assumes that the likelihood of the document given the class is a multivariate normal distribution. Hence, given the class, the likelihood of witnessing the document features can be estimated as follows:

(5)

Where:

* π is the mathematical constant pi
* µy is the factor that determines the features' covariance matrix.
* x is the vector of feature values in the document
* mean is the vector of mean feature values for the class
* σy is the covariance matrix of the features for the class

You may get the class's likelihood function from the training examples by calculating the percentage of data training set that basically consists. The mean as well as covariance of a features for every class can also be estimated via probabilistic estimation using the training data.

1. **MultinomialNB**

The data in the MultinomialNB method are represented by counts, where each feature is a word in the vocabulary and the count denotes how many times that word appears in the document. For example, if we are classifying movie reviews as positive or negative, a feature might be the frequency of the word “great” in the review.

The algorithm computes the possibility of a document belonging to a specific class (e.g., positive) as follows:

(6)

Where:

* P (x | y) shows how likely it is that, given the document's characteristics, it belongs to the class.

|

* P(x) is the prior probability for the class.
* The document's probability, P (y | x), is determined by the class.
* P(y) is the document’s evidentiary probability

The following formula can be used to determine the probability of the document given the class:

(7)

Where,

* Nyi based on the class, what is the likelihood of detecting feature i.
* θyi is how many of feature I there are in the document.

The probabilities of the features given the class can be estimated from the training data using maximum likelihood estimation or smoothing techniques such as Laplace smoothing.

1. **MLPClassifier**

The MLPClassifier (Multi-Layer Perceptron Classifier) algorithm for classification tasks, is a well-known machine learning algorithm. It is a form of artificial neural network composed of numerous layers of nodes, each of which is a simple computational unit that receives input signals and generates an output signal.

By analyzing a collection of training data and labels, the algorithm learns to map input features to output labels. The approach adjusts the biases & weights of the network's node during training to reduce the discrepancy between the expected outcome and the actual label.

The MLPClassifier method produces a probability distribution over all possible classifications. As shown below, a data point's likelihood of belonging to a particular class is calculated.:

(8)

Where:

* ya is the ath data point's actual label.
* The data point of ath element is Xa.
* j is jth class
* K is total number of classes
* zaj is input to the activation

A weight value of the output from the layer prior to this is passed through a non-linear and activation function to form an activation function for the output layer. The network can learn complex decision limitations because of the activation function's addition of nonlinearity.

To update the weights and biases of the network's nodes, backpropagation is used during training. This involves analyzing the gradient of the gradient descent with appropriate weights and biases and using learning algorithm to update the parameters. For issues related to classification, the loss of cross-entropy function is commonly utilized:

(9)

Where:

* yaj is the true label of each ath data point for jth class
* ŷaj is given for predicted probability of each ath data point for the jth class
* The overall sample size is n.

The MLPClassifier algorithm is a powerful as well as flexible algorithm that can learn complex decision boundaries and handle high-dimensional input data.

1. **DecisionTreeClassifier**

The decision tree is constructed by recursively separating the data depending on the best attribute that gives the greatest amount of information gain. The information gain quantifies the decrease in entropy or impurity of the data following the split. Every node portrays a potential value or option depending on each node's representation of a feature characteristic.

Information gain is:

(10)

Where:

* IG (Y, A) is attribute A’s calculated information gain on data set Y.
* H(Y) is the data set Y’s entropy.
* H (Y|A) is the data set Y's conditional entropy given attribute A.

|

The entropy to calculate the Information gain of a set is given by:

(11)

Where:

* S is a collection of samples.
* π is proportion of samples belonging to class i.

As soon as the decision tree is constructed, it can be used to categories fresh data by moving through the tree according to the values of its characteristics until the end is met that correlates to the class label.

1. **RESULTS**

This section describes the simulation setup, performance metrics used for result analysis, and output results.

1. **Simulation Setup**

Python was used to run the code on a Jupyter notebook. The experiment was carried out on a system equipped with an Intel 2.50-GHz, i5 CPU and 8.0-GB RAM. The simulation is run 30 times, and the average results are displayed graphically and tabulated.

1. **Performance Indicators**

This section describes the metrics chosen to evaluate the performance of the suggested methodology.

1. ***F1-score performance***: The Predicted accuracy of a Machine Learning model on almost any set of data is assessed using the F-score, also called the F1-score. It is employed to assess the classification issue. Additionally, it is used to assess information retrieval systems, including machine learning models. In essence, it is a method for combining machine learning models' recall and accuracy. It is described mathematically as that of the harmonic average of the recall as well as accuracy of the machine learning model. The F1-Score changes that puts precision before recall or the other way around. To get an F1 score, you must first understand recall and precision.

(12)

1. ***Precision performance***: Precision is calculated as the proportion of appropriately significant positive observations to all anticipated positive discoveries. Precision refers to a classifier's capacity to classify a data as positive even when it is probably negative.

(13)

Precision, on the contrary, calculates the percentage of true positive prediction that correspond to the overall positive class.

1. ***Recall performance***: It shows positive predictions obtained from positive examples using a connected data collection. Recall is the number of True Positives that our model identifies as Positive (True Positive). Using the same logic, we can argue that when the cost of False Negative is higher, we will use Recall as a model metric to predict our best machine learning model.

(14)

1. ***Accuracy performance***: In general, the proportion of correct observations compared to all observations is the most relevant performance metric. Accuracy is an invaluable statistic where the data sets are homogeneous and the number of false positives and negatives are approximately equal.

(15)

1. **DESCRIPTION OF OBTAINED RESULTS**

Tables 2 and 3 showcase the results of the proposed models applied, as well as various essential classifiers. It has been discovered that the proposed strategy, which utilizes a Decision Tree Classifier model, achieves up to 98% accuracy. The F1 score, recall, precision, and accuracy values are shown in Table 3.

Table 2 - DIFFERENT MACHINE LEARNING MODEL’S RESULT

|  |  |
| --- | --- |
| **Machine Learning Models** | **ACCURACY ACHIEVED** |
| BernoulliNB | 93% |
| GaussianNB | 94% |
| MultinomialNB | 75% |
| DecisionTreeClassifier | 98% |
| MLPClassifier | 86% |

Table 3 - CLASSIFICATION TABLE OF DECISION TREE CLASSIFIER

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribute** | **Precision** | **Recall** | **F1 score** | **Support** |
| 25 | 1.00 | 0.97 | 0.99 | 78 |
| 50 | 0.50 | 1.00 | 0.67 | 2 |
| 70 | 1.00 | 1.00 | 1.00 | 2 |
| 90 | 1.00 | 1.00 | 1.00 | 2 |
| Accuracy |  |  | 0.98 | 84 |
| Macro avg | 0.88 | 0.99 | 0.91 | 84 |
| Weighted avg | 0.99 | 0.98 | 0.98 | 84 |

The information in Fig. 3 shows the effectiveness measures of various categorization models applied to a dataset. The Recall, Precision, and F1-score values for each model are displayed on the line graph. In this graph, the BernoulliNB, GaussianNB, MLPClassifier, MultinomialNB, and Decision Tree models are contrasted. Metric values are shown on the y-axis, and model names are shown on the x-axis.

Figure 3 - Comparison Graph of accuracy in Percent

Figure 4 Precision, Recall, and F1-Score Metrics for Performance Comparison of Classification Models

As shown in Fig. 4, the Decision Tree model outperforms all others having precision scores of 0.99, recall scores of 0.98, as well as an F1-score of 0.98. The GaussianNB model likewise performs well, with precision, recall, and F1 scores of 0.98, 0.94, and 0.95, respectively. In comparison to the MultinomialNB model, which used this data, the BernoulliNB model has a precision score of 0.86, a recall score of 0.93, and an F1-score of 0.89. The MLPClassifier technique fared the worst of all models, with precision, recall, and F1 scores of 0.73, 0.86, and 0.79, respectively.

Figure 5 - Graph of DecisionTree's Performance

The performance graph shows the values for the DecisionTree classifier are shown in a bar graph in Fig. 5. The model's predictions of true positives are accurate 99 percent of the time, as shown by the calculated precision value of 0.99. The model correctly detected 98 percent of the dataset's positive cases, according to its recall value of 0.98. The model appears to be effective at identifying positive occurrences, according to the F1-score of 0.98 which is the harmonic mean of recall and precision.

Figure 6 – Bernoulli’s performance bar graph

The bar graph in Fig. 6 displays the performance metrics of the BernoulliNB classifier, which were evaluated using performance factors. The 0.86-inch-high bar, which represents the classifier's precision, represents the classifier's performance. This demonstrates that 86% of all the occurrences the model identified as positive were in fact true positives. With a recall of 0.93, indicated by the length of the bar, the model was capable of recognizing 93 percent of the real positive events. The F1 score is 0.89, and it is represented by the length of the green bar. This statistic provides a broad assessment of the model's effectiveness as a weighted harmonic average of the recall and precision. An F1-score of 0.89 suggests that the classifier has a decent balance of recall and precision in this case.

Figure 7 - GaussianNB's Performance Graph

The Gaussian Naive Bayes algorithm's performance factor metrics are shown in the bar graph in Fig. 7. This algorithm has a precision score of 0.98, meaning that it is highly accurate at picking out true positives from all other positive predictions. With a recall score of 0.94, GaussianNB can properly identify a significant portion of true positives among all actual positives. The precision and recall-weighted harmonic mean, or F1-score, is 0.95., indicating excellent ability in forecasting positive events I.e., the Gaussian Naive Bayes method appears to perform quite well in classification tasks, scoring highly in all three criteria.

Figure 8 – Multinomial’s Performance Graph

Bar graph of Fig. 8 represents the performance factors for the MultinomialNB classifier. The algorithm's precision score is 0.94, which indicates that 94 percent of the time when it produces a forecast, it is accurate. Recall is 0.75, which means that 75% of positive samples are correctly identified by the model. The precision and recall components of the F1-score are both 0.83. With this result, the MultinomialNB classifier appears to strike a reasonable compromise between precision and recall for the dataset being used. The graph indicates that the model is significantly biased towards generating false negatives because the accuracy score is higher than the recall score.

Figure 9 – MLPCLassifier’s performance graph

The Performance Factorts for the MLPClassifier model are displayed as bars in Fig. 9. Recall is 0.86, F1-score is 0.79, and accuracy is 0.73. According to the precision value, 73% of the occurrences that the model identified that the positive are indeed positive. While the recall number demonstrates that the model accurately detected 86% of the positive events. The harmonic means of classifiers, known as the F1-score, provides an overall evaluation of the model's performance. These results are shown graphically by a bar graph with the model's performance parameter as the x-axis. The graph shows that MLPClassifier has relatively lower precision and F1-score compared to the other models in the dataset, indicating that it may not perform as well as other models in predicting the target variable.

1. **CONCLUSION**

According to the evaluation metrics for the Decision Tree Classifier model, the model is effective at forecasting the glass transition (Tg) of polymeric composites. The model achieved high precision, F1-score and recall of 0.99, 0.98, and 0.98. The model correctly predicted the samples with a 98 percent accuracy. These results demonstrate that (Tg) may be precisely and efficiently predicted using machine learning methods, which can aid in the development of novel polymer materials with desired thermal properties. To enhance the model’s functionality and broaden its applicability to various polymer datasets, future research can investigate different feature engineering and selection techniques. As a result, this study contributes to the development of polymer materials and the realm of materials science.

1. **FUTURE WORK**

The accurate prediction of transition temperature is one possible area for machine learning developments in polymer material fabrication (Tg).

A crucial factor that impacts the mechanical, thermal, and physical properties of polymers is the temperature where glass transitions. For the planning and improvement of polymeric composites for numerous applications, including in the automotive, aerospace, and electronics sectors, it is crucial to accurately forecast (Tg).

1. **REFERENCES**
2. D. Thakur and P. Lalwani, “Human joints analysis system: A machine  
   learning approach,” in 2022 IEEE World Conference on Applied Intelli-  
   gence and Computing (AIC). IEEE, 2022, pp. 410–415
3. Y. Liu, T. Zhao, W. Ju, and S. Shi, “Materials discovery and design using machine learning,” Journal of Materiomics, vol. 3, no. 3, pp. 159–177, 2017.
4. C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, and R. Ramprasad, “Polymer genome: a data-powered polymer informatics platform for property predictions,” The Journal of Physical Chemistry C, vol. 122, no. 31, pp. 17 575–17 585, 2018.
5. B. Zuo, S. Zhang, C. Niu, H. Zhou, S. Sun, and X. Wang, “Grafting density dominant glass transition of dry polystyrene brushes,” Soft Matter, vol. 13, no. 13, pp. 2426–2436, 2017.
6. Q. Yang, Y. Qian, Z. Fan, J. Lin, D. Wang, J. Zhong, and M. Oeser, “Exploiting the synergetic effects of graphene and carbon nanotubes on the mechanical properties of bitumen composites,” Carbon, vol. 172, pp. 402–413, 2021.
7. W. Sha, Y. Li, S. Tang, J. Tian, Y. Zhao, Y. Guo, W. Zhang, X. Zhang,S. Lu, Y.-C. Cao et al., “Machine learning in polymer informatics,” InfoMat, vol. 3, no. 4, pp. 353–361, 2021.
8. Y. Zhang and X. Xu, “Machine learning glass transition temperature of polymers,” Heliyon, vol. 6, no. 10, p. e05055, 2020.
9. D. R. Cassar, A. C. de Carvalho, and E. D. Zanotto, “Predicting glass transition temperatures using neural networks,” Acta Materialia, vol. 159, pp. 249–256, 2018.
10. C. Wen, B. Liu, J. Wolfgang, T. E. Long, R. Odle, and S. Cheng, “Determination of glass transition temperature of polyimides from atom- istic molecular dynamics simulations and machine-learning algorithms,” Journal of Polymer Science, vol. 58, no. 11, pp. 1521–1534, 2020.
11. L. A. Miccio and G. A. Schwartz, “Localizing and quantifying the intra- monomer contributions to the glass transition temperature using artificial neural networks,” Polymer, vol. 203, p. 122786, 2020.
12. S. Goswami, R. Ghosh, A. Neog, and B. Das, “Deep learning-based approach for prediction of glass transition temperature in polymers,” Materials Today: Proceedings, vol. 46, pp. 5838–5843, 2021.
13. J. A. Pugar, C. M. Childs, C. Huang, K. W. Haider, and N. R. Washburn, “Elucidating the physicochemical basis of the glass transition temperature in linear polyurethane elastomers with machine learning,” The Journal of Physical Chemistry B, vol. 124, no. 43, pp. 9722–9733, 2020.
14. N. Andraju, G. W. Curtzwiler, Y. Ji, E. Kozliak, and P. Ranganathan, “Machine-learning-based predictions of polymer and postconsumer re- cycled polymer properties: A comprehensive review,” ACS Applied Materials & Interfaces, vol. 14, no. 38, pp. 42 771–42 790, 2022.
15. S. Goswami, R. Ghosh, A. Neog, and B. Das, “Deep learning-based approach for prediction of glass transition temperature in polymers,” Materials Today: Proceedings, vol. 46, pp. 5838–5843, 2021.
16. G. Pilania, “Machine learning in materials science: From explainable predictions to autonomous design,” Computational Materials Science, vol. 193, p. 110360, 2021.
17. D. Bhattacharya and T. K. Patra, “dpoly: Deep learning of polymer phases and phase transition,” Macromolecules, vol. 54, no. 7, pp. 3065– 3074, 2021.
18. W. B. Chaabene, M. Flah, and M. L. Nehdi, “Machine learning predic- tion of mechanical properties of concrete: Critical review,” Construction and Building Materials, vol. 260, p. 119889, 2020.
19. T. K. Patra, “Data-driven methods for accelerating polymer design,” ACS Polymers Au, vol. 2, no. 1, pp. 8–26, 2021.
20. S. Arrhenius,“U¨ ber die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch sa¨uren,” 1889.
21. D. Thakur and P. Lalwani, “A cuckoo search-based optimized ensemble  
    model (csoem) for the analysis of human gait,” Journal of Intelligent &  
    Fuzzy Systems, no. Preprint, pp. 1–14.