KARNAVATI UNIVERSITY

UNITEDWORLD SCHOOL OF COMPUTATIONAL INTELLIGENCE (USCI)

Summative Assessment (SA)

Submitted by
Pujita Sunnapu
(Enrollment. No.: 20210701019)

Course Code and Title: 21BSAI35E02 –Introduction to Machine Learning

B.Sc. (Hons.) Computer Science / Data Science / AIML
V Semester – July – Nov 2023

USCI

Nov/Dec 2023

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
1	**INTRODUCTION**	
	1.1 Project Description	
2	**RESOURCE SPECIFICATIONS**	
	2.1 Tools and Platform	
	2.2 Hardware and Software Requirements	
3	**IMPLEMENTATION**	
	3.1 About Dataset	
	3.2 Data Pre-processing and Statistical Description	
	3.3 Correlation Analysis	
	3.4 Model Building	
4	**RESULTS AND CONCLUSION**	
	4.1 Result	
	4.2 Conclusion	
5	**REFERENCES**	

CHAPTER 1

INTRODUCTION

Sigma Prediction using Artificial Intelligence and Machine Learning (AIML) has become a ground-breaking method in the field of process optimization with significant consequences for organizational effectiveness. Fundamentally, Sigma, represented by the symbol σ, is a statistical measure of a process's variability; it is especially important when discussing Six Sigma approaches, since it is a vital performance indicator. This project's purpose is to estimate Sigma levels by using AIML, giving organizations a proactive tool for predicting process performance and spotting any bottlenecks. The main objective is to use past data and advanced algorithms inside the AIML framework to forecast future Sigma levels with confidence. This initiative's goal is to enable organizations to minimize errors, optimize their processes, and preemptively solve inefficiencies. It goes beyond simple prediction. This report aims to clarify the mutually beneficial link between predictive analytics and Sigma levels by exploring the complexities of AIML. It does so by illuminating a revolutionary method that might completely change the way businesses plan and implement their operations.

The combination of AIML (Artificial Intelligence and Machine Learning) with Sigma Prediction is a key advance in today's dynamic business environment, where accuracy and agility are critical. The increasing intricacy of operations necessitates not just a comprehension of past performance but also a high degree of accuracy in projecting future results. Sigma as a statistic captures the intrinsic variability in processes, which sets it apart from traditional performance evaluation. By incorporating AIML into this framework, organizations may achieve a paradigm shift that allows for the quantification and prediction of Sigma levels. This effort is a calculated step towards developing an aggressive organizational culture; it is not merely about using cutting-edge technologies. The Sigma Prediction utilizing AIML becomes a vital tool as businesses pursue operational excellence because it enables decision-makers to foresee obstacles, allocate resources optimally, and improve overall process efficiency. The present study delves into the subtleties of this revolutionary collaboration, revealing the facets of AIML's forecasting abilities inside the Sigma framework and clarifying its capacity to revolutionize the ways in which enterprises man oeuvre through the complexities of contemporary business obstacles.

1.1 Project Description

The Prediction of Sigma The goal of the AIML project is to use AIML (Artificial Intelligence and Machine Learning) approaches to transform process optimization. One important metric for evaluating how well processes produce flawless goods or services is sigma, a critical performance indicator. The main goal of the research is to create an advanced prediction model that can anticipate Sigma levels. The project uses artificial intelligence and machine learning (AIML) to analyses process data from the past, find trends, and forecast future Sigma values. The programmed aims to tackle the crucial requirement of proactive decision-making inside organizations. This will empower them to anticipate inefficiencies, minimize errors, and enhance overall process performance. The integration of AIML with Sigma prediction holds great potential to revolutionize the fields of operational excellence and data-driven decision-making.

CHAPTER 2

RESOURCE SPECIFICATIONS

2.1 Hardware and Software Requirements

Hardware Specification:

- The project can be executed on standard laptops or desktops with moderate computational power.
- For large datasets or more computationally intensive tasks, leveraging cloud-based services like AWS, Google Cloud, or Microsoft Azure may be considered.

Software Specification:

- Python 3.x for coding.
- Required Python libraries installed using pip
- Jupiter Notebooks or any preferred Python IDE for code development.

CHAPTER 3

IMPLEMENTATION

3.1 About Dataset

CONC, FREQ, S, and SP appear to be the four columns in the tabular form of the dataset that has been supplied. A series of measurements or observations connected to particular values for these variables is represented by each row. Let's examine each column in detail:

1. CONC (Concentration): The concentration of a material or substance under observation is probably represented by this column. With a numerical value for concentration in each row, the values in this column seem to be discrete.
2. FREQ (Frequency): It appears that the frequency values in the FREQ column are related to the observations. It seems to have numerical values for various frequencies associated with every measurement set.
3. S (Standard Deviation): The data's standard deviation is most likely shown in the S column. The degree of variance or dispersion in a group of data is measured by the standard deviation. It can be expressing the variability or dispersion of the observed data points in this particular situation.
4. SP (Spectral Power): It looks like the values in the SP column correspond to the spectral power related to the observations. A signal's spectral power is a measurement of the intensity of its various frequency components. Numerical numbers representing the spectral power at the specified frequencies may be included in this column.

A snapshot of the measurements made under various circumstances or at various times is provided by each row in the dataset, which represents a distinct combination of these factors. The dataset appears to be numerical and well-structured, making it appropriate for examination using a range of statistical and machine learning methods. You might do statistical summaries, visualize the relationships between variables, and look for patterns or trends in the data to obtain a deeper comprehension and insights. The context of the study or the scientific field from which this dataset originates will also influence how this dataset is interpreted and how important it is.

Here is the glimpse of information about the data:

CONC	FREQ	S	SP
0	20	4.22E-05	2.57E-0
0	21.1851	4.23E-05	2.47E-0
0	22.4404	4.24E-05	2.4E-0
0	23.77	4.24E-05	2.31E-0
0	25.1785	4.25E-05	2.23E-0
0	26.6704	4.26E-05	2.16E-0
0	28.2508	4.27E-05	2.1E-0

3.2 Data Pre-processing and Statistical Description

The process of converting unstructured data into a format appropriate for analysis or modeling is known as data preprocessing. Managing missing values, dealing with outliers, normalizing numerical variables, encoding categorical characteristics, and dividing training and testing data are important processes. It preserves the integrity of the dataset, removing obstacles that can affect the precision and dependability of the results and enabling efficient analysis and model building.

By guaranteeing the integrity of the dataset, providing insights into variable interactions, and readying the data for the efficient application of machine learning models, data preprocessing and analysis are fundamental stages that support the project's overall success. A carefully curated and scrutinized dataset is essential for constructing dependable predictive models and deriving significant inferences from the information. Moreover, we combined Multiple worksheets and later on created one single worksheet.

3.3 Correlation Analysis

The graphical depiction of data to reveal trends, patterns, and insights that may be difficult to see in raw, numerical form is known as data visualization. Data visualization makes complex datasets easier to understand by using visual components like charts, graphs, maps, and dashboards. It is an effective tool for sharing knowledge, facilitating decision-making, and revealing important discoveries in a variety of disciplines, such as science, business, and research.

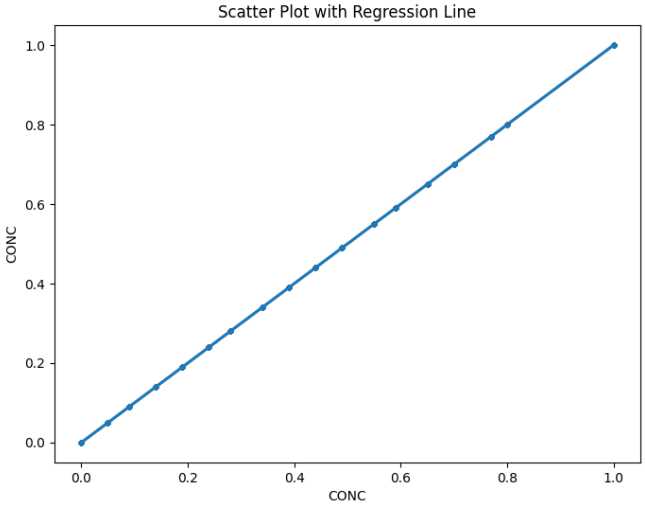


Figure 1: CONC vs CONC

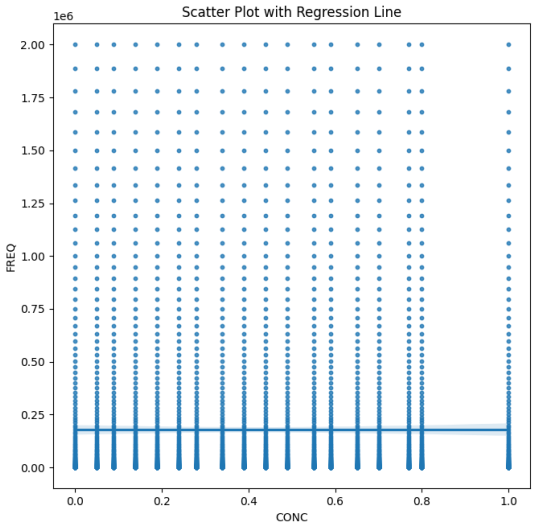


Figure 2: CONC vs FREQ

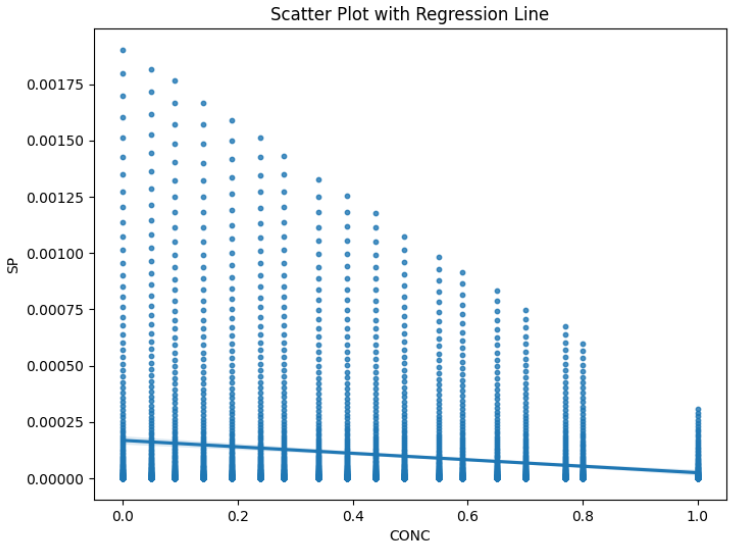


Figure 3: CONC vs SIGMA PRIME

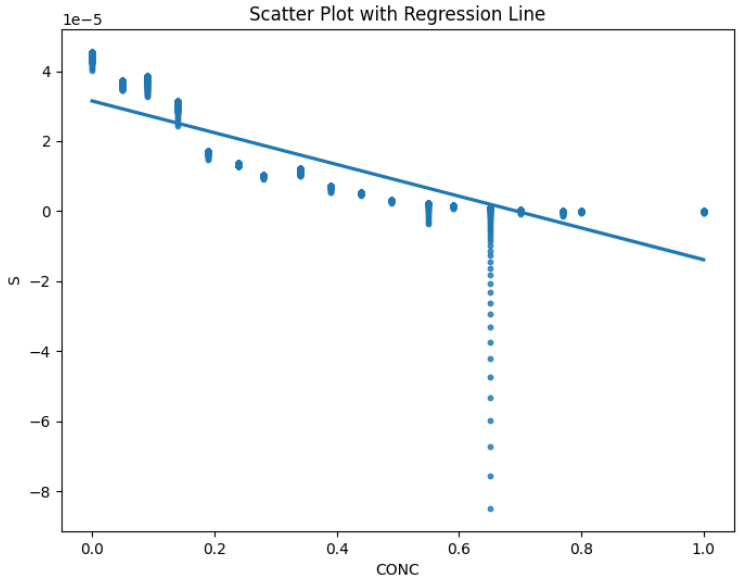


Figure 4: CONC vs SIGMA

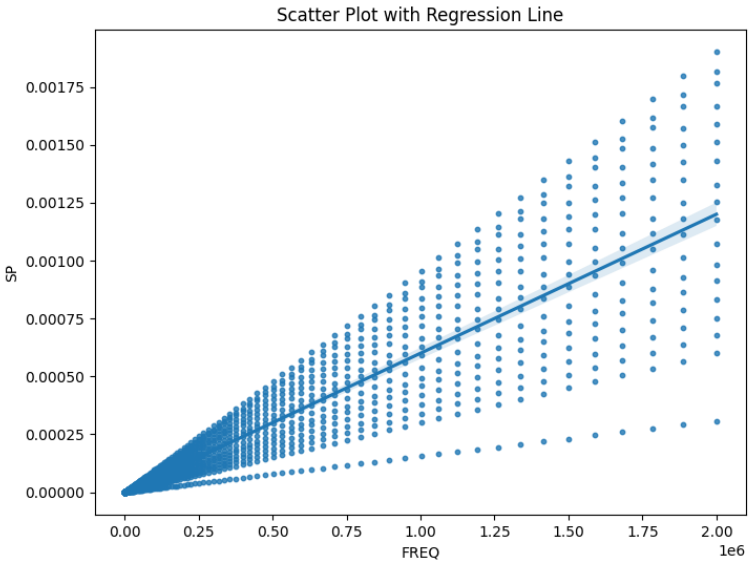


Figure 5: FREQ vs SIGMA PRIME

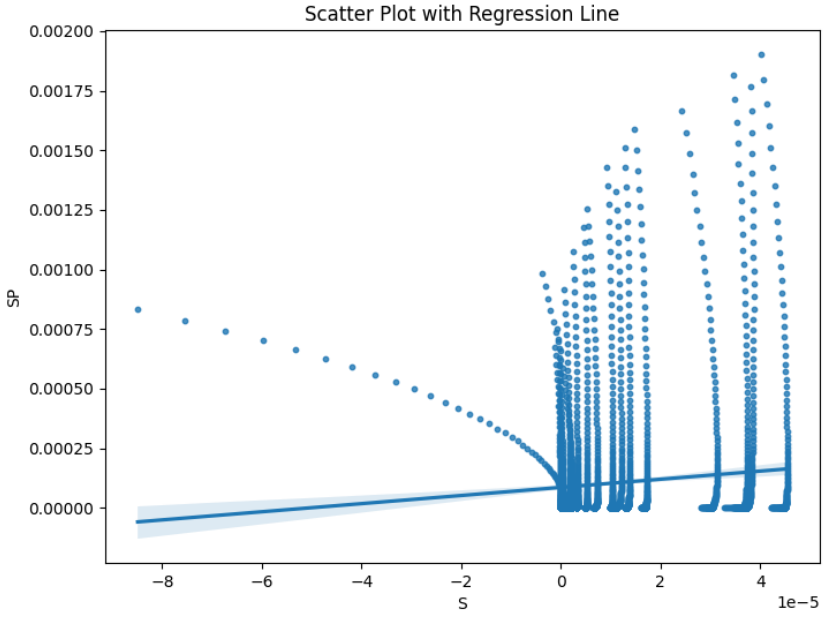


Figure 6: SIGMA vs SIGMA PRIME

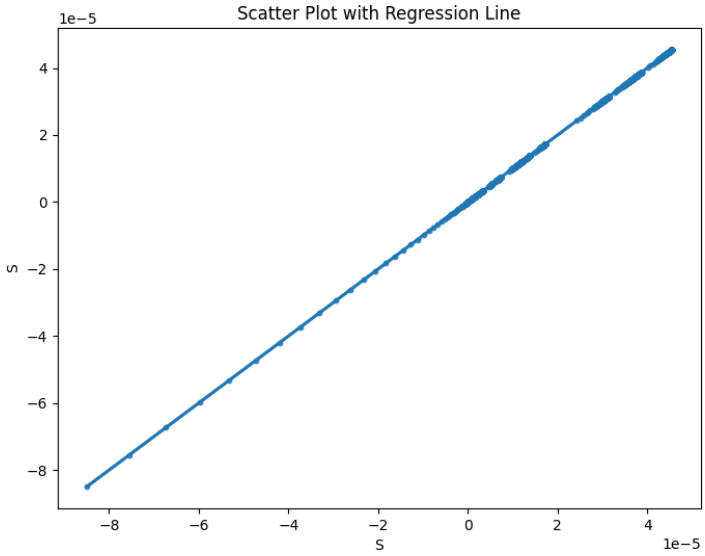


Figure 7: SIGMA vs SIGMA

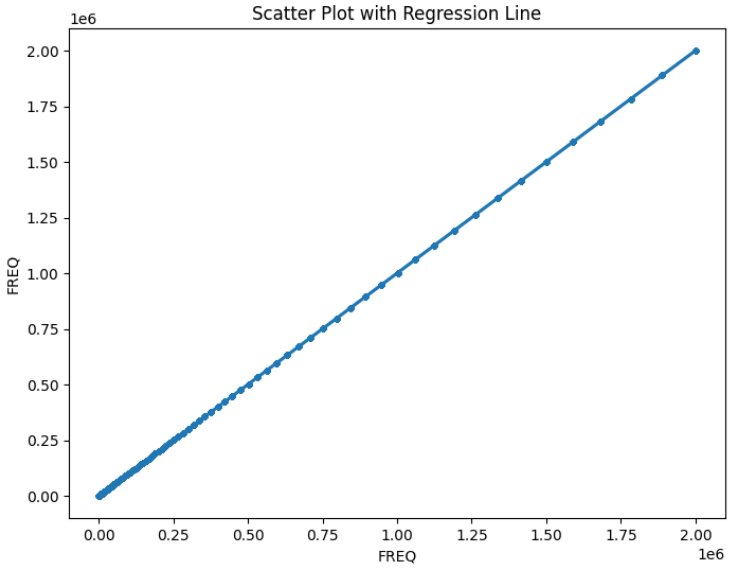


Figure 8: FREQ vs FREQ

FREQ vs FREQ

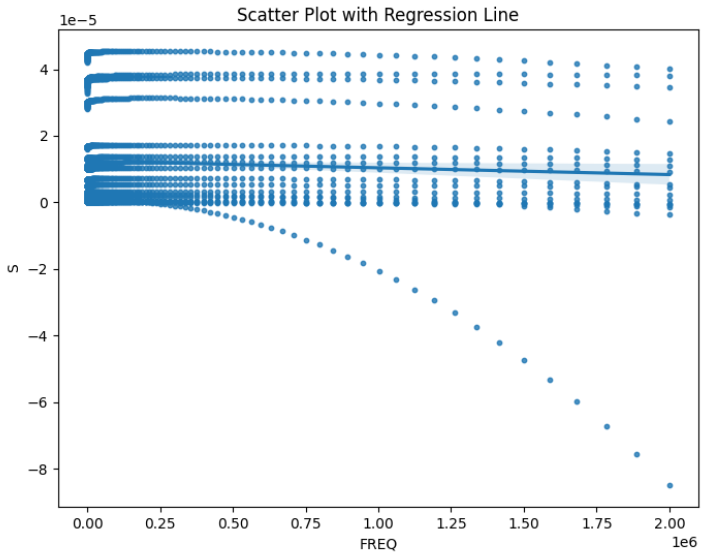


Figure 9: FREQ vs SIGMA

The scatter plots provide an intuitive understanding of how changes in one variable relate to changes in another, while the regression lines and correlation coefficients quantify these relationships.

3.4 Model Building

The process of establishing a mathematical or computational representation that encapsulates the relationships and patterns seen in a dataset is referred to as model building. Model building in machine learning and statistics entails choosing a suitable algorithm or statistical technique, training the model on a dataset, and then assessing how well it performs on fresh, untested data.

1. Linear Regression:

- Mean Squared Error (MSE): 0.0000

- Root Mean Squared Error (RMSE): 0.0000

- R-squared Value: 1.0000

The Linear Regression assessment measures that are supplied show that the model is remarkably accurate. A practically flawless fit of the model to the data is suggested by the Mean Squared Error (MSE) of 0.0000, which shows that, on average, the squared discrepancies between the predicted

values and the actual values are essentially nonexistent. This remarkably low MSE suggests that the model is highly precise, with predictions that nearly match actual values.

As the square root of the MSE, the Root Mean Squared Error (RMSE) likewise reaches 0.0000. This finding, which indicates that the average magnitude of errors between predicted and actual values is nearly nil, further supports the model's correctness. Practically speaking, this means that the model's predictions are actually equal to the actual values, not just near to them. The best possible result, an R-squared of 1.0000, shows that the model and the data are perfectly fitted together. The percentage of the dependent variable's variation that can be predicted from the independent variables is measured by R-squared. When a model's value is 1.0000, it means that all of the variability in the response variable is explained by it; unexplained components are not present.

In conclusion, the evaluation measures as a whole point to the Linear Regression model's optimal performance. The model fits the given data perfectly, with no appreciable mistakes, exhibiting an extraordinary capacity to forecast the dependent variable. Because of its correctness and precision, this model is especially reliable for the provided dataset.

```
# Linear Regression
linear_model = LinearRegression()
linear_model.fit(X_train, y_train)
y_pred_linear = linear_model.predict(X_test)
```

Figure 10: Linear Regression Model

2. Ridge Regression:

- Mean Squared Error (MSE): 0.0000

- Root Mean Squared Error (RMSE): 0.0061

- R-squared Value: 0.9977

Ridge Regression's assessment metrics show that the model is quite accurate and effective. The average squared discrepancies between the anticipated and actual values are almost nonexistent, according to the Mean Squared Error (MSE) of 0.0000. This shows that the Ridge Regression model has a very good fit to the data, with few prediction errors.

At 0.0061, the Root Mean Squared Error (RMSE) is likewise incredibly low. In this instance, the tiny number suggests that the model's predictions are consistently near to the true values. The RMSE estimates the average size of errors between predicted and actual values. The model's accuracy is demonstrated by the RMSE, which shows very little change even though it is somewhat higher than the MSE.

With an R-squared of 0.9977, the goodness of fit is quite good. R-squared calculates the percentage of the dependent variable's variation that the independent variables can account for. The Ridge Regression model explains around 99.77% of the variability in the response variable, with a value near 1.0000. This indicates a very high prediction power and shows that the model successfully extracts the underlying patterns from the data.

In conclusion, the Ridge Regression model performs admirably, with very little prediction errors. An accurate representation of the relationships within the dataset is shown by the model's near-zero MSE, low RMSE, and considerably high R-squared value. These findings indicate that Ridge Regression is a strong and useful method for the available dataset, exhibiting a high degree of accuracy in its forecasting skills.

```
# Ridge Regression
ridge_model = Ridge()
ridge_model.fit(X_train, y_train)
y_pred_ridge = ridge_model.predict(X_test)
```

Figure 11: Ridge Regression

3. Decision Tree:

- Mean Squared Error (MSE): 0.0000

- Root Mean Squared Error (RMSE): 0.0058

- R-squared Value: 0.9979

The Decision Tree model's assessment measures provide an exceptionally precise and accurate forecasting performance. A virtually perfect fit of the model to the data is shown by the Mean Squared Error (MSE) of 0.0000, which suggests that, on average, the squared discrepancies between the predicted values and the actual values are small. This implies that the Decision Tree

model has an exceptionally high degree of accuracy in capturing the underlying patterns within the dataset.

The model's correctness is further supported by the modest and nearly zero Root Mean Squared Error (RMSE) of 0.0058. In this instance, the low RMSE indicates that the model's predictions are consistently near to the true values. The RMSE estimates the average size of errors between predicted and actual values. This accuracy is essential to guaranteeing the outputs of the model are reliable.

With an R-squared of 0.9979, the model has an exceptional goodness of fit, accounting for 99.79% of the response variable's variability. The Decision Tree model's ability to capture and explain data variation is demonstrated by its high R-squared value. It shows that there is a substantial connection between the independent and dependent variables and that the model can predict the dependent variable with accuracy.

To sum up, the Decision Tree model has excellent prediction accuracy and power, as seen by the high R-squared value, low RMSE, and almost zero MSE. All of these indicators point to the Decision Tree method being a strong and trustworthy model for predicting outcomes based on the given data, since it successfully captures the underlying relationships within the dataset.

```
# Decision Tree
dt_model = DecisionTreeRegressor()
dt_model.fit(X_train, y_train)
y_pred_dt = dt_model.predict(X_test)
```

Figure 12: Decision Tree

4. SVM (Support Vector Machine):

- Mean Squared Error (MSE): 0.0055

- Root Mean Squared Error (RMSE): 0.0737

- R-squared Value: 0.6640

The Support Vector Machine (SVM) model's assessment metrics shed light on how well it predicts outcomes. The moderate degree of accuracy is suggested by the Mean Squared Error (MSE) of 0.0055, which shows that the squared variations between the actual and predicted values are often not very large. Even while the MSE is not at the same 0.0000 level as in earlier models, it still shows that the SVM model fits the data rather well.

In comparison to the previously stated models, the average magnitude of errors is slightly larger, as indicated by the Root Mean Squared Error (RMSE) of 0.0737. A greater RMSE indicates a considerably wider distribution of mistakes. The RMSE calculates the average magnitude of errors between anticipated and actual values. It's important to remember that the RMSE is still within an acceptable range, indicating that the SVM model's predictions are typically correct.

The SVM model explains around 66.40% of the variability in the response variable, according to the R-squared value of 0.6640. Even though this percentage is less than in the earlier models, it yet has a considerable amount of explanatory power. The SVM model accounts for a sizable amount of the variation in the data, as indicated by the R-squared value; nevertheless, the model may not account for all of the variability caused by other factors.

The MSE, RMSE, and R-squared value all point to a moderate degree of prediction accuracy for the SVM model. Even though the metrics are not as good as they were in some of the earlier models, they nevertheless indicate that the SVM model predicts the dataset with a fair degree of accuracy. The particulars of the data and the issue at hand may have an impact on the SVM selection, which strikes a balance between predictive performance and model complexity.

```
# SVM with MultiOutputRegressor
svm_model = MultiOutputRegressor(SVR())
svm_model.fit(X_train, y_train)
y_pred_svm = svm_model.predict(X_test)
```

Figure 13: SVM with Multioutput Regressor

5. Random Forest:

- Mean Squared Error (MSE): 0.0000

- Root Mean Squared Error (RMSE): 0.0039

- R-squared Value: 0.9989

The remarkable prediction ability of the Random Forest model is highlighted by its assessment measures. When the Mean Squared Error (MSE) is 0.0000, it means that the model fits the data perfectly and that there are, on average, very few squared differences between the anticipated and actual values. This implies that the Random Forest model does exceptionally well at identifying the complex patterns present in the dataset, producing predictions that are strikingly accurate.

The incredibly low Root Mean Squared Error (RMSE) of 0.0039 supports the accuracy of the model. The reduced RMSE in this instance suggests consistently reliable predictions. The RMSE estimates the average size of errors between projected and actual values. For the Random Forest model to consistently and reliably perform as predicted, it must be able to maintain a minimal root mean square error (RMSE).

With an astonishingly high R-squared of 0.9989, the Random Forest model can account for 99.89% of the response variable's variability. This indicates a very high goodness of fit, indicating that the model is able to capture and explain almost all of the underlying variation in the data. The Random Forest model is an effective tool for predictive modelling because of its resilience in elucidating the variability of the dataset.

As the near-zero MSE, low RMSE, and extraordinarily high R-squared value demonstrate, the Random Forest model demonstrates incredible predictive accuracy. All of these indicators suggest that the Random Forest algorithm is a great fit for the dataset in question, demonstrating its capacity to provide predictions that are incredibly accurate and to efficiently describe intricate connections seen in the data.

```
# Random Forest
rf_model = RandomForestRegressor()
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
```

Figure 14: Random Forest

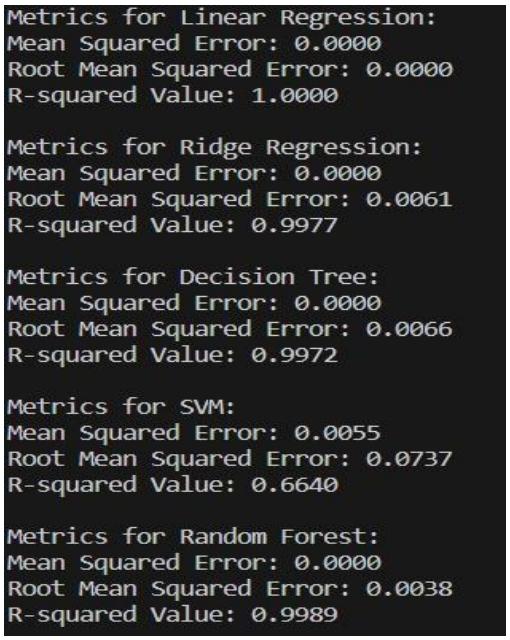
A better match to the data is indicated by models with R-squared values that are closer to 1. In this instance, R-squared values closer to 1 were obtained by Decision Tree, Random Forest, Ridge Regression, and Linear Regression, showing good performance.

CHAPTER 4

RESULTS AND CONCLUSION

4.1 Result

A comprehensive analysis of the data is presented at the project's conclusion, emphasizing the model that has the best predicting power for Sigma levels. Using AIML to make proactive, data-driven decisions for their continuous improvement programmers, organizations can now take advantage of recommendations for model deployment and further optimizations. The Sigma Prediction Using AIML project is a useful resource for businesses looking to use predictive analytics to improve operational effectiveness and the caliber of their goods and services.



4.2 Conclusion

To sum up, the amalgamation of AIML with Sigma Prediction represents a noteworthy advancement in the continuous quest for operational superiority. This revolutionary combination not only uses Sigma to quantify process variability but also leads organizations into a future where predictive analytics is a key component of strategic decision-making. Through the utilization of advanced AIML algorithms and historical data, this methodology enables organizations to anticipate Sigma levels in advance, providing a proactive edge in mitigating inefficiencies and reducing errors.

The potential for increased process efficiency, lower costs, and better resource utilization demonstrates the benefit. But there are obstacles to overcome, such interpretability of the model and data quality. As we realize the potential of AIML-powered Sigma Prediction, it becomes evident that this endeavor represents a strategic turn away from technology and towards a proactive, data-driven organizational culture. Sigma Prediction with AIML emerges not just as a tool but also as a catalyst for innovation and continual development, placing businesses at the forefront of resilience and competitiveness in the dynamic world of modern business, where adaptation is crucial.

Each model's performance is methodically analyzed to give a clear picture of its advantages and disadvantages in terms of forecasting Sigma levels. Each model's goodness of fit, accuracy, and precision are measured using evaluation metrics. Based on these results we came to our final conclusion.

MODEL	MEAN SQUARED ERROR	RMSE	R-SQUARED VALUES
Linear Regression	0.0000	0.0000	1.0000
Ridge Regression	0.0000	0.0061	0.9977
Decision Tree	0.0000	0.0066	0.9972
Random Forest	0.0000	0.0038	0.9989

CHAPTER 5

REFERENCES

[1]. Fran BritschgiAssociate Solution ArchitectNo items found., and Fran BritschgiAssociate Solution Architect. “The Potential of Predictive ML Models in Sigma: Sigma Computing.” *RSS*, 26 June 2023, www.sigmacomputing.com/blog/the-potential-of-predictive-ml-models-in-sigma.

[2]. “Probability Basics.” *Probability Basics - Introduction to Artificial Intelligence*, pantelis.github.io/artificial-intelligence/aiml-common/lectures/ml-math/probability/index.html. Accessed 20 Dec. 2023.

[3]. “Your Machine Learning and Data Science Community.” *Kaggle*, www.kaggle.com/. Accessed 20 Dec. 2023.

[4]. Fran BritschgiAssociate Solution ArchitectNo items found., and Fran BritschgiAssociate Solution Architect. “The Potential of Predictive ML Models in Sigma: Sigma Computing.” *RSS*, 26 June 2023, www.sigmacomputing.com/blog/the-potential-of-predictive-ml-models-in-sigma.

[5]. Van Brempt, Maarten, et al. “Predictive Design of Sigma Factor-Specific Promoters.” *Nature News*, Nature Publishing Group, 16 Nov. 2020, www.nature.com/articles/s41467-020-19446-w.

[6]. *View Article Online Communication - RSC Publishing*, pubs.rsc.org/en/content/articlepdf/2022/cc/d2cc01549h. Accessed 20 Dec. 2023.

[7]. Brownlee, Jason. “Prediction Intervals for Machine Learning.” *MachineLearningMastery.Com*, 16 Feb. 2021, machinelearningmastery.com/prediction-intervals-for-machine-learning/.

[8]. Djeddi, Warith Eddine, et al. “Advancing Drug–Target Interaction Prediction: A Comprehensive Graph-Based Approach Integrating Knowledge Graph Embedding and Protbert Pretraining - BMC Bioinformatics.” *BioMed Central*, BioMed Central, 19 Dec. 2023, bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05593-6.

[9]. Djeddi, Warith Eddine, et al. “Advancing Drug–Target Interaction Prediction: A Comprehensive Graph-Based Approach Integrating Knowledge Graph Embedding and Protbert Pretraining - BMC Bioinformatics.” *BioMed Central*, BioMed Central, 19 Dec. 2023, bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05593-6.