

PROJECT TITLE: NET ASSET VALUE (NAV) FORECASTING

DATASET USED: UTTAMIS DATASET

GROUP NUMBER: 04

GROUP MEMBERS

ASYA H HAJI

GABRIEL D MINZEMALULU

GLORY BARAKA MPONDA

KARIM M. NYUMBA

WATSON JAMES

# ABSTRACT

Investment management demands precision and foresight in predicting the Net Asset Value (NAV) of funds to make informed decisions and manage risk effectively. This project focuses on developing a comprehensive framework for forecasting NAV, catering to a range of investment schemes managed by UttaMIS, including the Umoja Fund, Wekeza Maisha Fund, Watoto Fund, Jikimu Fund, and Liquid Fund.

TABLE OF CONTENTS

Contents

[ABSTRACT i](#_Toc146207501)

[2.2.1 DATA CLEANING AND TRANSFORMATION 4](#_Toc146207502)

[2.2 UNDERSTANDING VARIABLES 6](#_Toc146207503)

[2.3 DATA TRANSFORMATION 6](#_Toc146207504)

[3 FEATURE SELECTION 14](#_Toc146207505)

[3.2 FEATURE LIST 14](#_Toc146207506)

[3.3 FEATURE IMPORTANCE TECHNIQUES 15](#_Toc146207507)

[3.4 FEATURE IMPORTANCE RESULTS d 15](#_Toc146207508)

[3.4 CHOSEN FEATURE SUBSET AND EXCLUDED FEATURES 16](#_Toc146207509)

[4 MODEL DEVELOPMENT 17](#_Toc146207510)

[4.1 FEATURE ENGINEERING 17](#_Toc146207511)

[4.2 DATA SPLITTING 17](#_Toc146207512)

[4.3 MODEL SELECTION 17](#_Toc146207513)

[4.4 MODEL TRAINING AND EVALUTION 18](#_Toc146207514)

[5.RESULTS AND FINDINGS 19](#_Toc146207515)

[6.CONCLUSION 21](#_Toc146207516)

[7. REFERENCES 22](#_Toc146207517)

1. **INTRODUCTION**

**1.1 BACKGROUND**

The Unit Trust of Tanzania (UTT) Asset Management and Investor Services (AMIS) is a key player in the Tanzanian financial landscape, established to promote and develop capital markets by pooling together funds from various investors, both individual and institutional. These funds are then invested in a diversified portfolio comprising equities, bonds, and other financial instruments, with the goal of maximising returns for the investors.

**Net Asset Value (NAV) - The Heartbeat of Mutual Funds:** NAV represents the per-unit value of a mutual fund and serves as a fundamental metric that reflects the health and performance of the fund. For investors, NAV offers a snapshot of the fund's current standing, helping them gauge their investments' growth or decline. For institutions like UTT AMIS, maintaining a consistent and growing NAV is indicative of sound investment strategies and efficient fund management, which in turn solidifies investor trust and attracts further investments.

**1.2 PROBLEM STATEMENT**

In the intricate world of investments, a central challenge confronted by investors, fund managers, and stakeholders of mutual fund schemes is the accurate prediction of their future Net Asset Value (NAV). This forecasting is not merely a numbers game; it is the backbone of informed decision-making, effective portfolio management, and precise risk assessment. For the company managing these funds, understanding the future trajectory of NAVs is pivotal for optimising their offerings and fine-tuning their strategic outlook. Similarly, investors benefit immensely from these forecasts, gaining valuable insights into potential returns, thereby enabling them to adjust their investment strategies accordingly. The intricacies of financial markets, a myriad of influencing factors, and the pressing need for real-time decision-making further accentuate the forecasting challenge. Traditional approaches to NAV prediction, while foundational, often grapple with assimilating dynamic market conditions and nascent trends. The core challenge, then, is to adeptly harness historical data to foresee the NAVs of diverse schemes, ensuring both the company and its investors remain a step ahead in the ever-evolving financial landscape.

**1.3 OBJECTIVE**

The primary objective of this study is to address the forecasting challenges associated with predicting the future Net Asset Value (NAV) of schemes managed by UTT AMIS. Specifically, the study aims to:

**1.3.1 Analyse Historical Data:** Examine the past performance data of various schemes to identify patterns, trends, and anomalies that might influence NAV predictions.

**1.3.2 Evaluate Current Forecasting Techniques**: Assess the efficacy of traditional NAV forecasting methodologies in the Tanzanian context and identify areas of potential improvement.

**1.3.3 Develop Advanced Forecasting Models**: Leverage modern analytical tools, including machine learning and statistical techniques, to enhance the accuracy and reliability of NAV forecasts.

**1.3.4 Understand Influencing Factors**: Investigate both domestic and global factors that might impact the NAV of schemes, thereby providing a holistic perspective for better predictions.

**1.3.5 Provide Recommendations**: Offer actionable insights and recommendations for UTT AMIS, investors, and other stakeholders to make more informed investment decisions based on the forecasted NAV values.

**1.3.6 Enhance Investor Confidence**: By providing a robust and reliable forecasting model, aim to bolster investor confidence in UTT AMIS's schemes, promoting sustained investment and growth.

1. **DATA COLLECTION AND PREPROCESSING**

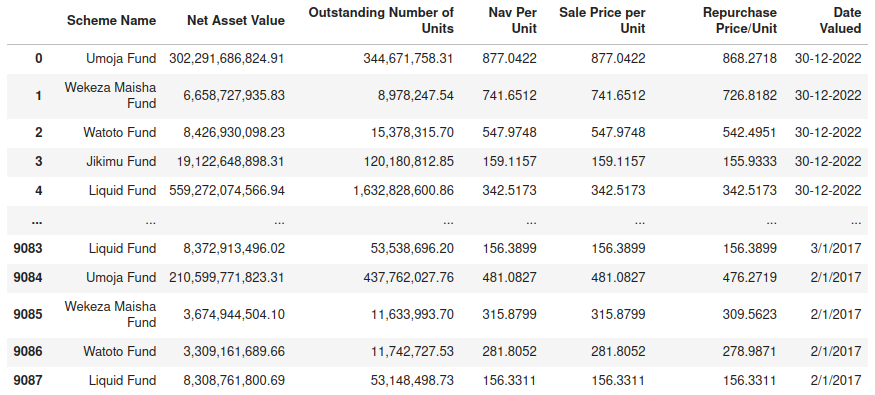
**2.1 DATA COLLECTION**

The dataset for this study was sourced directly from the official UTT AMIS website's ***“Fund Performance”*** section. To automate the data extraction process, a web scraping technique was utilised, leveraging the capabilities of the **Selenium** library in Python.

**Web Scraping Procedure**:

1. **Headless Mode**: Selenium operated in headless mode, ensuring data scraping without the need for visual rendering of web pages.
2. **Threaded Approach**: Data collection was accelerated by employing multiple threads to scrape different data segments concurrently.
3. **Pagination Management**: The scraping script adeptly navigated through multiple pages, ensuring comprehensive data extraction.
4. **CSV Storage**: Post-extraction, the data was promptly saved into segmented CSV files, guaranteeing data integrity.

The entire scraping process was encapsulated in the ***scrape\_pages*** function, ensuring modularity and reusability.



2.2 DATA PREPROCESSING

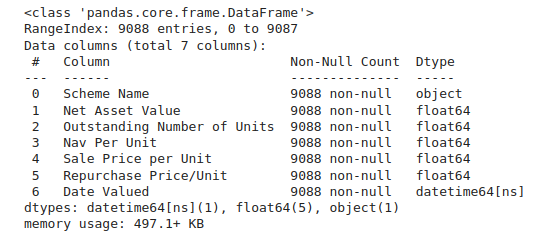
With the data in hand, preprocessing steps were undertaken to transform it into a suitable format for subsequent analysis.

## 2.2.1 DATA CLEANING AND TRANSFORMATION

Data cleaning refers to the process of identifying and correcting or removing inaccurate, incomplete, or irrelevant data from a dataset. The goal of data cleaning is to ensure that the data is accurate, complete, and consistent so that it can be used effectively for analysis, modelling, or visualization. During data cleaning, several steps were taken as follows:

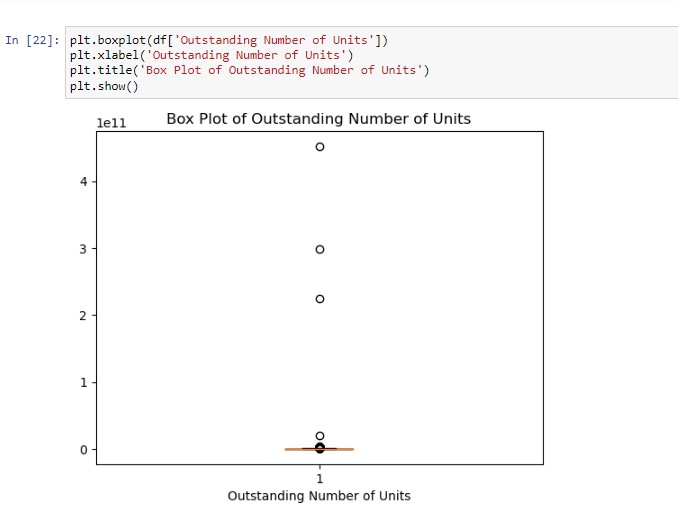
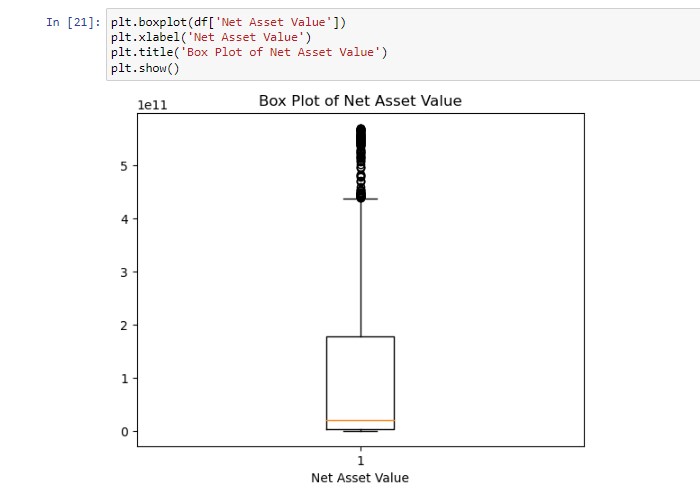
1. Dataset Structure: The dataset consists of 9088 rows and 7 columns, each providing insights into different financial schemes.
2. Data Types: The dataset has a mix of object (string) and float data types. Some columns, like "Net Asset Value" and "Outstanding Number of Units", are in string format due to the presence of comma separators.
3. Duplicates: There are 921 duplicate rows in the dataset. Depending on the context and nature of the data, we may need to decide whether to retain or remove these duplicate entries.
4. Missing Values: There are no missing or null values in any of the columns, as confirmed by our analysis and visualization.
5. Features: The dataset provides various metrics, such as net asset value, sale price per unit, and repurchase price per unit, for different schemes. The "Date Valued" column indicates the date of the provided data, which may be useful for time series analysis.

After preprocessing, a thorough inspection was conducted using the ***df.info()*** function to ascertain the data types and recognize any missing or inconsistent values.



Handling outliers

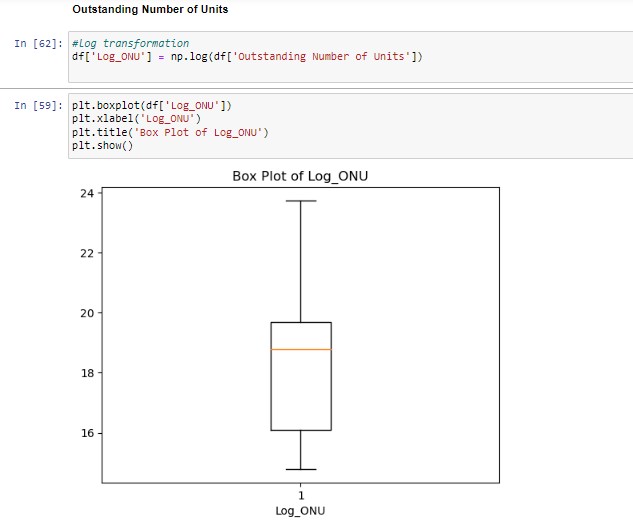
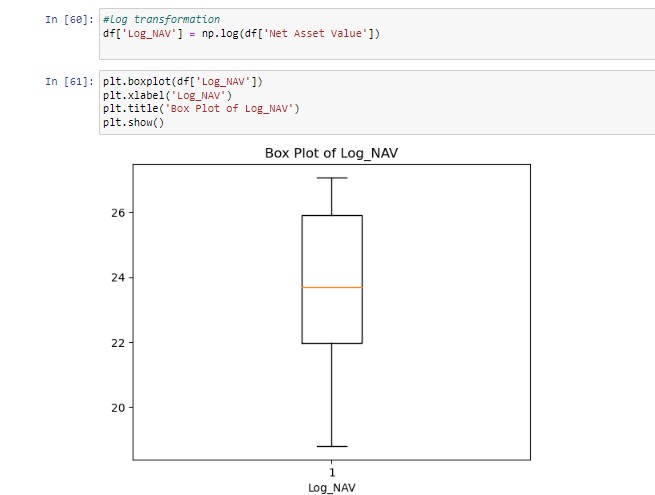
Detecting outliers by using box plot



*Figure 2: Box plots showing presence of outliers*

As shown in the boxplots above, outliers are detected in net asset value and Outstanding number of units.

Removing outliers



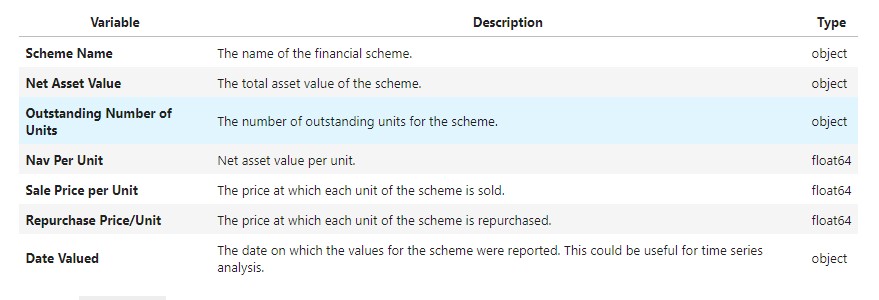
*Figure 3: Box plots after removing outliers*

We used Log transformation method to handle outliers. Log transformation involves taking the logarithm of each data point. This has the effect of compressing large values and expanding small values. This can be helpful for reducing the influence of outliers, which are often very large or very small values. The figures below show the box plots of net asset value and outstanding number of units after handling outliers.

## 2.2 UNDERSTANDING VARIABLES

2.2.1 Variables Description

The dataset consists of the following variables:

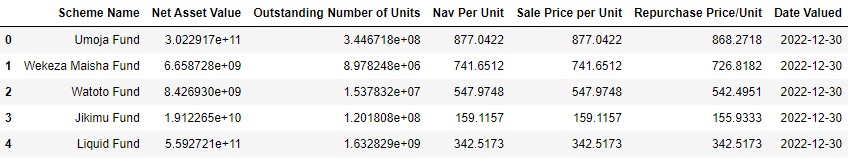


*Figure 4: Figure showing the description of variables*

The figure shows different variables and their data types. In our dataset we have variables which are either objects or floats

## 2.3 DATA TRANSFORMATION

Data transformation is a fundamental step in data preprocessing and analysis. It involves converting data from its original format or structure into a different format that is more suitable for a specific analysis, modelling, or visualization task. Data transformation is essential for improving data quality, making data compatible with analytical tools, and uncovering hidden patterns or relationships in the data.



*Figure 5: Figure displaying data after transformation*

Based on our dataset we observed that the "Date Valued" column should be converted to the datetime data type to enable time-series analysis and date-based filtering. Also, columns like Net Asset Value and outstanding number of units are typically represented as numbers but are currently of type "object." Therefore, they should be converted to appropriate numeric data types like float

The figure above shows first few rows of dataset after being transformed.

2. EXPLORATORY DATA ANALYSIS

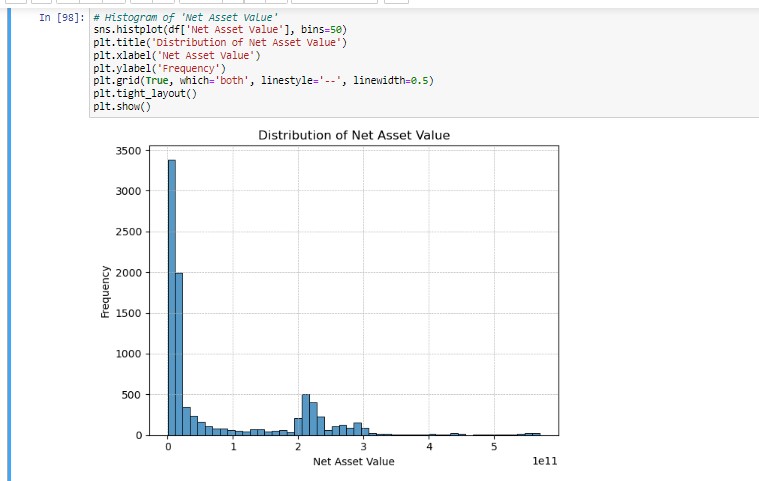
# 

Exploratory Data Analysis (EDA), also known as Data Exploration, is a step in the Data Analysis Process, where a number of techniques are used to better understand the dataset being used. It helps you gather insights and make better sense of the data, and removes irregularities and unnecessary values from data. In this project we will perform Exploratory data analysis in four steps Which are Data Collection, Data cleaning, Data visualization and Gathering insights.

2.4 DATA VISUALIZATION

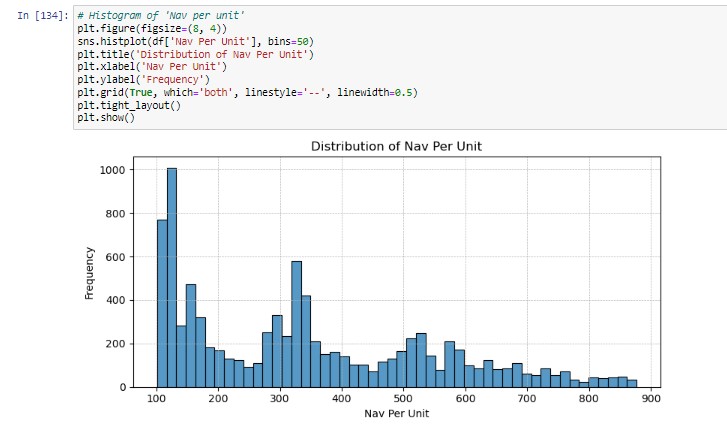
Data visualization in data science refers to the process of generating graphical representations of information often known as plots or charts which can be used for effective analysis and interpretation.

Distribution of Net asset value



*Figure 6: figure showing the distribution of net asset value*

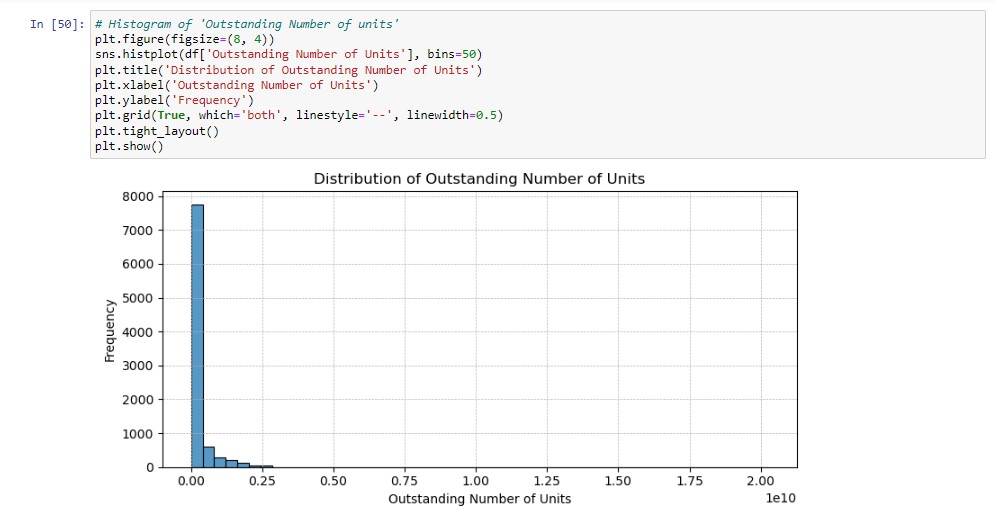
* The figure shows that, the distribution of Net asset value is mostly right-skewed, indicating that most of the schemes have their values in the lower range.
* While there are schemes with higher values, they are less frequent in the dataset.

Distribution of Nav per unit

*Figure 7: Figure showing the distribution of nav per unit*

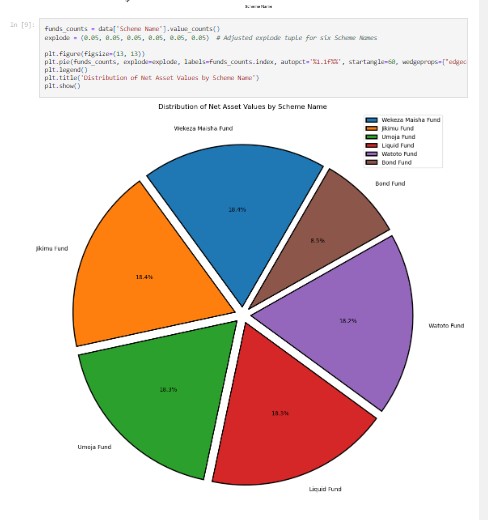
* The figure above shows that, the distribution of Nav per unit is ranges mostly up to 300 to 400, and a significant peak is observed around the value of 1000, such distribution might indicate that while most schemes maintain a standard values per unit, a few schemes might have specialized or premium offerings or pricing.

Distribution of outstanding Number of units



*Figure 8: Figure showing the distribution of Outstanding number of units*

* The figure shows that, the distribution of Outstanding number of units is also right-skewed, indicating that most of the schemes have their values in the lower range.
* A significant peak is observed around the value of 1000, suggesting that a majority of the schemes have their values close to this mark

Distribution of Net asset value by scheme name

*Figure 9: Figure showing the distribution of Net asset value by scheme name*

From the figure above, you can easily compare the total NAV across different schemes. This can help you identify which schemes have the highest or lowest total NAV, providing insights into the relative size or popularity of each scheme. You can also see the relative contribution of each scheme to the total NAV.

Trends over time

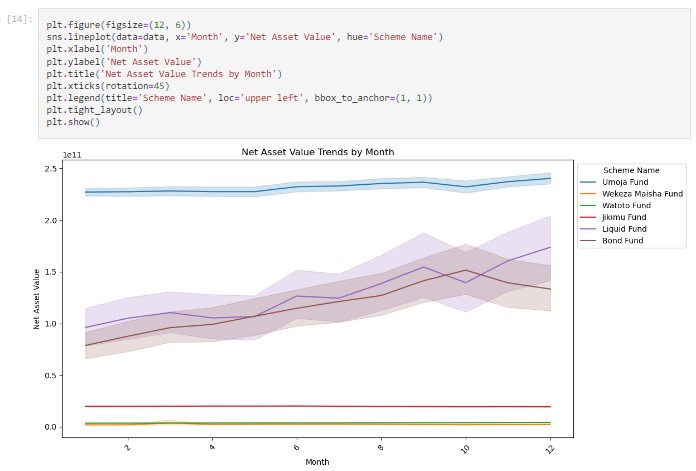
This can be used to show how data changes over time. It can be used to identify patterns and trends in data, to make predictions, and to communicate information about data in a clear and concise way.

 Time series plot of Net asset value

*Figure 10: Time series plot of Net asset value*

* From the figure above, The NAV of different funds has shown a general upward trend over time, indicating overall growth.
* There may be seasonality or cyclical patterns in the NAV data, which can be identified through time series analysis.
* Certain funds have higher volatility (fluctuations) in their NAV compared to others, indicating different risk levels

Net asset value trends by month



*Figure 11: Figure showing the Net asset value trends by month*

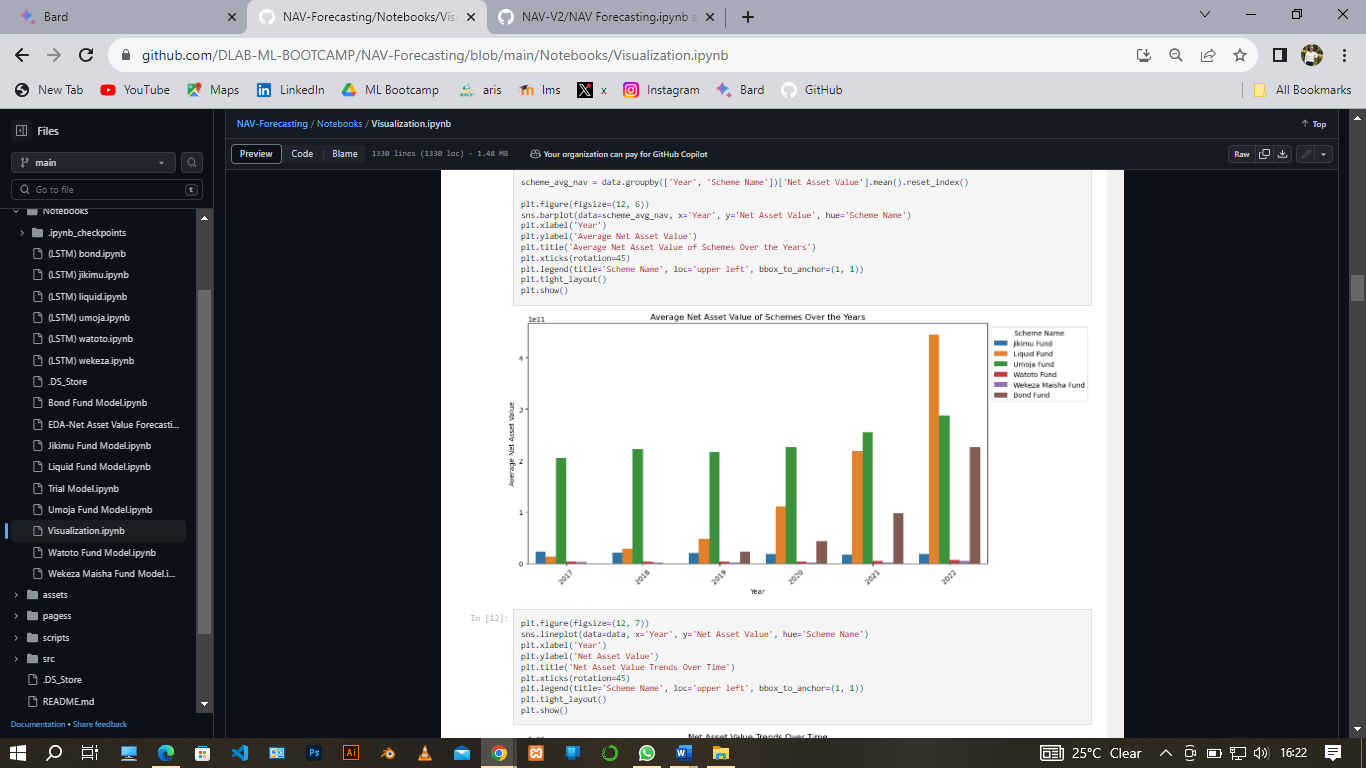
From the figure above:

* Analysing monthly data allows you to detect seasonal patterns. For example, if there are certain months or seasons where the average NAV tends to be higher or lower this could be due to market or economic factors.
* Examine monthly fluctuations to identify any schemes that are more sensitive to short-term market changes. Some schemes may exhibit higher volatility or sensitivity to specific months.
* Look for patterns of consistency or variability in scheme performance. Example, if there are months where most schemes perform similarly, and other months where there is significant divergence in performance.
* Consider how investor behaviour or market events might influence monthly trends.

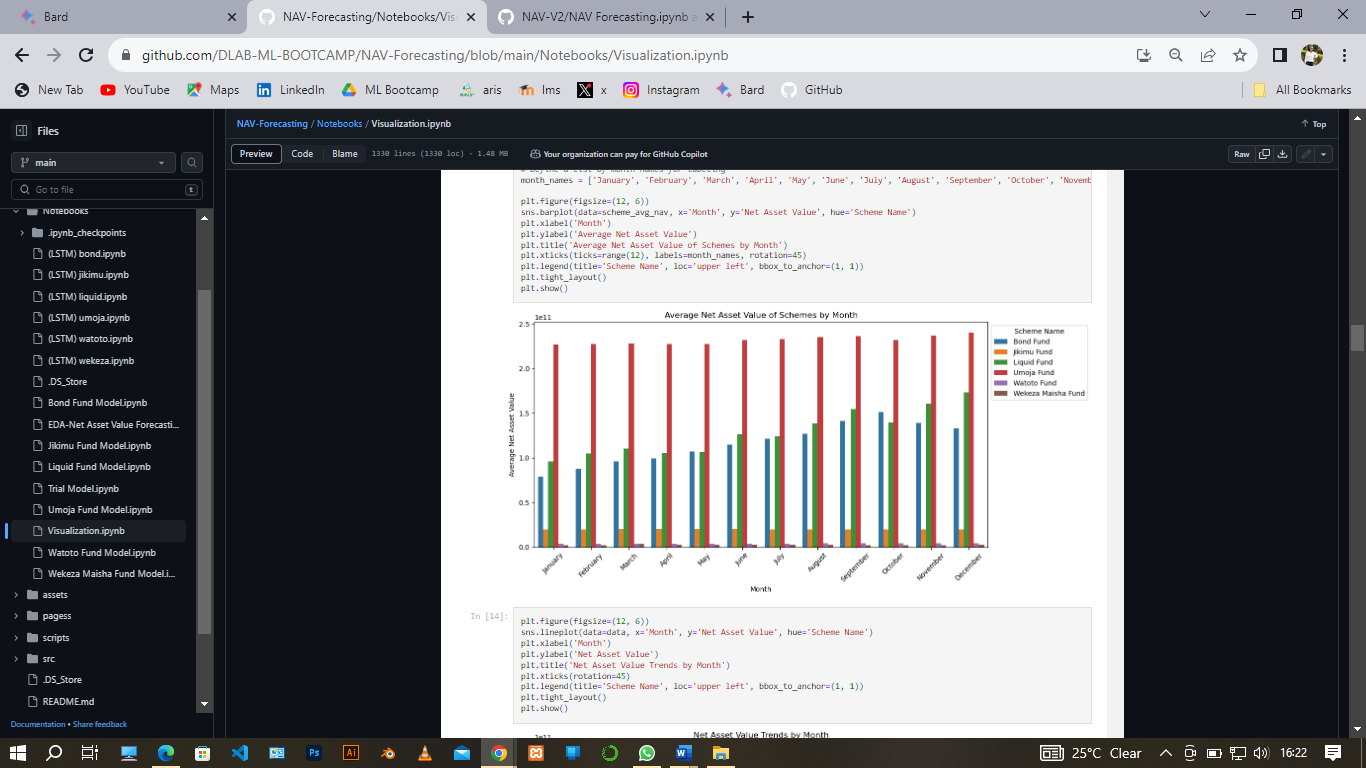
For example, tax seasons, holidays, or significant economic events could impact NAV on a monthly basis.

* Analyse whether schemes experience changes in net asset value due to redemption or purchase behaviours by investors during specific months.

Average net asset value of schemes over the years

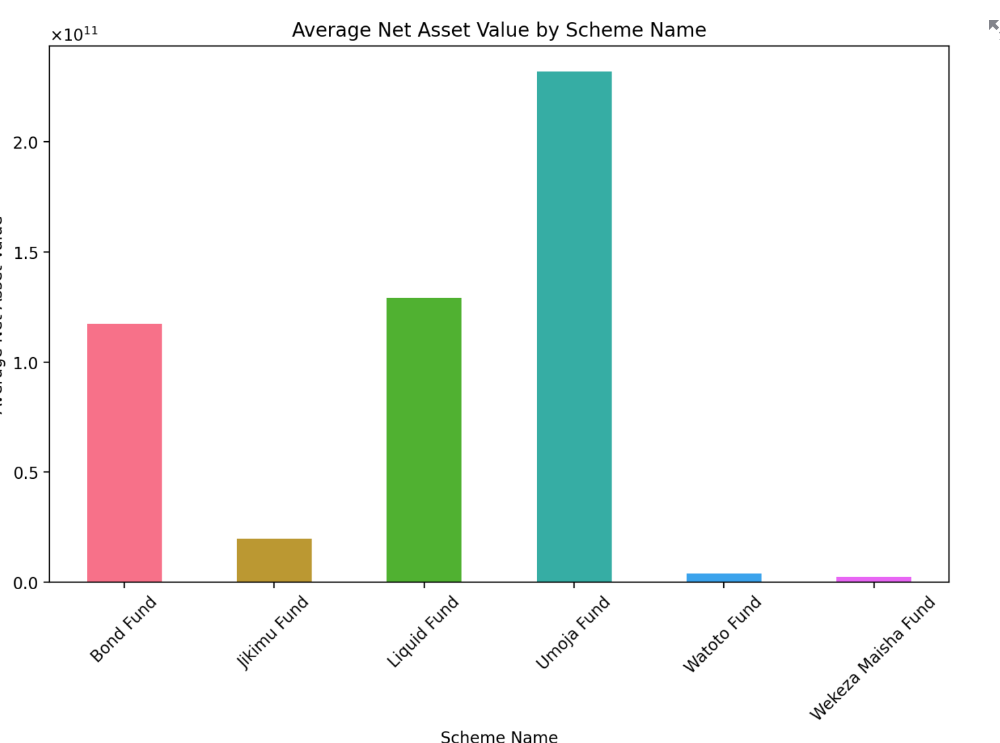


The figure above shows the average of net asset value of each scheme over the years from 2017 to 2022 which shows how Net asset value of each scheme varies in each year.

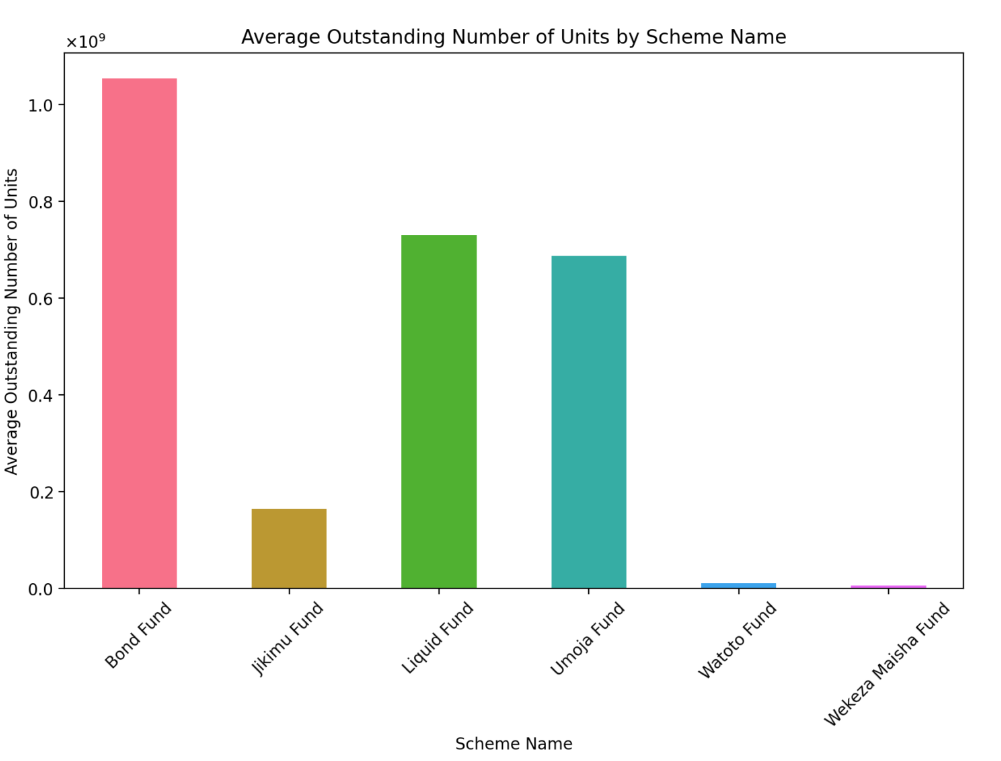
Average net asset value of schemes by month

The figure above shows distribution of net asset value of each scheme over months

Average net asset value by scheme name



The figure above shows the distribution of average net asset value of each scheme

Average net asset value by scheme name

The figure above shows the distribution of average outstanding number of units of each scheme

# 3 FEATURE SELECTION

Crucial step for explaining how you chose the most relevant variables (features) for your analysis or modelling task. It helps improve the efficiency and effectiveness of your analysis by focusing on the most informative attributes. Here are some tasks accompanied with feature selection

3.1 OBJECTIVE AND JUSTIFICATION

The primary objective of the feature selection phase is to identify and choose the most relevant features that significantly influence the Net Asset Value (NAV) forecasting model. By selecting a subset of features, we aim to improve the model's accuracy, reduce overfitting, and enhance its interpretability. Feature selection is a crucial step in ensuring that the model focuses on the most informative attributes while avoiding noise and redundancy.

## 3.2 FEATURE LIST

Before proceeding with feature selection, let's review the list of features available in our dataset after performing feature engineering:

* Outstanding number of units /Log\_ONU
* Repurchase price / Unit
* Nav Per Unit
* Scheme\_Liquid Fund
* Scheme\_Jikimu Fund
* Scheme\_Umoja Fund
* Scheme\_Wekaza Maisha Fund
* Scheme\_Watoto Fund
* Scheme\_Bond Fund
* Sale Price per unit
* Repurchase Price/Unit (Numeric)
* Date Valued

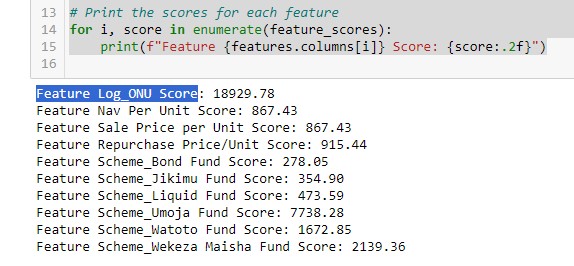
## 3.3 FEATURE IMPORTANCE TECHNIQUES

In this analysis, we employ several techniques to assess the importance of each feature:

• Correlation Analysis: We calculate the correlation between each numeric feature (Net Asset Value, Outstanding Number of Units, Nav Per Unit, Sale Price per Unit, Repurchase

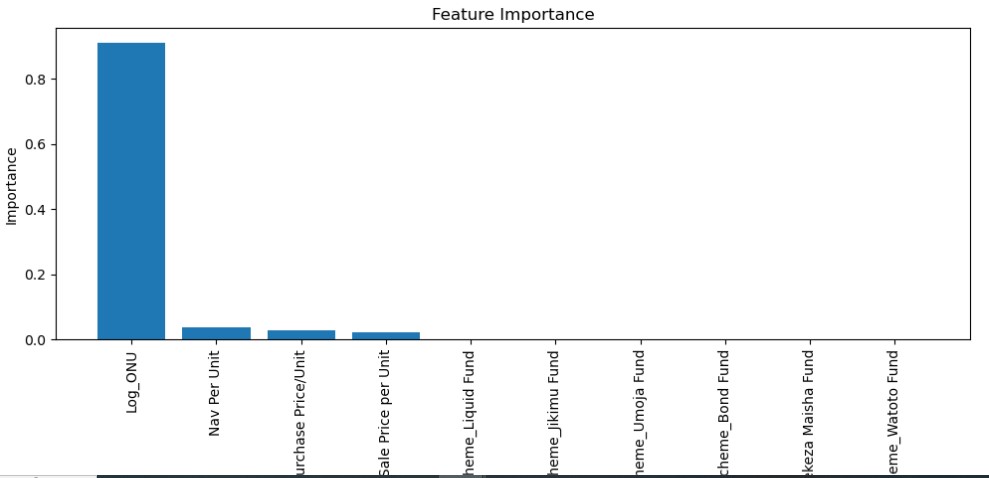
Price/Unit) and the target variable (Net Asset Value). This helps identify linear relationships. • Feature importance score: We utilize tree-based models, such as Random Forest, to compute feature importance scores. These scores reveal the contribution of each feature to model performance.

## 3.4 FEATURE IMPORTANCE RESULTS d

The feature importance analysis yielded the following results: the below shows the output we noticed we analysing the features importance

*Figure 12:figure displaying features with their corresponding scores*

Where also we deploy bar plot to display the distribution of features with their corresponding score. Below is the bar plot



*Figure 13: figure displaying bar plot graph of feature importance score*

From the figure above feature Log\_ONU which was transformed from Outstanding Number of Units appear to have highest score compared to other hence it is the highest feature importance.

## 3.4 CHOSEN FEATURE SUBSET AND EXCLUDED FEATURES

Based on the feature importance analysis features Log\_ONU which was transformed from Outstanding Number of Units was selected for inclusion in the NAV forecasting model as it appeared with highest score. The inclusion of these features is expected to provide the model with a comprehensive set of attributes for accurate NAV forecasting. Where the rest features were excluded

# 4 MODEL DEVELOPMENT

The model development phase is a critical component of our project, where we create and fine-tune a machine learning model to forecast the Net Asset Value (NAV) of investment funds managed by UttaMIS. This phase builds upon the feature selection process and leverages the selected features to construct a robust predictive model. The objective is to provide accurate and reliable NAV forecasts for the Umoja Fund, Wekeza Maisha Fund, Watoto Fund, Jikimu Fund, and Liquid Fund, enabling informed investment decisions and risk management.

## 4.1 FEATURE ENGINEERING

Feature engineering in Machine Learning involves extracting useful features from given input data following the target to be learned and the machine learning model used. In feature engineering, Additional features such as market indices, economic indicators, and temporal information (Year, Month, Day, Weekday) were engineered to capture the complex relationships within the data. Logarithmic transformation of the "ONU" feature was applied to normalize its distribution.

## 4.2 DATA SPLITTING

Data splitting is another important aspect we explored, particularly for creating models based on data. This technique helps ensure the creation of data models and processes that use data models are accurate. The dataset was split into training and testing sets

(80% training and 20% testing) using the `train-test-split` function.

## 4.3 MODEL SELECTION

Several models were explored to find the best fit for the NAV forecasting task. These models included:

* LSTM (Long Short-Term Memory): is a deep learning model designed for sequence prediction tasks. It was considered for its ability to capture complex temporal dependencies in the data, making it a suitable candidate for NAV forecasting
* ARIMA (Autoregressive Integrated Moving Average): ARIMA, a traditional time series model, served as a baseline for comparison. It was chosen to assess whether simpler models can provide competitive performance.
* SARIMA (Seasonal ARIMA): SARIMA was applied to account for seasonality in the NAV data, acknowledging that financial time series often exhibit seasonal patterns.
* Ensemble Model (Lasso Regression + Ridge Regression): An ensemble model combining Lasso and Ridge Regression was selected. This ensemble model leverages the strengths of both regression techniques, including feature selection and regularization capabilities.

## 4.4 MODEL TRAINING AND EVALUTION

4.4.1 Ensemble model: Lasso Regression and Ridge Regression models were trained with the best hyperparameters determined through experimentation (Lasso Alpha = 0.01, Ridge Alpha = 0.1). The ensemble model, combining these two models using the

Voting Regressor, was trained on the training dataset.

4.4.2. Model Evaluation: The performance of the models was rigorously evaluated using the following key metrics:

* Mean Absolute Error (MAE): MAE measures the average absolute prediction error and provides insights into the model's accuracy.
* Mean Squared Error (MSE): MSE quantifies the average squared prediction errors, emphasizing larger errors.
* Root Mean Squared Error (RMSE): RMSE is the square root of MSE, providing a more interpretable error metric.
* R-squared (R²): R² measures the proportion of variance in the NAV explained by the model, indicating goodness of fit.
* Mean Absolute Percentage Error (MAPE): MAPE assesses the percentage error in predicting NAV values, providing insights into relative prediction accuracy.

Sharpe Ratio and Annualized Returns: Daily returns were computed from the predicted NAV values, and annualized returns were calculated to gauge investment performance. The Sharpe ratio was employed to measure risk-adjusted returns, a crucial metric for investors

# 5.RESULTS AND FINDINGS

After extensive experimentation and evaluation, the ensemble model combining Lasso Regression and Ridge Regression exhibited superior performance:

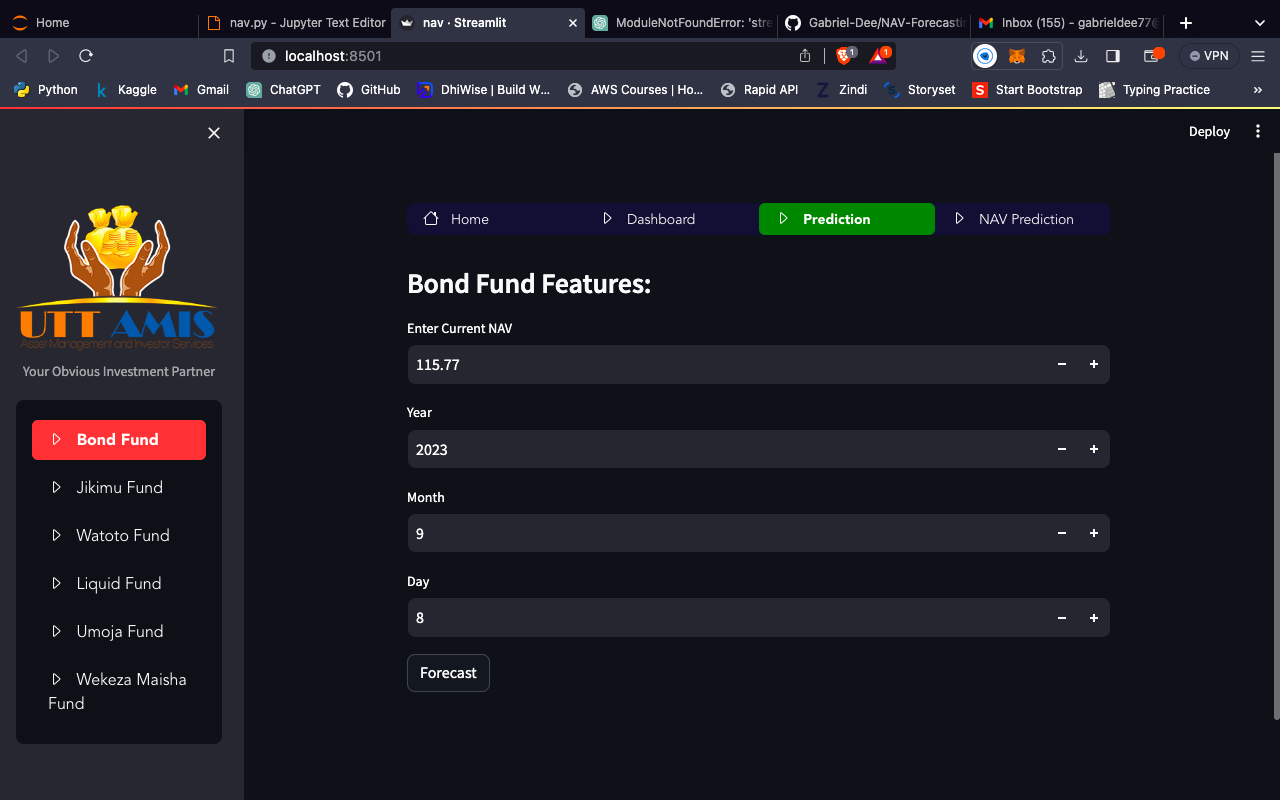
* It outperformed other models in terms of forecasting accuracy, as evidenced by lower MAE, MSE, and RMSE.
* The model achieved higher R² values, indicating a better fit to the data.
* Its robustness in handling feature selection and regularization made it the preferred choice.
* Here are some figures displaying the findings obtained from nav forecast predictive model.

Figure 14 Figure showing nav predictions of bond fund:

The above figure shows a sample of an when NAV of 115.77 invested in Bond fund schema after a given duration of time was forecasted with accuracy value of 99.11% are the predicted NAV per unit becomes 117.54. the result shown on the next figure

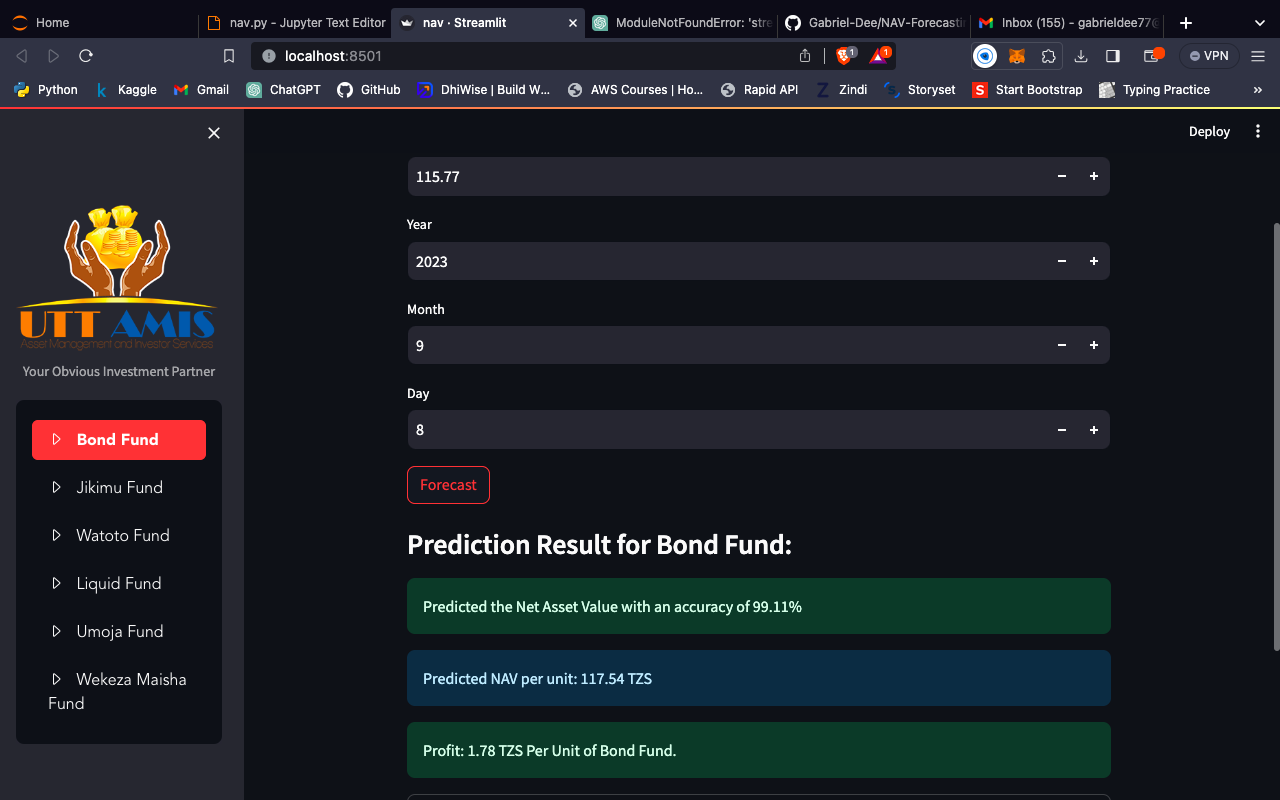


Figure 15: Figure showing the results of nav predictions for bond fund

Now the above figure shows the predictive result as stated earlier. Through these findings we obtain from our model to accurately forecasting some values hence main part of project goal was successfully achieved

# 6.CONCLUSION

The model development phase for forecasting Net Asset Values of UTT AMIS schemes involved a systematic approach to data preparation, model selection, and evaluation. The ensemble model combining Lasso and Ridge Regression proved to be a reliable and accurate solution for NAV forecasting across all six schemes. This approach ensures that UTT AMIS can make informed investment decisions, mitigate financial risk, and maximize returns for its investors. Continued monitoring and periodic retraining of the model will be essential to adapt to evolving market dynamics and ensure the accuracy and relevance of predictions. This model development effort exemplifies the significance of data-driven decision-making in the financial industry, benefiting both fund managers and investors.

# 7. REFERENCES

https://github.com/DLAB-ML-BOOTCAMP/NAV-Forecasting https://www.uttamis.co.tz/fund-performance