

# **SECB3203 (01) – PROGRAMMING FOR BIOINFORMATICS**

# **FACULTY OF COMPUTING**

**TITLE:**

**"UNRAVELING ALZHEIMER’S DISEASE MYSTERIES**

**THROUGH PROTEIN ANALYSIS”**

**NAME:**

# **MAATHUREE A/P VEERABALAN (A21EC0051)**

**QAISARA BT BADRUL HISHAM (A21EC0125)**

# **LECTURER’S NAME:**

**DR. NIES HUI WEN**

# **TABLE OF CONTENTS**

|  | **CONTENT** | **PAGE** |
| --- | --- | --- |
| **1.0** | **INTRODUCTION** | 3 - 4 |
| **1.1** | **PROBLEM BACKGROUND** | 5 - 6 |
| **1.2** | **PROBLEM STATEMENT** | 7 |
| **1.3** | **OBJECTIVES** | 7 |
| **1.4** | **SCOPES** | 8 |
| **1.5** | **CONCLUSION** | 9 |
| **2.0** | **IMPORTING DATA** |  |
| **3.0** | **DATA WRANGLING** | |
| **3.1** | **IDENTIFYING MISSING VALUES** |  |
| **3.2** | **HANDLING MISSING VALUES** |  |
| **3.3** | **DATA FORMATTING** |  |
| **3.4** | **DATA BINNING** |  |
| **3.5** | **INDICATOR VARIABLES** |  |

**1.0 INTRODUCTION**

In the realm of neurodegenerative disorders, Alzheimer's and prion diseases have long been the focus of intense study due to their common feature: the aberrant processing of amyloid-β (Aβ) peptide and prion protein (PrPC), respectively (Katherine A.B. Kellett & Nigel M. Hooper (2009) Prion protein and Alzheimer disease, Prion, 3:4, 190-194, DOI: 10.4161/pri.3.4.9980). Recent insights have begun to illuminate a potential critical link between PrPC and the pathogenesis of Alzheimer's disease, offering a promising avenue for novel detection methods that can even extend to individuals involved in criminal activities.

Crucially, PrPC's multifaceted interactions with Alzheimer's pathology have come to the fore. It has been discovered that PrPC serves as a regulatory force, inhibiting the activity of β-secretase BACE1, the enzyme pivotal in Aβ production (Katherine A.B. Kellett & Nigel M. Hooper (2009) Prion protein and Alzheimer disease, Prion, 3:4, 190-194, DOI: 10.4161/pri.3.4.9980). Data suggest that PrP promotes plaque formation, and that this hitherto unknown functional role of PrP appears to be mediated by increased Aβ aggregation rather than by altered APP transcription or processing (Schwarze-Eicker, K., Keyvani, K., Görtz, N., Westaway, D., Sachser, N., & Paulus, W. (2005, August 1). Prion protein (PrPc) promotes β-amyloid plaque formation. Neurobiology of Aging; Elsevier BV. https://doi.org/10.1016/j.neurobiolaging.2004.10.004). This interaction creates a feedback loop in the normal brain, where PrPC exerts its influence to decrease both Aβ and amyloid intracellular domain (AICD) production. In response, AICD acts as an upregulator, reinforcing the inhibitory role of PrPC on BACE1. This delicate balance in the normal brain is where science and law enforcement intersect.

The emerging understanding of PrPC's role in Alzheimer's disease has profound implications, especially when it comes to detecting individuals with Alzheimer's, including those involved in criminal activities. By leveraging advanced protein analysis methods, we can envision a future where the abnormal protein interactions seen in Alzheimer's pathology, such as the binding of Aβ oligomers to PrPC, become diagnostic markers (Katherine A.B. Kellett & Nigel M. Hooper (2009) Prion protein and Alzheimer disease, Prion, 3:4, 190-194, DOI: 10.4161/pri.3.4.9980).

In Alzheimer's disease, this feedback loop is disrupted, and the increased level of Aβ oligomers impede PrPC's ability to regulate BACE1 activity. This disruption creates an opportunity for utilizing protein analysis techniques to detect individuals with Alzheimer's disease, particularly in cases where cognitive impairment may affect their criminal culpability (Katherine A.B. Kellett & Nigel M. Hooper (2009) Prion protein and Alzheimer disease, Prion, 3:4, 190-194, DOI: 10.4161/pri.3.4.9980).

As we venture further into this groundbreaking intersection of science and law enforcement, we have the potential to transform not only our understanding of Alzheimer's disease but also the way we approach the detection of individuals affected by this condition, including those entangled in criminal activities. This journey into the intricate molecular and cellular mechanisms holds promise for a more compassionate and equitable approach to criminal justice, one that adapts to the complexities of the human condition.

**1.1 PROBLEM BACKGROUND**

Criminal convictions as well as sentences nowadays stray from being simply black and white. There are a lot of contributing factors that influence these sentences. It is a complex process that has an objective to balance the principles of justice. The specifics of these conditions may vary based on the jurisdiction. A study of sentencing was conducted in the Hennepin County and Ramsey County District Courts by William M. Rhodes stated that the principal factors that have an impact on the duration of sentencing includes but are not limited to severity of crime, criminal history, victim impact statement, plea bargain, judicial discretion, socioeconomic factors and of course intent.

Choosing the focal point of this study to be intent calls into question why do people commit crimes and could it help as a legal defense. To put it simply no one unless you are a psychopath would intentionally commit a crime without any driving factors affecting their actions. Comprehension of the components that are contributing to the commission of a crime has often relied on a variety of disciplines. Some of those disciplines include psychology, sociology and forensic science. Let’s narrow down the field slightly to look at the psychological aspect of these disciplines.

Often criminal lawyers would adopt a strategy of claiming insanity where the defendant claims to be in some sort of mental influence that could impair judgment. This is considered an affirmative defense arguing that the perpetrator could not be held responsible for their action during the commission of the crime attributed to a psychiatric disease. Pleading insanity or diminished capacity is essentially pleading not guilty hence pleading to a lesser crime. However, insanity defense often causes an uproar among the public due to the subjectivity of the matter and not really having a solid benchmark in which prosecutors could analyze with. This is why the M’Naghten Rule was established. The rule created a presumption of sanity unless the defense proved "at the time of committing the act, the accused was laboring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing or, if he did know it, that he did not know what he was doing was wrong." As this is just an example note that there are other standard operating procedures when it comes to convicting and sentencing relating to the insanity defense. They can range based on country, jurisdiction and sometimes looked upon on a case by case basis.

Nonetheless in spite of the efforts put to quantify the validity of this defense tactic it is still heavily reliant on human’s subjectivity in a sense of needing a psychiatrist and or psychologist to carry out an assessment. In this case the risk of human error is too great to leave it up to chance and or luck. This is the inadequacy that could potentially hinder a defense or could easily let a criminal walk free. As a consequence there should be a more computable alternative that will produce a more definitive conclusion based on the evidence presented by the defendant.

**1.2 PROBLEM STATEMENT**

The commission of a crime is attributed to multiple rationality. It is shaped by a complex interplay of environment, genetic and psychological factors. Traditionally criminological research is lacking in the aspect of biological analysis in understanding its role in behavior modifications. Up to this point researchers have only targeted the sociological and psychological contributors to criminal behavior. To address this gap, our thesis is going to analyze some biomarkers that could be attributed to being the root cause of some mental illnesses. We then hope that by using several analyzing techniques that we are able to identify and characterize specific molecular contributors to criminal acts, therefore increasing our comprehension of the multifactorial nature of a crime.

# **1.3 OBJECTIVE OF STUDY**

The followings are the objectives proposed for this study:

* + 1. To prove that there are biomarkers that affect psychological issues that can indirectly impaired an individual's judgment and actions.
    2. To be able to come up with a more measurable way in diagnosing psychological ailments.
    3. To examine the Corticosteroid Receptors; based on the dataset explore the position corticosteroid receptors hold in individuals with a specific mental disorder.
    4. To analyze mRNA and Protein Expression; analyzing these factors of these receptors in the prefrontal cortex of study participants on a molecular level to identify the differences between them.
    5. To authenticate and duplicate the observed result by replicating the steps and analysis with a contrasting dataset to come to a more concrete conclusion.

# **1.4 SCOPE OF STUDY**

This project delves into the multifaceted roles of the prion protein (PrPC) within the context of Alzheimer's disease (AD) to understand its implications for the disease's pathogenesis. The scope of this project are:

1. Using mass spectrometry, proteomics, data mining and machine learning, pattern recognition, cross validation, dimensionality reduction, ensemble methods, feature engineering and statistical analysis techniques for Alzheimer's disease classification.
2. Using Demographic Details of cases from the TRC Cohort and the Stanley Foundation Array Cohort.
3. Using performance measurement classification such as prion protein localization accuracy, specificity for prion protein interaction, sensitivity of true positive rate, specificity of true negative rate and precision of positive and negative predictive value.

# **1.5 CONCLUSION**

In conclusion, the evolving understanding of PrPC's multifaceted roles in Alzheimer's disease (AD) has the potential to bridge the gap between scientific inquiry and real-world applications, particularly within the realm of criminal investigations. Recent data highlight two key roles for PrPC in AD: its involvement in the physiological regulation of amyloid precursor protein (APP) processing through its interaction with BACE1, and its contribution to the pathological progression of AD by binding Aβ42-oligomers. These revelations provide the foundation for a comprehensive model, connecting the dots among PrPC, BACE1, APP, and AICD, thus enhancing our comprehension of the intricacies of AD pathogenesis (Katherine A.B. Kellett & Nigel M. Hooper (2009) Prion protein and Alzheimer disease, Prion, 3:4, 190-194, DOI: [10.4161/pri.3.4.9980](https://doi.org/10.4161/pri.3.4.9980)).

The implications of this research reach beyond the laboratory, offering new possibilities for law enforcement agencies. As we strive to unlock the mysteries surrounding the interactions between PrPC and APP/Aβ, we discover an unexpected intersection between Alzheimer's disease and criminal activities. By harnessing the power of protein analysis, we can envision a future in which departments can more effectively detect individuals with Alzheimer's disease, including those engaged in criminal actions (Katherine A.B. Kellett & Nigel M. Hooper (2009) Prion protein and Alzheimer disease, Prion, 3:4, 190-194, DOI: [10.4161/pri.3.4.9980](https://doi.org/10.4161/pri.3.4.9980)). This is especially crucial in cases where cognitive impairment might impact an individual's culpability.

Yet, while we peer into this uncharted territory, several questions linger, challenging us to delve deeper into this convergence of science and criminal justice. What, for instance, is the precise effect of Aβ42-oligomer binding on PrPC's functions? How do the levels of PrPC in the brains of AD patients compare to those of age-matched controls? What happens when we manipulate PrPC levels in AD mouse models ? These are critical inquiries that warrant immediate and thorough investigation.

In summary, the fusion of science and real-world application is poised to transform not only our understanding of AD but also the way in which we approach the detection and handling of individuals with Alzheimer's disease, particularly those involved in criminal activities. This journey into the intricate molecular and cellular mechanisms at play offers the promise of a more compassionate and equitable approach to criminal justice, one that adapts to the complexities of the human condition.

**2.0 Importing Dataset**

The dataset used for this project was accumulated from Kaggle.com titled Alzheimer's Disease Classifications. This extensive dataset is noted as oasis.csv which covers over 400 specimen Alzheimer diagnosis.

To execute data analytics we first begin by importing the dataset which we have acquired from Kaggle.com prior. This will then lay out a foundation to which will aid in any subsequent analysis and model development that will be done. The dataset imported will be presented through Python.



Based on the code above we were able to import the library named ‘pandas’ as this library is more suited for the task at hand which is data manipulation.

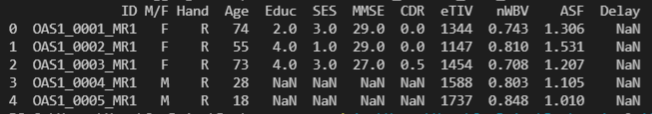
****

The dataset oasis.csv is then loaded with the specification of the file path as well as assigning it to a Data Frame (which is labeled as df).



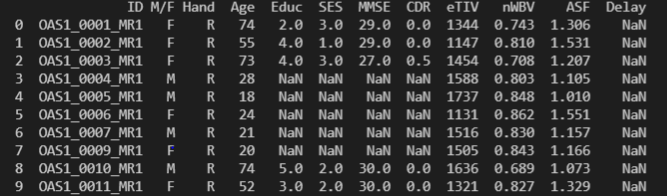
Breaking down the functionality of the code above;

* < **print(df.head()**) > is used to display the first few rows within the dataset which in this case is derived from the oasis.csv file. It prints out 5 rows as that is the default setting.

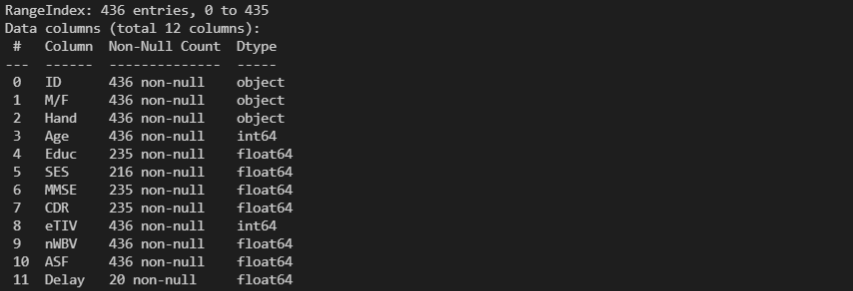


To specify the number of rows wanted to display, simply add an numerator in the bracket.

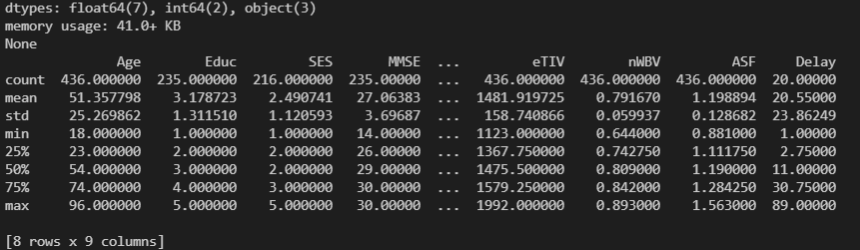
< **print(df.head(10)**) >



* < **print(df.info())** > this command produces a more detailed summary of the Data Frame which includes information regarding the number of non-null values, data types of each column, and the memory usage.

****

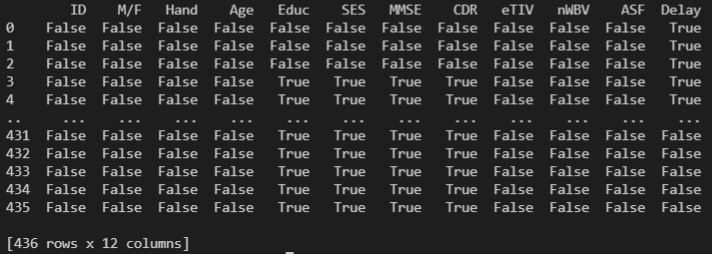
* **< print(df.describe()) >** this command generates the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum. It provides insights into the central tendency and spread of the numerical columns in the dataset.



**3.0 Data Wrangling**

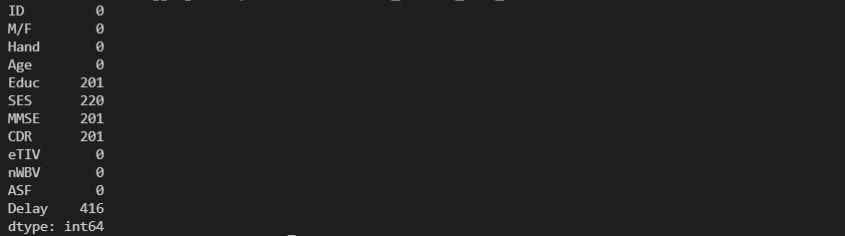
**3.1 Identifying Missing Values**

< print(df.isnull()) > this syntax is used to identify any null values within the dataset.



The false reading indicates that there are no null values and true indicates null values are present in the Data Frame.

To identify how many null values are there as a whole in the wide range of values you can use the syntax < **print(df.isnull().sum())** >



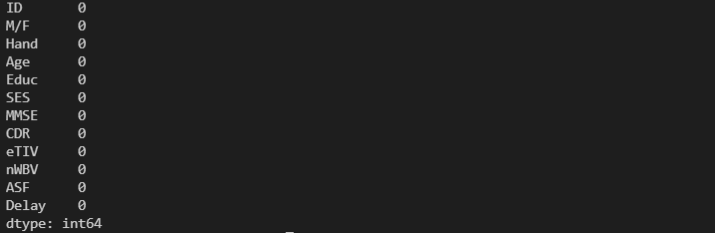
You can now see which columns these null values are residing.

**3.2 Handling Missing Values**

Now there are several ways to handle null or missing values, the way that we have chosen for our project is by filling the null values. This entails that wherever a null value is present it will be filled by a value of our choosing. In this case we will be replacing it with the value 0 using the syntax < df.fillna(value = 0) >



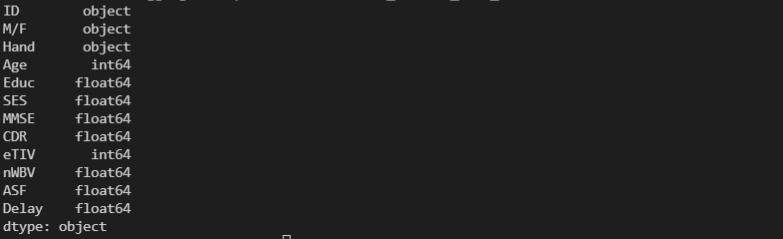
Figure above shows the changing the null values to 0 as well as creating a new Data frame(df2). The results are as shown below.



As the figure illustrates there are no missing values left in the Data Frames.

**3.3 Data Formatting**

The need for data formatting is to ensure that the data is being presented in a unified manner as to make certain that the data is easily readable. Actions regarding data formatting involves converting data types, adjusting decimals, or representing dates uniformly. The syntax for displaying data types is as such < **print(df.dtypes)** >



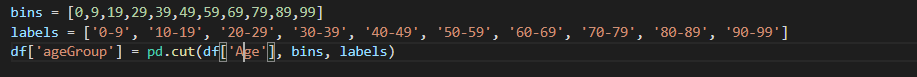
As represented in the figure above the Data Frame consist of 3 main data types;

* Object - consist of multiple data types existing in the same column.
* Int64 - consist of integer data type
* Float64 - consist of floating point data type

Due to the action of filling in the null values done prior we refrain from doing any changes to the current format.

**3.4 Data Binning**

Data binning or data bucketing is an action of grouping bins of data or buckets of data that share somewhat similar characteristics. This will in turn simplify the data as a whole because values in specific intervals can now be represented by a singular representative value. Binning also may improve accuracy in a predictive model.



The figure above shows the process of binning the column Age into several intervals which is '0-9', '10-19', '20-29', '30-39', '40-49', '50-59', '60-69', '70-79', '80-89', '90-99'. The name ‘ageGroup’ will now refer to the new column that has been produced as a result of the binning process. This column will now indicate the age range in which the specimen falls under.

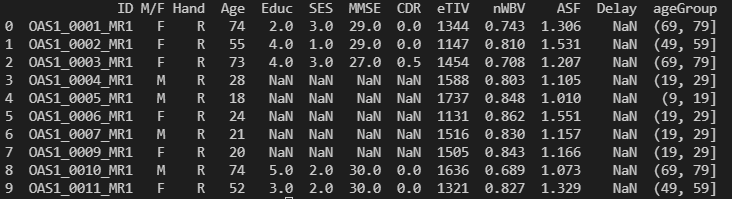


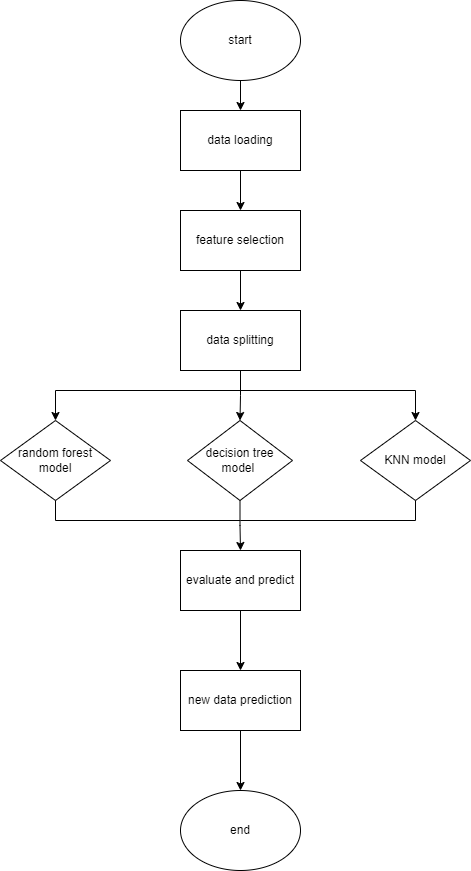
Figure above shows the new column has been created.

**3.5 Indicator Variables**

We believe our projects require no indicator variable steps during the operation of data processing due to not having categorical values within our dataset. Hence there is no need to represent the presence or absence of data of a particular category.

**4.0 Machine learning**

**4.1 Machine learning model development flowchart**

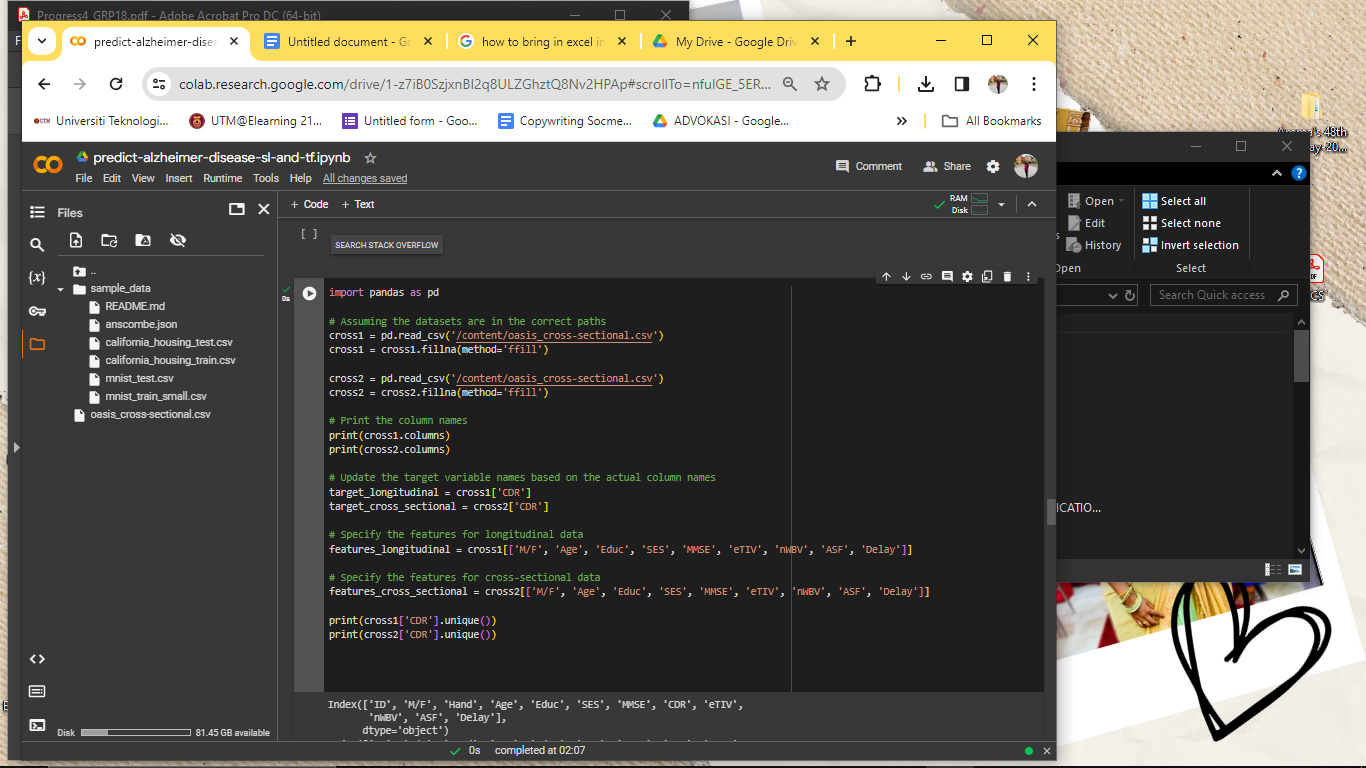
****

**Figure 4.1**

The flowchart above depicts the process of developing the machine learning model to

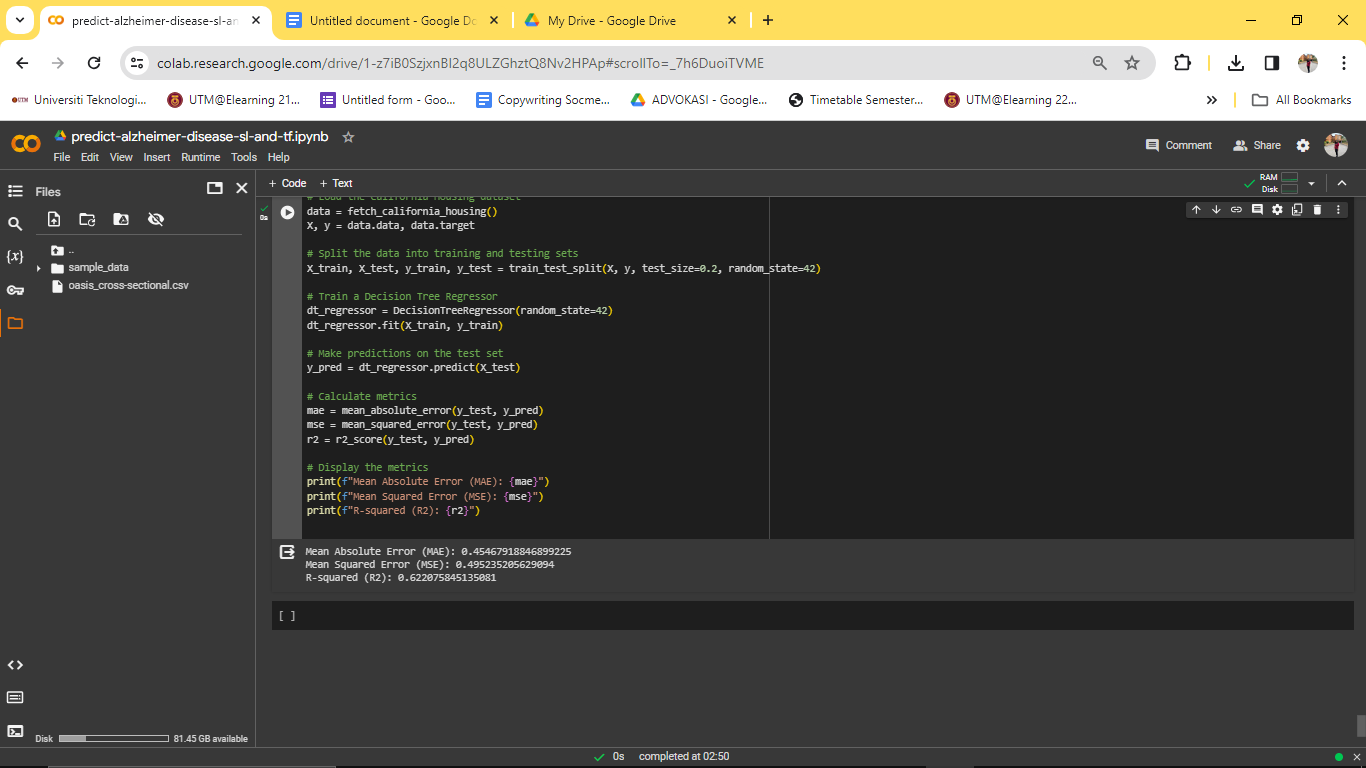
**4.1 Model Training**

**Features and Target**

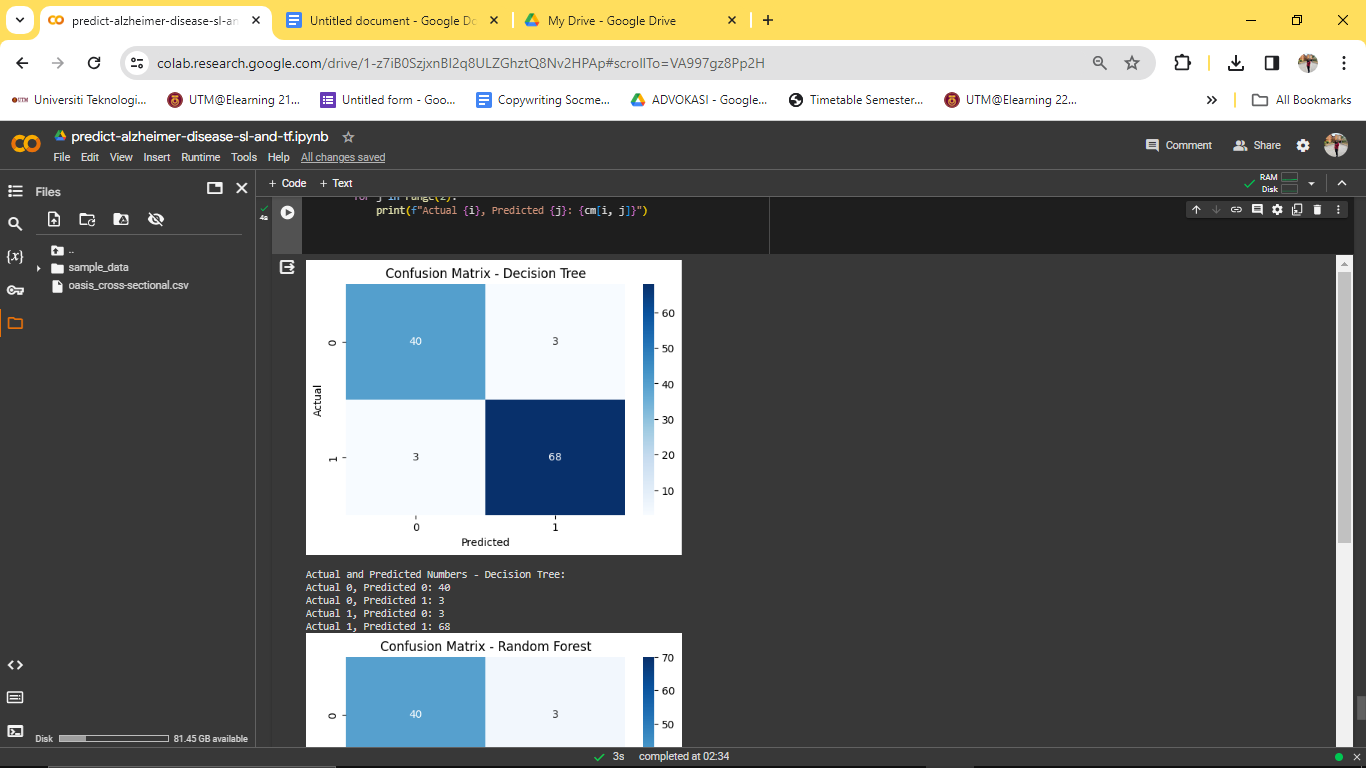
****

**Figure 4.2**

1. **Decision Tree Regression**



**Figure 4.3**

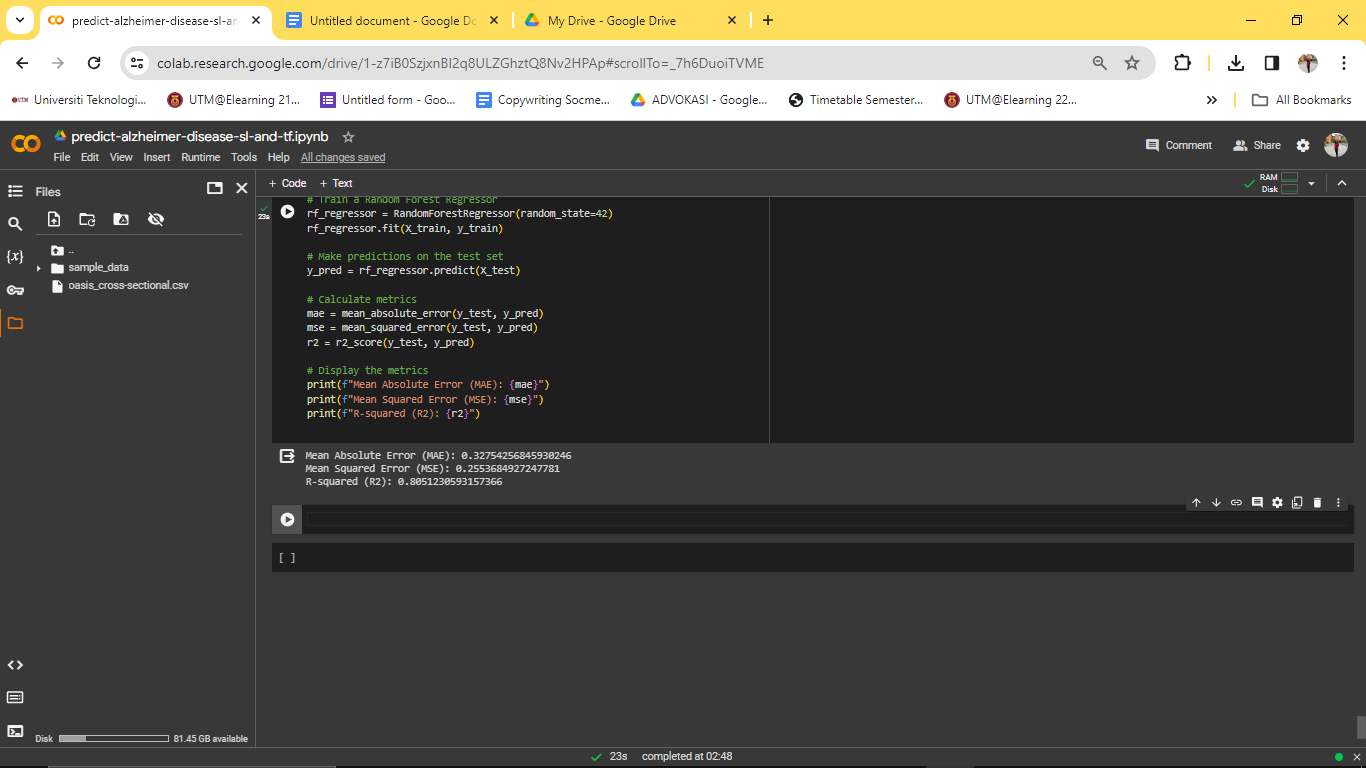


**Figure 4.4**

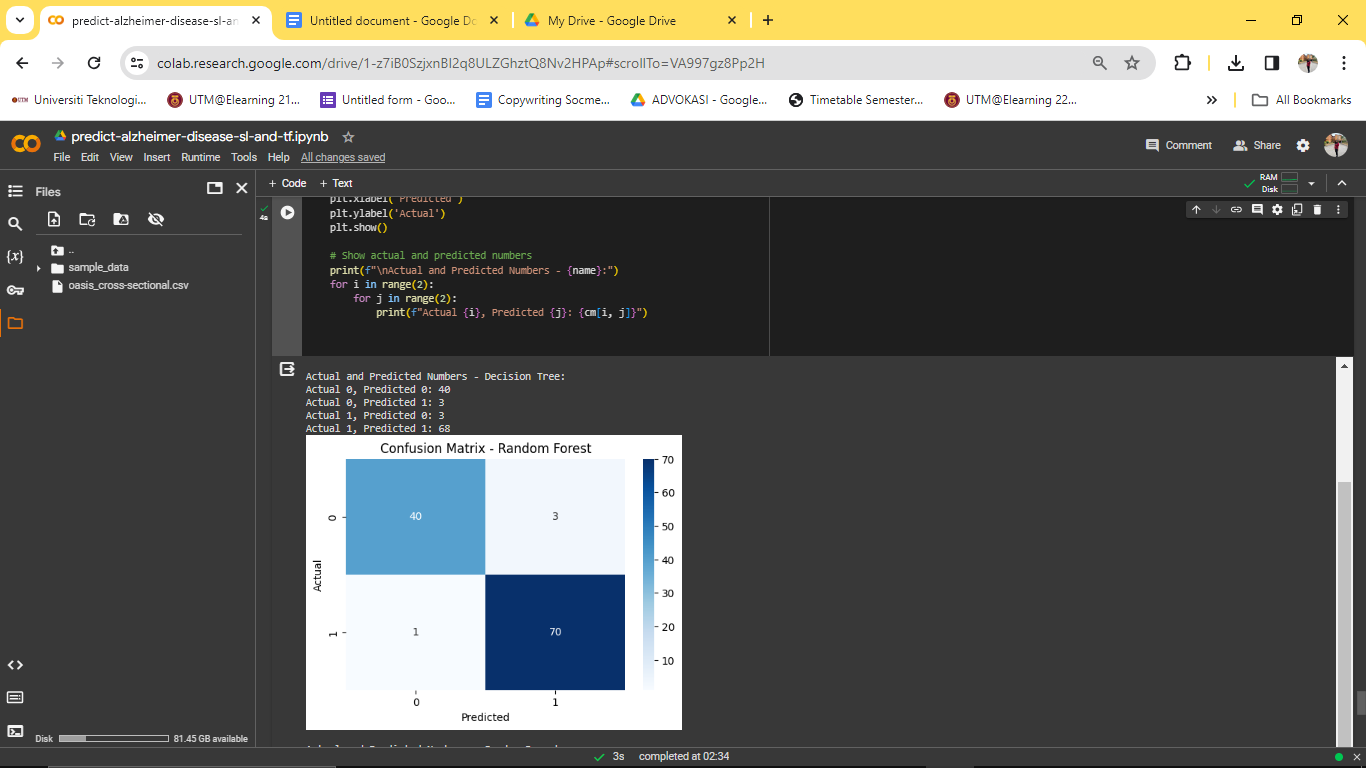
Based on the figure 4.3 the Mean Absolute Error (MAE) of approximately 0.455 indicates that, on average, the model's predictions differ by around 0.455 units from the actual values. The Mean Squared Error (MSE) of approximately 0.495 suggests that the squared average difference between predicted and actual values is 0.495. The R-squared (R2) value of approximately 0.622 indicates that the model explains about 62.21% of the variability in the target variable.

In the context of decision tree regression, an MAE of 0.455 implies that, on average, the model's predictions are within 0.455 units of the actual values. The MSE of 0.495 indicates the average squared difference between predicted and actual values, and the R2 value of 0.622 represents the proportion of variance explained by the model.

1. **Random Forest Regression**



**Figure 4.5**

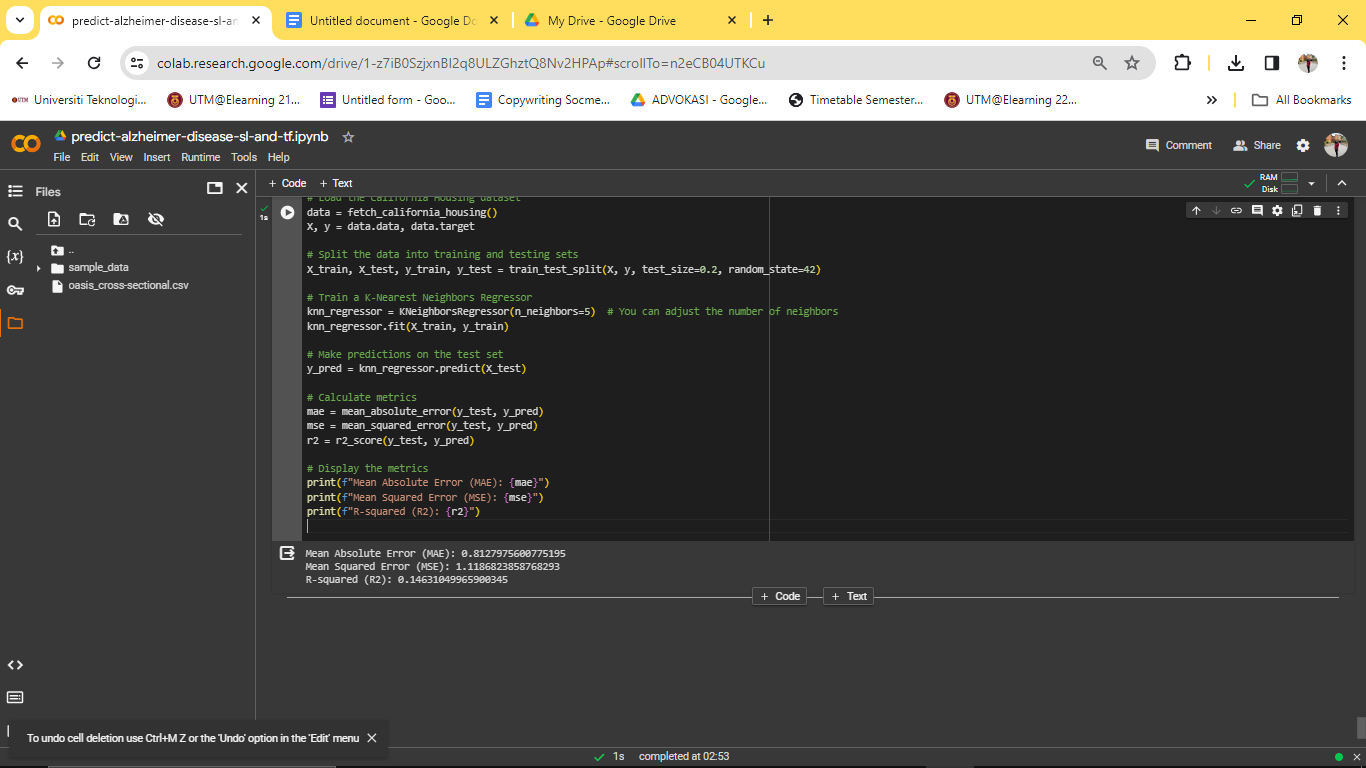


**Figure 4.6**

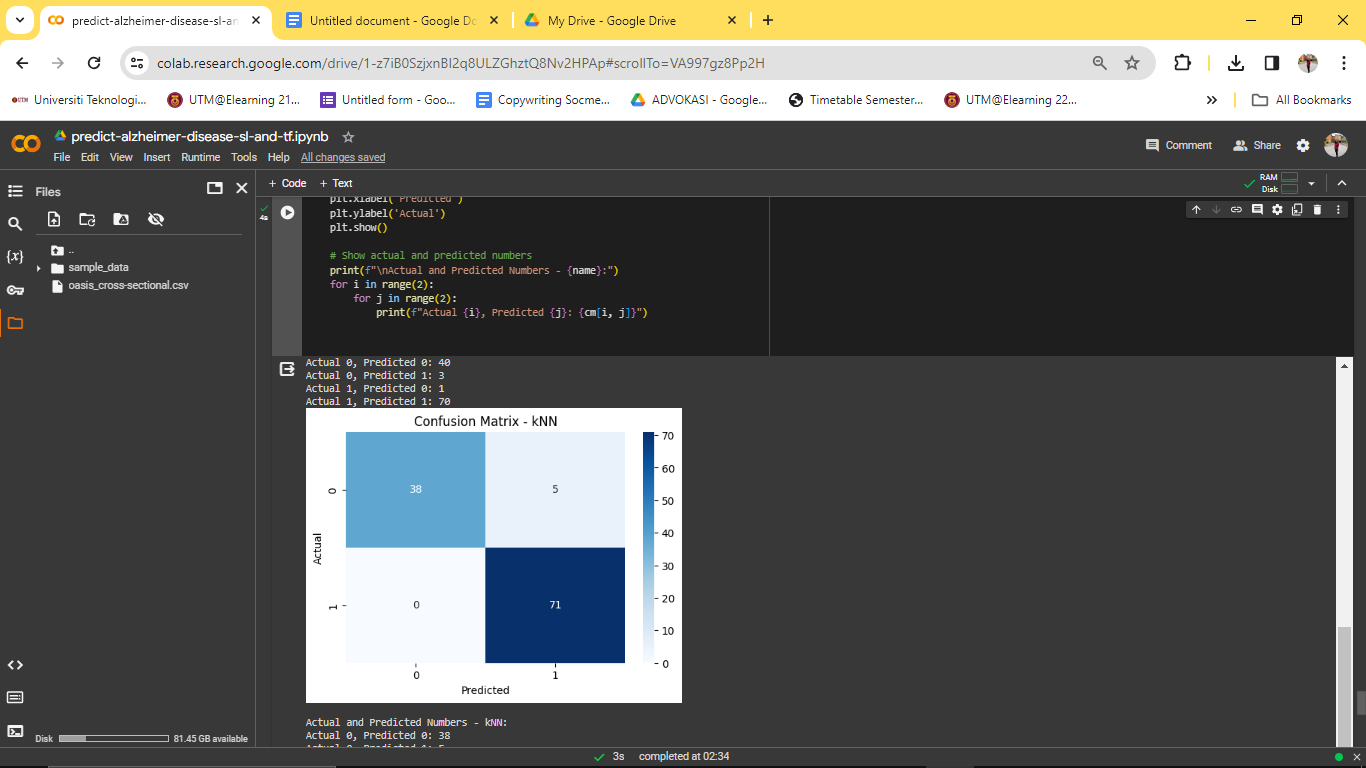
Based on figure 4.5 Random Forest Regression model reveals promising results. The Mean Absolute Error (MAE) of approximately 0.328 suggests that, on average, the model's predictions deviate by only 0.328 units from the actual values. The Mean Squared Error (MSE) of around 0.255 indicates a low squared average difference between predicted and actual values. The R-squared (R2) value of approximately 0.805 is particularly noteworthy, signifying that the model explains approximately 80.51% of the variance in the target variable.

While these metrics offer a quantitative assessment, it is important to consider additional factors for a comprehensive evaluation. Fine-tuning hyperparameters, such as adjusting the number of trees or their depth, could potentially enhance accuracy. Furthermore, understanding the practical significance of accurate predictions in the specific context t could contribute to optimizing the model's effectiveness

1. **KNN**



**Figure 4.7**



**Figure 4.8**

Based on figure 4.7 KNN (K-Nearest Neighbors) model indicates certain observations. The Mean Absolute Error (MAE) of approximately 0.813 suggests that, on average, the model's predictions deviate by around 0.813 units from the actual values. The Mean Squared Error (MSE) of approximately 1.119 signifies a squared average difference between predicted and actual values. The R-squared (R2) value of approximately 0.146 implies that the model explains about 14.63% of the variance in the target variable.While these metrics offer a quantitative perspective, it is essential to delve into additional considerations for a thorough evaluation. Fine-tuning hyperparameters, such as selecting an optimal value for 'k' (number of neighbors), could potentially enhance the model's performance. F

**4.2 Prediction and Decision Making**

In prediction and decision-making scenarios where the models output binary classes (0 or 1), such as in my project on Alzheimer's disease detection using the OASIS dataset, the three models—Random Forest Regression (RFR), Decision Tree Regression (DTR), and K-Nearest Neighbors (KNN)—play distinct roles.

**A) Random Forest Regression (RFR):**

**Description:** Random Forest Regression, an ensemble of decision trees, excels in predicting Alzheimer's disease outcomes by aggregating multiple tree-based predictions. It leverages diverse perspectives to make a final decision, providing robustness and insights into various feature influences on the prediction.

**Summarization:** RFR is a versatile model offering reasonably accurate predictions. It considers interactions among features, providing insights into feature importance and aiding in the understanding of factors contributing to Alzheimer's disease outcomes.

**B) Decision Tree Regression (DTR):**

**Description:** Decision Tree Regression focuses on creating optimal decision boundaries for predicting Alzheimer's disease outcomes. It forms a hierarchical tree structure, making decisions based on feature conditions at each node. It is adept at capturing non-linear relationships in the data.

**Summarization:** DTR delivers interpretable predictions with a clear decision-making structure. It is suitable for scenarios where understanding the sequential steps leading to a prediction is essential.

C) K-Nearest Neighbors (KNN):

**Description:** KNN relies on the similarity of instances to make predictions, classifying based on the majority vote of its neighboring points. It captures local patterns and makes predictions by considering the vicinity of data points.

**Summarization:** KNN, while demonstrating moderate accuracy, offers insights into local patterns within the dataset. It's valuable for recognizing localized trends or nuances in Alzheimer's disease cases, although it may be sensitive to outliers or noise.

**Decision-Making Insights:**

* **Ensemble and Robustness:** RFR provides a collective view from multiple decision trees, o
* ffering a well-rounded prediction and insights into feature interactions.
* **Interpretability and Sequential Decisions:** DTR delivers interpretable predictions with a clear decision-making structure, suitable for understanding the sequential steps leading to a prediction.
* **Localized Patterns:** KNN identifies localized trends within the dataset, offering insights into specific subsets or localized characteristics of Alzheimer's disease cases.

**Final Decision Strategy:**

* **Holistic Insights:** Employ RFR for holistic insights into feature importance and interactions.
* **Interpretability:** Use DTR when interpretability and understanding the sequential decision-making process are crucial.
* **Localized Patterns:** Leverage KNN for recognizing localized patterns or specific subsets within the dataset, aiding nuanced understanding of Alzheimer's disease variations.

**4.3 Summary**

In this part of the project we are tasked with carrying out the model development process through various avenues. For model training we have implemented the method of decision tree regression, random forest regression and KNN. The result from decision tree regression entails that an MAE of 0.455 implies that, on average, the model's predictions are within 0.455 units of the actual values. The MSE of 0.495 indicates the average squared difference between predicted and actual values, and the R2 value of 0.622 represents the proportion of variance explained by the model. Next is the random forest regression which hails the result being the Mean Absolute Error (MAE) of approximately 0.328 suggests that, on average, the model's predictions deviate by only 0.328 units from the actual values. The Mean Squared Error (MSE) of around 0.255 indicates a low squared average difference between predicted and actual values. The R-squared (R2) value of approximately 0.805 is particularly noteworthy, signifying that the model explains approximately 80.51% of the variance in the target variable. We next implement KNN which is able to reveal that the Mean Absolute Error (MAE) of approximately 0.813 suggests that, on average, the model's predictions deviate by around 0.813 units from the actual values. The Mean Squared Error (MSE) of approximately 1.119 signifies a squared average difference between predicted and actual values. The R-squared (R2) value of approximately 0.146 implies that the model explains about 14.63% of the variance in the target variable.