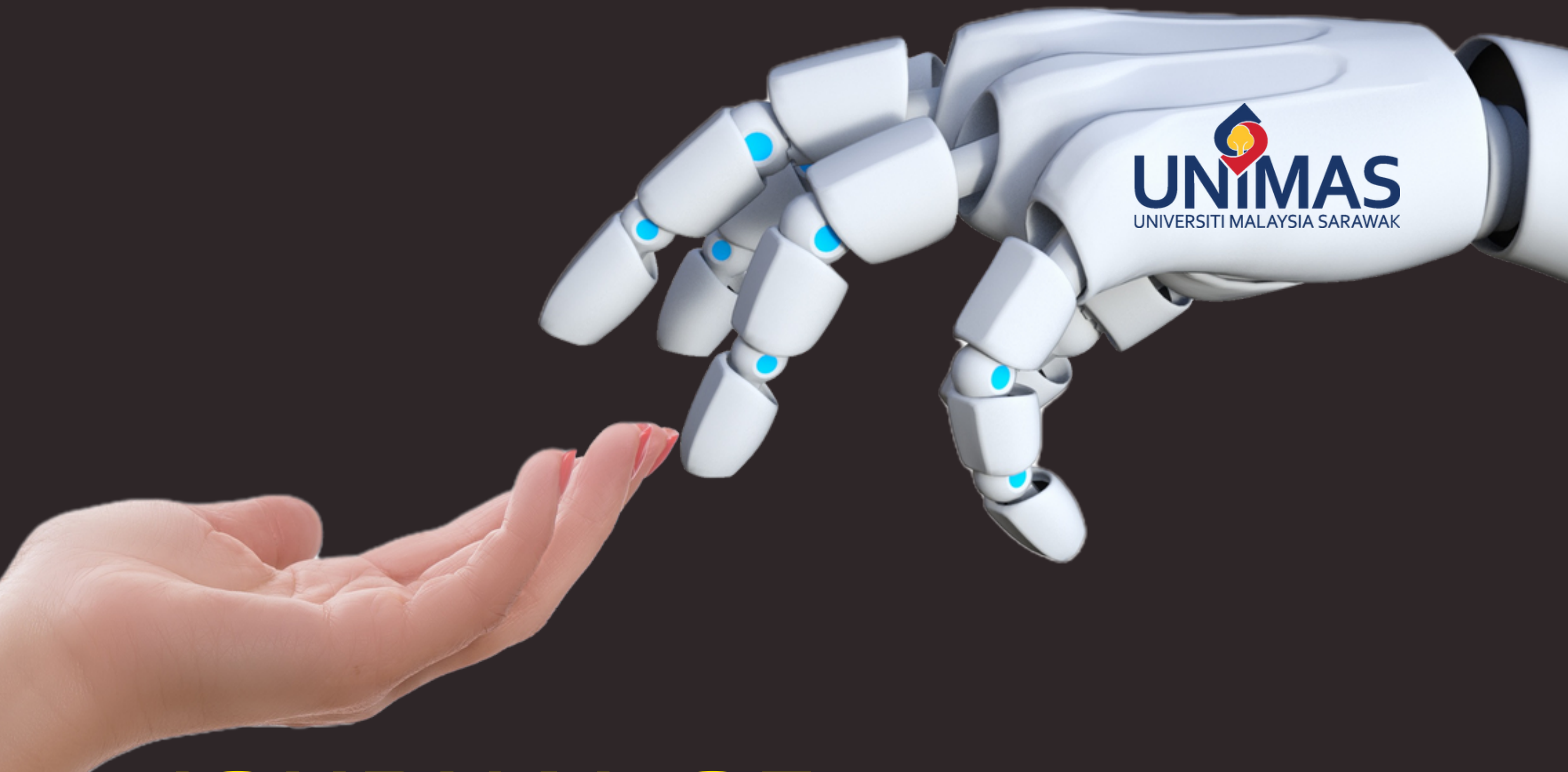


UNIMAS
UNIVERSITI MALAYSIA SARAWAK

JOURNAL OF COMPUTING AND SOCIAL INFORMATICS

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITI MALAYSIA SARAWAK

eISSN 2821-3777
9772821377005

Editorial Committee

Advisory Board

Reviewers

Journal of Computing and Social Informatics

The Journal of Computing and Social Informatics (JCSI) is an international peer-reviewed publication that focuses on the emerging areas of Computer Science and the overarching impact of technologies on all aspects of our life at societal level. This journal serves as a platform to promote the exchange of ideas with researchers around the world.

Articles can be submitted via *www.jcsi.unimas.my*

Assoc Prof Dr Chiew Kang Leng

Chief Editor
Journal of Computing and Social Informatics
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan
Sarawak, Malaysia

Contents

Generation Z and the New Economic Reality: A Machine Learning Perspective on Financial Challenges

[1*]Abdullah Aljishi, [2]Matin Marjani, [3]Arash Latifi and [4]Lior Shamir

[1, 2, 3]Department of Electrical and Computer Engineering, Kansas State University, Kansas, United States
[4]Department of Computer Science, Kansas State University, Kansas, United States

email: [1*]aljishi@ksu.edu, [2]matinmarjani@ksu.edu, [3]arashlatifi@ksu.edu, [4]lshamir@ksu.edu

**Corresponding author*

Received: 07 February 2025 | Accepted: 25 April 2025 | Early access: 01 May 2025

Abstract - *This study explores the socioeconomic disparities and financial challenges faced by different generational cohorts, with a focus on Generation Z. The research aims to identify patterns in socioeconomic features, such as income distribution and housing affordability, that distinguish generations and impact their financial outcomes. Machine learning models were used, with classification models that predicted generational membership and regression models that estimated the year of birth as a continuous variable. Using mutual information for feature selection, the Explainable Boosting Machine (EBM) achieved the highest classification accuracy of 74.78%, as evaluated using 10-fold cross-validation, while regression analysis demonstrated moderate predictive power (R^2 = 0.6005) with an average absolute error of eight years. The results highlight significant generational differences, with Generation Z experiencing the highest median rent-to-income burden (60.0%) and substantial barriers to homeownership. Despite higher participation in the workforce compared to previous generations at similar life stages, systemic economic challenges, such as rising housing costs and stagnant wages, disproportionately affect Generation Z. These findings underscore the utility of machine learning in identifying generational trends and socioeconomic disparities, offering a framework for further research to refine models and explore additional socioeconomic variables to enhance understanding of generational dynamics. Code and data to reproduce the results are available in GitHub, as detailed in the Dataset Overview subsection.*

Keywords: Generation Z, machine learning, socioeconomic disparities, financial challenges, income inequality.

1 Introduction

For decades, the idea of generational progress has been a cornerstone of societal development: each generation has historically achieved better financial outcomes, improved living standards, and greater opportunities than the one before (Chetty et al., 2017). From the Silent Generation, who endured the hardships of the Great Depression and World War II but benefited from post-war economic growth (Elder, 2018), to the Baby Boomers, who enjoyed access to affordable housing, stable jobs, and rising wages (Patterson, 1996), this upward trajectory seemed inevitable. However, recent discourse suggests that Gen Z may be the first generation to diverge from this trend, facing significant challenges in achieving financial stability and upward mobility relative to their parents and predecessors (Gregory, 2023; Lev, 2021).

Understanding these challenges requires examining the distinct historical, cultural, and socioeconomic contexts that have shaped each generation. Generational divisions, categorized by birth years and corresponding age ranges, are illustrated in Figure 1 (Ipsos, 2023). While Baby Boomers benefited from the economic prosperity of the post-war era, Generation X saw the rise of dual-income households (Bianchi, 2000), increased access to higher education (Bound & Turner, 2007), and advancements in technology that began to reshape industries (Autor et al., 1998). Millennials grew up during the rapid expansion of the internet and digital technologies, which created new opportunities for innovation and connectivity (Palfrey & Gasser, 2011). However, as economic landscapes have evolved, Generation Z is now widely believed to be grappling with skyrocketing housing costs, stagnant wages, mounting student debt, and a job market increasingly dominated by gig work and automation. These economic shifts are thought to have compounded the financial difficulties of Gen Z, potentially leaving them with fewer opportunities for upward mobility compared to prior generations (Twenge, 2023). This paper seeks to investigate whether these claims hold true by examining economic data and generational trends.

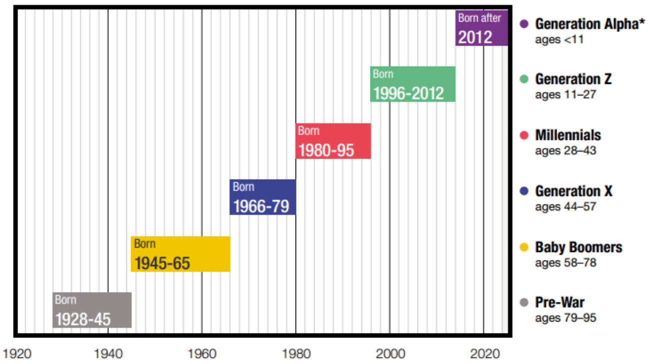


Figure 1: Generational categories by birth year and age in 2023. Adapted from Ipsos (2023)

Previous generations often benefited from pathways to financial security, such as affordable housing and stable income growth. In contrast, Generation Z appears to face significant barriers to wealth-building, particularly through homeownership, which has historically been a key driver of economic mobility (Harvard Joint Center for Housing Studies, 2006). Systemic factors, including growing income inequality, reduced union representation, and living costs consistently outpacing wage growth, are thought to exacerbate these challenges (Mishel & Bivens, 2021; Western & Rosenfeld, 2011).

The issue of income inequality and its evolution across generations has been the subject of extensive research. Prior studies have explored the economic disparities between generational cohorts, such as Baby Boomers, Generation X, Millennials, and Generation Z, focusing on factors like wage stagnation, inflation, and wealth accumulation. For instance, Charles and Hurts (2003) studied how wealth is passed between generations, finding that children's wealth is strongly influenced by their parents' wealth, with lifetime income and asset ownership playing the biggest roles. Gallipoli et al. (2020) examined the joint evolution of cross-sectional inequality in income and consumption across generations.

Research on generational economics has also highlighted the influence of macroeconomic factors, such as recessions and housing market trends, on the financial well-being of different generations. Studies such as Green and Lee (2016) have provided insights into the demand for housing based on age and demographics. However, many existing analyses rely on aggregated data or limited timeframes, which fail to capture the granular differences across individual-level datasets.

In the field of predictive modelling, recent advancements have been made in using machine learning to analyze socioeconomic patterns. For example, Fan et al. (2023) demonstrated the utility of classification and regression models in predicting socioeconomic outcomes based on complex interactions of urban features. While these approaches have proven effective, their application to understanding generational disparities in income remains underexplored.

This study aims to build on this body of work by investigating key factors such as income distribution, housing affordability, and inflation-adjusted wages. By identifying trends and disparities, this research seeks to determine whether Gen Z represents a significant deviation from historical patterns and to understand the systemic changes required to address these challenges.

The remainder of this paper is organized as follows: Section 2 details the dataset and preprocessing procedures, while Section 3 outlines the methodology employed in this study. Section 4 describes the economic metrics used to evaluate generational economic conditions and disparities. Section 5 presents and discusses the findings, and finally, Section 6 concludes the paper and offers recommendations for future research.

2 Data Source and Acquisition

Data for this study was obtained from the IPUMS USA database (Ruggles et al., 2024), a comprehensive resource providing integrated microdata for social and economic research. The dataset comprises anonymized individual-level records with variables covering a range of topics, including demographics, housing characteristics, and employment status. The data span multiple years, enabling longitudinal analysis of generational trends.

2.1 Dataset Overview

The raw dataset obtained from IPUMS spans a comprehensive temporal range, covering the years from 1970 through 2023, which includes individual survey data from the years 1970, 1980, 1990, 2000, and annually from 2001 to 2023. This dataset encapsulates various individual and household attributes across approximately 100,000 respondents, representing a diverse cross-section of the U.S. population.

Key features in the dataset include `YEAR` (year of the survey), `BIRTHYR` (birth year), `RENT` (monthly rent), `VALUEH` (home value), `INCTOT` (total personal income), and categorical variables such as `STATEFIP` (state), `RACE` (race), and `EDUC` (educational attainment). The extensive span and substantial size of the dataset necessitated the implementation of efficient data management techniques, particularly crucial due to the presence of both continuous and categorical variables, as well as coded missing values across several features.

All the code and related files used in the preprocessing pipeline are available in the project's GitHub repository (Aljishi et al., 2025). The repository includes all scripts and resources necessary to replicate the data processing steps, including data splitting, feature engineering, handling missing values, sampling, balancing, and normalization. Additionally, it provides the final clean and balanced dataset used in the analysis, along with detailed instructions on setting up and running the pipeline.

2.2 Data Preparation

Several new features were engineered to facilitate analysis. To ensure comparability of monetary values across the years, features such as `INCTOT` (total personal income) and `VALUEH` (home value) were adjusted for inflation. Using official inflation indices, all monetary values were converted to constant 2023 dollars, providing a standardized economic baseline for longitudinal analysis. A `GENERATION` feature was derived from `BIRTHYR`, assigning individuals to specific generational cohorts: Baby Boomers (1946-1964), Generation X (1965-1980), Millennials (1981-1996), and Generation Z (1997-2012). Individuals under the age of 18 were excluded from the analysis, as the study focuses exclusively on adults. Finally, categorical features such as `STATEFIP`, `SEX`, and `EDUC` were transformed into binary (one-hot encoded) representations to enable their use in machine learning models. This ensured a consistent numerical format across the dataset.

2.3 Handling Missing Values

A systematic approach was implemented to address the missing data. First, explicitly coded missing values (e.g., `9999` for `RENT` or `9999999` for `VALUEH`) were identified and replaced with standard NaN (Not a Number) representations for consistent handling across analyses. Subsequently, rows containing missing values in key features were removed. Finally, features exhibiting a high proportion of missing data, deemed non-essential for the analysis, were excluded entirely.

2.4 Sampling and Balancing

The dataset exhibited imbalances in representation across both years and generations. To mitigate these imbalances and ensure fair comparisons, several sampling strategies were employed. First, to achieve equitable representation across years, the data were sampled to ensure a consistent number of observations per year, accounting for variations in the original yearly dataset sizes. Second, within each year, the generational balance was addressed through a combination of undersampling of overrepresented generations and oversampling of underrepresented generations, particularly Generation Z. Finally, to further refine generational representation and account for age-related biases, an age-based sampling approach was implemented, undersampling older individuals from dominant generations such as Baby Boomers while retaining younger individuals across all generations.

2.5 Final Dataset Structure

The resulting dataset comprises balanced samples across years and generations. Continuous variables, including `INCTOT` and a derived poverty indicator, were standardized using z-score normalization. Categorical variables were fully one-hot encoded, as described previously. Temporal features, specifically `YEAR` and `AGE` features, were retained to facilitate longitudinal and age-based analyses. The detailed descriptions of the features included in the dataset is provided in Table 1.

Table 1: Features Descriptions

FEATURE	DESCRIPTION
GENERATION	Generation of person
YEAR	Census year
ROOMS	Number of rooms
NFAMS	Number of families in household
NCHILD	Number of own children in the household
YNGCH	Age of youngest own child in household
AGE	Age
BIRTHYR	Year of birth
POVERTY	Poverty status
OCCSCORE	Occupational income score
ERSCOR50	Occupational earnings score, 1950 basis
NPBOSS50	Nam-Powers-Boyd score, 1950 basis
STATEFIP	State (FIPS code)
OWNERSHP	Ownership of dwelling (tenure)
KITCHEN	Kitchen or cooking facilities
PLUMBING	Plumbing facilities
UNITSSTR	Units in structure
PHONE	Telephone availability
CBNSUBFAM	Number of subfamilies in household
SEX	Sex
MARST	Marital status
RACE	Race
BPL	Birthplace
SCHOOL	School attendance
EDUC	Educational attainment
EMPSTAT	Employment status
INCTOT_ADJUSTED	Total personal income adjusted for inflation
FTOTINC_ADJUSTED	Total family income
HOUSING_VALUE_ADJ	Monthly contract rent or House value

3 Methodology

The methodology employed in this study integrates statistical analysis and machine learning techniques to examine generational income disparities and predict socioeconomic trends. The analysis incorporates feature selection and correlation analysis to enhance the accuracy and interpretability of the models. It focuses on three main tasks: predicting generational categories and birth years using classification and regression models, analyzing income distributions to assess inequality across generations, and examining how different variables correlate with birth year to identify significant predictors of economic outcomes.

3.1 Feature Selection

For feature selection, we employed mutual information (MI) as a metric to assess the dependency between individual features and the target variable. MI measures both linear and non-linear relationships, making it a robust tool for identifying features with strong predictive potential (Cover & Thomas, 2006). To determine the most relevant features, we applied the filter method, ranking all features based on their MI scores and selecting the top 20 features. This approach ensures that only the most informative features are retained, improving the efficiency and accuracy of the subsequent modeling process (Peng et al., 2005).

In addition to MI, we incorporated the Fisher discriminant score as a supplementary measure. The Fisher score evaluates the class separability of each feature, quantifying how well each feature differentiates between the target classes (Bishop, 2006). Although we did not directly use the Fisher score to select features, it provided valuable insights into the discriminative power of the selected features. By assessing class separability, the Fisher score contributes to a better understanding of which features are most effective in distinguishing between the target classes, thereby enhancing the interpretability of the feature selection process.

To address the potential issue of categorical variables represented by one-hot encoding, we ensured that when a one-hot encoded feature was selected, all related binary features within the one-hot group were also included. This practice prevents the loss of information that could occur if only a single binary feature from the one-hot group were selected. By retaining the full set of related one-hot encoded features, we preserve the consistency of categorical variable representation, ensuring the feature selection process remains aligned with the original structure of the data.

This multi-faceted approach to feature selection—utilizing both MI scores and Fisher discriminant scores—ensures that the chosen features are not only predictive but also contribute meaningfully to the model's ability to differentiate between target classes, while also maintaining consistency in how categorical variables are represented.

3.2 Predictive Models

This study utilized predictive modeling techniques, specifically classification and regression, to analyze and understand socioeconomic trends across generations. Both approaches play a crucial role in uncovering patterns and relationships in the dataset, offering complementary insights into the financial disparities and socioeconomic dynamics that define each generational cohort.

3.2.1 Classification Models

Classification models were employed to predict generational categories (e.g., Baby Boomers, Generation X, Millennials, and Generation Z) based on a range of socioeconomic features. These features included income levels, employment status, homeownership rates, and education attainment. The classification task involved dividing individuals into predefined generational groups based on their birth years and exploring how socioeconomic characteristics differed between these groups.

The classification process began by encoding the generational labels as categorical variables, allowing models to identify patterns within the data. The primary goal of using classification models was to uncover how distinct socioeconomic factors contribute to generational differences. For instance, a classification model might reveal that income levels and housing affordability are significant predictors of generational membership, underscoring their importance in shaping the economic identity of a cohort.

Popular pre-existing classification algorithms—such as Random Forests, Explainable Boosting Machine (EBM), and Logistic Regression—were selected for their balance between predictive accuracy and interpretability. The Explainable Boosting Machine (EBM), in particular, exemplifies this balance—delivering strong classification accuracy and the potential for transparent, feature-level insights. Unlike black-box models such as neural networks, EBM is well-suited for applications where model transparency is valuable, which aligns with our broader research interest in understanding generational patterns. These models were evaluated using accuracy as the primary metric and confusion matrices to provide additional insight into classification performance. Accuracy served as an overall measure of the model's ability to correctly classify individuals into their respective generations, offering a straightforward and interpretable evaluation of performance. The confusion matrices were particularly useful for identifying misclassifications and understanding the model's ability to distinguish between closely related cohorts, such as Millennials and Generation Z, which often share overlapping socioeconomic characteristics. To strengthen the robustness of our evaluation, we employed both an 80/20 train-test split and 10-fold cross-validation. Together, these metrics provided a comprehensive assessment of model effectiveness.

3.2.2 Regression Models

Regression models were utilized to predict the birth year of individuals as a continuous variable, offering a complementary perspective to the classification task. While classification focuses on grouping individuals into discrete generational categories, regression allows for a more granular analysis by identifying how socioeconomic features vary across a continuous timeline. This approach is particularly useful for detecting subtle trends and shifts in financial characteristics over time.

The regression process involved selecting socioeconomic features, such as inflation-adjusted income, education levels, and housing costs, as predictors of birth year. By treating birth year as a continuous outcome, regression models provided insights into how specific economic and demographic factors evolve over time, reflecting broader structural changes in society. For example, a regression analysis might reveal that rising student debt is strongly correlated with more recent birth years, indicating the increasing financial burden on younger generations.

Linear regression was employed in this study to model the relationship between predictors and birth year. We selected linear regression to maintain interpretability and to support the explanatory focus of our analysis. While more complex models—such as ensemble regressors—can offer improved predictive performance, they were not prioritized in this study. Our primary objective was to highlight transparent relationships between socioeconomic factors and birth year, rather than to maximize predictive accuracy. The model was evaluated using several metrics to quantify its performance. The Mean Squared Error (MSE), and Average Absolute Error (AAE) measured the magnitude of deviations between predicted and actual birth years, while the Average Relative Error highlighted the accuracy of predictions relative to the true values. The R^2 Score assessed the proportion of variance in the birth year explained by the model, providing insight into its overall goodness-of-fit. These metrics collectively offered a comprehensive evaluation of the model's performance and its ability to generalize to unseen data.

3.3 Correlation Analysis

Correlation analysis was performed to examine the linear relationships between selected features and the target variable, birth year, using Pearson correlation coefficients. Positive coefficients indicate a direct relationship with more recent birth years, while negative coefficients suggest an inverse relationship with earlier birth years.

To assess statistical significance, p-values were calculated for each coefficient, with values below 0.05 considered significant, highlighting meaningful associations. This analysis aids in identifying relevant features for modeling and complementing feature selection techniques, providing deeper insights into feature importance and interpretability.

4 Economic Metrics

To comprehensively analyze generational economic disparities, this study employed a range of economic metrics to examine income distribution, labor market participation, rent burden, and housing affordability. These metrics provide quantitative insights into the financial realities faced by different generations, highlighting systemic economic shifts and their implications. By evaluating income inequality, employment trends, rent-to-income ratios, and homeownership challenges, this study explores how structural changes have shaped the financial prospects of Baby Boomers, Generation X, Millennials, and Generation Z. Each metric was chosen for its ability to capture specific aspects of generational financial outcomes, enabling a holistic assessment of economic progress and equity. This section details the methods used to calculate these metrics, their significance, and the insights they provide into evolving generational economic dynamics.

4.1 Income Distribution

Income distribution was analyzed using Lorenz Curves, Gini Coefficients (Gini, 1997), income percentile breakdowns, and P90/P10 ratios to capture the breadth of economic inequality across generations. The Lorenz Curve graphically represents cumulative income distribution, offering a visual depiction of inequality. The Gini Coefficient quantifies this inequality, with values ranging from 0 (perfect equality) to 1 (maximum inequality), enabling straightforward comparisons of income disparities between generations. Income percentile breakdowns focused on the bottom 50%, top 10%, and top 1% of earners to illustrate the concentration of wealth within specific groups over time. The P90/P10 ratio, calculated as the income of individuals at the 90th percentile divided by the income of individuals at the 10th percentile, provides a measure of income disparity within a generation. A higher P90/P10 ratio indicates greater gaps between high and low earners, offering additional insight into the extent of inequality beyond the averages. These metrics collectively reveal how income distribution has shifted and whether younger generations experience more pronounced income inequality compared to their predecessors.

4.2 Labor Market Participation

Labor market trends were evaluated by examining employment rates for individuals aged 18–27, as well as across all age groups for each generation. Employment rates, calculated as the proportion of the population engaged in paid work, offer insights into economic activity and engagement. These rates were compared across generations to assess whether younger cohorts, particularly Generation Z, are actively participating in the workforce at levels comparable to or exceeding those of older generations.

4.3 Rent Burden

Rent affordability was assessed using the rent-to-income ratio, calculated as the proportion of income spent on housing costs. This metric highlights the financial strain of meeting basic living expenses, particularly for younger generations such as Millennials and Generation Z. A higher rent-to-income ratio reflects greater economic pressure, inhibiting savings and wealth accumulation.

4.4 Housing Affordability and Homeownership

Housing affordability was evaluated by comparing average inflation-adjusted house values to average inflation-adjusted incomes over time. The maximum affordable mortgage, estimated as three times the household income, was used to assess whether generational earnings align with housing costs. Furthermore, the income increase required to bridge the gap between affordable mortgages and average home values was calculated, highlighting barriers to homeownership for Generation Z. Homeownership is widely recognized as a key pathway to wealth accumulation, particularly for low-income households, making this analysis critical to understanding the long-term financial implications of declining housing accessibility for younger generations.

5 Results

This section presents key findings from both statistical and machine-learning analyses, organized into five subsections. Feature Selection identifies variables most relevant to generational differences. Regression and Classification Results evaluate model performance in predicting birth years and generational cohorts. Correlation Results uncover patterns among socioeconomic variables, while Economic Metrics highlight income inequality and financial disparities across generations. Together, these results provide a comprehensive view of the factors shaping generational differences.

5.1 Feature Selection Results

To understand what factors are most strongly connected to a person's birth year, we used several methods: mutual information (MI), Fisher scores, and p-values. These tools help us measure how much each feature in the data relates to birth year, both in terms of general strength (MI), how well each feature helps to distinguish between people born in different years (Fisher score), and whether the results are statistically meaningful (p-values).

Our results, shown in Table 2, highlight the top 20 features. HOUSING_VALUE_ADJ (monthly contract rent or house value) was the strongest predictor, meaning that housing values often reflect generational differences. Similarly, INCTOT_ADJUSTED (total personal income adjusted for inflation) and NPBOSS50 (Nam-Powers-Boyd score) were also highly important. The NPBOSS50 score reflects the social and economic position of individuals based on their occupations, capturing generational shifts in job types and education levels. These findings suggest that both economic conditions and occupational status play a significant role in distinguishing between generations.

Table 2: Selected Features (Ranked by MI Score)

Feature name	MI Score	Fisher Score	P-Value
HOUSING_VALUE_adjusted	0.7386	9.4550	$< 10^{-5}$
INCTOT_adjusted	0.3768	22.8839	$< 10^{-5}$
NPBOSS50	0.3407	3.1002	$< 10^{-5}$
ERSCOR50	0.3237	3.0257	$< 10^{-5}$
FTOTINC_adjusted	0.1842	13.1118	$< 10^{-5}$
PLUMBING_21	0.1247	1289.6349	$< 10^{-5}$
PLUMBING_22	0.1247	4.3309	$< 10^{-5}$
PLUMBING_12	0.1247	9.4044	$< 10^{-5}$
PLUMBING_20	0.1247	1239.1866	$< 10^{-5}$
PLUMBING_14	0.1247	2.3760	$< 10^{-5}$
YNGCH	0.0990	127.2489	$< 10^{-5}$
OCCSCORE	0.0761	15.7726	$< 10^{-5}$
POVERTY	0.0622	29.5281	$< 10^{-5}$
KITCHEN_3	0.0418	2.2006	$< 10^{-5}$
KITCHEN_5	0.0418	410.0394	$< 10^{-5}$
KITCHEN_4	0.0418	253.6757	$< 10^{-5}$
MARST_6	0.0227	141.5336	$< 10^{-5}$
MARST_5	0.0227	10.4430	$< 10^{-5}$
MARST_4	0.0227	7.9465	$< 10^{-5}$
MARST_3	0.0227	2.3478	$< 10^{-5}$

MARST_2	0.0227	1.9914	$< 10^{-5}$

We also found that household characteristics, such as plumbing and kitchen facilities, were significant predictors. For example, PLUMBING_20 indicates households with complete plumbing, which reflects improved housing standards and modern living conditions often associated with younger generations. Meanwhile, PLUMBING_21 refers to plumbing facilities used only by the household, highlighting private access to utilities, another marker of better housing quality. In contrast, PLUMBING_22 represents plumbing shared with others, a condition more commonly found in older or lower-income housing arrangements, which were typical in past generations.

Kitchen features also proved relevant. KITCHEN_5 refers to households with an exclusive-use kitchen, typically found in more modern homes and reflecting higher living standards. The availability of private kitchen facilities provides further clues about generational shifts in housing quality.

Even features such as YNGCH (age of youngest own child in the household) were helpful for predicting birth year, as family structure and the presence of younger children often vary across generations, providing insights into household demographics and life stages.

By combining these different measures, we selected features that not only improve the accuracy of our predictions but also give us meaningful insights into how social, economic, and household factors vary across generations.

5.2 Correlation Analysis Results

The correlation analysis revealed key relationships between the features and the target variable, birth year. PLUMBING_20 emerged as the most positively correlated feature (Pearson correlation = 0.535291), indicating a strong direct relationship with birth year, followed by MARST_6 (Pearson correlation = 0.410951) and KITCHEN_4 (Pearson correlation = 0.227687). These features demonstrated their relevance in predicting more recent birth years, with high statistical significance ($p < 10^{-5}$).

On the negative side, PLUMBING_21 showed the strongest Pearson inverse correlation with birth year (Pearson correlation = -0.534888, $p < 10^{-5}$), followed by YNGCH (Pearson correlation = -0.307273, $p < 10^{-5}$). These negative correlations suggest that increases in these feature values are associated with older birth years. Other notable negative Pearson correlations include KITCHEN_5 (-0.255492) and POVERTY (-0.143595), reflecting systemic or socioeconomic patterns linked to earlier generations.

Most features demonstrated statistically significant Pearson correlations, with p-values below 10^{-5}, confirming the reliability of the observed relationships. However, a few features, such as MARST_2 (Pearson correlation = 0.006087, p = 0.249559) and KITCHEN_3 (Pearson correlation = -0.008190, p = 0.121364), exhibited weak and statistically insignificant correlations, indicating limited predictive utility.

The correlation analysis results, including Pearson coefficients and p-values for all features, are summarized in Table 3. These findings highlight the varying influence of features on birth year prediction, with both positive and negative correlations critical to the regression model's predictive capacity. These insights will guide feature selection and model optimization.

Table 3: Correlation and P-Values for Selected Features

FEATURE NAME	PEARSON CORRELATION	P-VALUE
PLUMBING_20	0.535291	$< 10^{-5}$
MARST_6	0.410951	$< 10^{-5}$
KITCHEN_4	0.227687	$< 10^{-5}$
MARST_2	0.006087	0.249559
KITCHEN_3	-0.008190	0.121364
ERSCOR50	-0.031943	$< 10^{-5}$
PLUMBING_14	-0.033424	$< 10^{-5}$
NPBOSS50	-0.034276	$< 10^{-5}$
MARST_3	-0.037826	$< 10^{-5}$
PLUMBING_22	-0.041601	$< 10^{-5}$
PLUMBING_12	-0.045062	$< 10^{-5}$
HOUSING_VALUE_ADJUSTED	-0.046916	$< 10^{-5}$
FTOTINC_ADJUSTED	-0.063988	$< 10^{-5}$

OCCSCORE	-0.069503	$< 10^{-5}$
MARST_5	-0.074974	$< 10^{-5}$
MARST_4	-0.094504	$< 10^{-5}$
INCTOT_ADJUSTED	-0.122010	$< 10^{-5}$
POVERTY	-0.143595	$< 10^{-5}$
KITCHEN_5	-0.255492	$< 10^{-5}$
YNGCH	-0.307273	$< 10^{-5}$
PLUMBING_21	-0.534888	$< 10^{-5}$

5.3 Regression Results

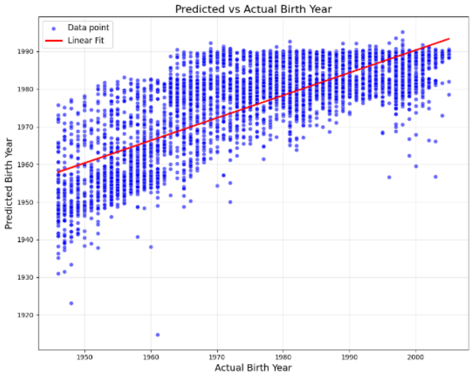


Figure 2: Result of regression illustrates the relationship between predicted and actual birth years, with the red line representing the linear fit. The data points cluster around the diagonal line, indicating that the model generally predicts birth years accurately.

The results of the regression analysis in Figure 2 reveal the model's capability to predict the birth year based on the given socioeconomic features. As shown in the scatter plot of Predicted vs. Actual Birth Year, the linear regression model demonstrates a clear positive trend, with the predicted values generally aligning with the actual birth years. The R^2 score of 0.6005 indicates that approximately 60.05% of the variance in birth year can be explained by the model, showcasing a moderate level of predictive power.

The Mean Squared Error (MSE) of 105.8127 and the Average Absolute Error (AAE) of 8.1880 highlight the magnitude of deviations between the predicted and actual birth years, with the model achieving an average prediction error of just over 8 years. Considering that the median generation length is 16 years, the AAE of 8 years suggests that, on average, the predicted birth year is likely to fall within the same generation as the actual birth year. This reinforces the model's utility in providing generation-level insights, even if exact birth year predictions are not perfect. The Average Relative Error of 0.41% further emphasizes the accuracy of the model in relative terms, suggesting minimal percentage-based deviations from actual values.

The Pearson correlation coefficient between predicted and actual birth years is 0.7749, indicating a strong positive linear relationship. The statistical significance of this correlation is confirmed by a p-value of ($p < 10^{-5}$), which underscores the reliability of the observed relationship. Despite these strengths, the scatter plot reveals some dispersion around the linear fit, particularly for older birth years, indicating areas where the model's predictions deviate more from the true values.

Overall, the results demonstrate that the regression model provides a reasonable and interpretable framework for predicting birth year, capturing meaningful relationships between socioeconomic predictors and the dependent variable. Moreover, the model's performance at the generation level, as indicated by the AAE relative to the median generation length, highlights its practical applicability for generational analysis while leaving room for further refinement to improve precision.

5.4 Classification Results

Table 4: Performance Comparison of Machine Learning Algorithms Based on Accuracy

ALGORITHM	ACCURACY (80/20 SPLIT)	ACCURACY (10-FOLD CV)
ZEROR (BASELINE)	29.72%	29.68%

RANDOM FOREST	70.35%	70.07%
GRADIENT BOOSTING	66.41%	65.61%
LOGISTIC REGRESSION	53.75%	53.51%
DECISION TREE	67.37%	67.72%
EBM	74.62%	74.78%

The ZeroR model achieved an accuracy of 29.72% by always predicting the majority class "Baby Boomers." In cross-validation, ZeroR achieved a similar accuracy of 29.68%, reflecting the baseline nature of the model. This model serves as a baseline for comparison, against which the performance of other models is evaluated. In contrast, all other models achieved significantly higher accuracy, indicating that they can detect meaningful signals in the data that differentiate between generational categories. This high accuracy, observed consistently across both the percentage split and cross-validation, highlights that the dataset contains distinct patterns or features that are characteristic of each generation, validating the relevance of the features used in the models. The classification results for the evaluated models are summarized in Table 4, and the corresponding normalized confusion matrices in Figures 3 provide detailed insights into their performance.

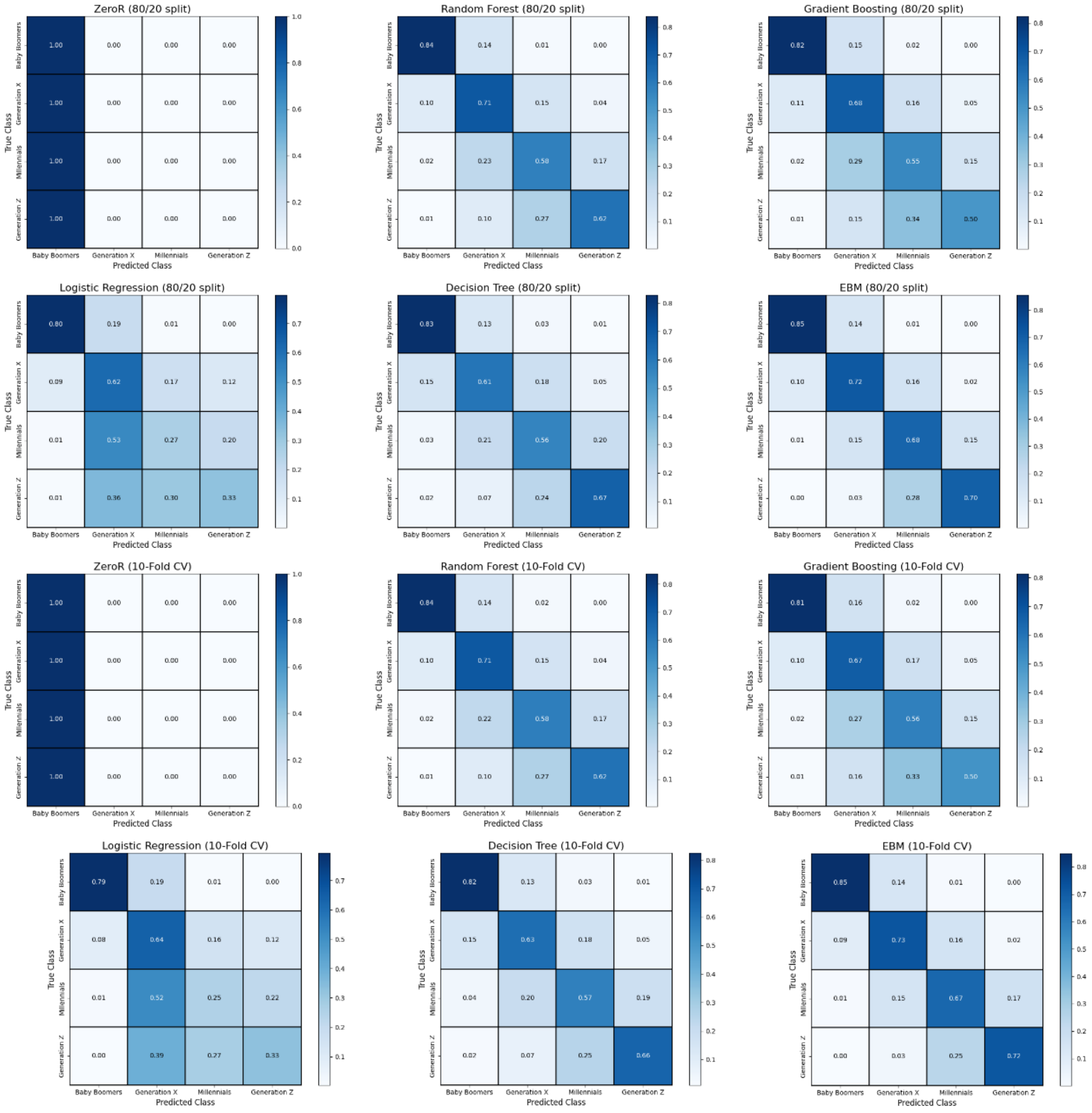


Figure 3: The normalized confusion matrices for the evaluated models, with the top two rows showing results from the 80/20 train-test split and the bottom two rows from 10-fold cross-validation, provide detailed insights into their performance as summarized in Table 4.

Explainable Boosting Machine (EBM) achieved the highest accuracy among all models, with 74.62% in the percentage split and a slightly higher 74.78% in cross-validation, demonstrating its strong capability to handle the dataset while maintaining interpretability. Random Forest followed with accuracies of 70.35% in the percentage split and 70.07% in cross-validation, reflecting solid performance with minimal misclassifications. The Decision Tree model achieved 67.37% in the percentage split and 67.72% in cross-validation, providing interpretable results alongside competitive accuracy. Gradient Boosting, while slightly lower, attained 66.41% in the percentage split and 65.61% in cross-validation, maintaining reasonable predictive power but trailing behind the tree-based models. Logistic Regression delivered lower performance, with accuracies of 53.75% in the percentage split and 53.51% in cross-validation, indicating its limited ability to capture complex patterns within the data.

As shown in the normalized confusion matrices in Figure 3, models such as EBM and Random Forest offer balanced predictions across generational categories, effectively minimizing misclassifications, particularly for "Generation Z" and "Baby Boomers." These results highlight the models' ability to leverage meaningful features that distinguish between generations, reinforcing the validity of the dataset and its potential for deeper generational insights and applications.

5.5 Economic Metrics Results

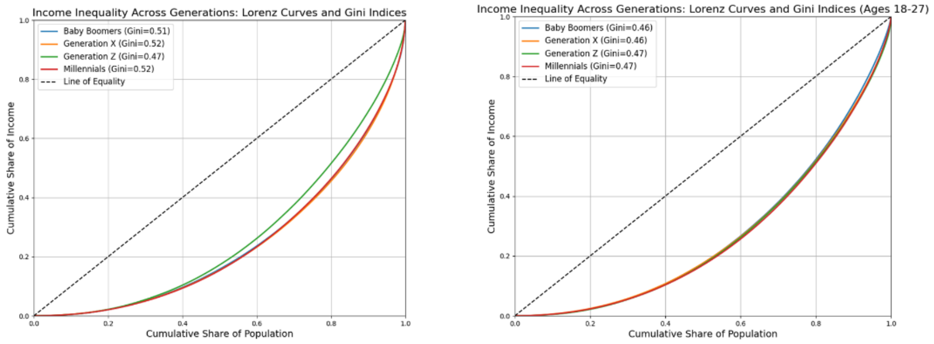


Figure 4: Lorenz curves and Gini indices illustrating income inequality across generations for all ages (left) and ages 18–27 (right) in the United States.

The results reveal significant trends and disparities across generations in income distribution, housing affordability, and economic opportunities. Generation Z experiences lower levels of overall income inequality compared with individuals of all ages from different generations, as reflected in their Gini index (0.47 versus 0.51–0.52 for other generations), as shown in Figure 4. However, for individuals aged 18–27, the Gini index for Generation Z (0.47) is comparable to Millennials and slightly higher than Baby Boomers and Generation X (both at 0.46), suggesting a narrower gap in income inequality among younger cohorts. Similarly, the P90/P10 ratio for Generation Z is the highest (19.9) across all ages, indicating significant income disparities between the highest and lowest earners within the generation, as depicted in Figure 5. For ages 18–27, the P90/P10 ratio for Generation Z remains the highest, further highlighting these disparities, even among younger individuals.

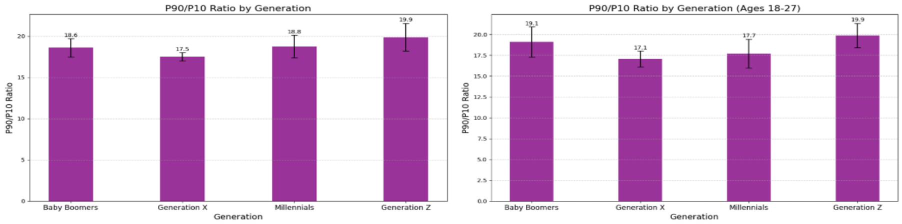


Figure 5: Bar chart illustrating the P90/P10 ratio, a measure of income inequality, across different generations (All ages vs ages 18-27) in the United States.

This dynamic is further evidenced by the distribution of income shares. Generation Z demonstrates a more equitable income distribution at the lower end compared to other generations across all ages, with 66.7% of total income concentrated within the bottom 50%. However, for individuals aged 18–27, Baby Boomers slightly surpass Generation Z with 67.3% of income held by the bottom 50%, compared to Generation Z's 66.7%, followed by Millennials (66.5%) and Generation X (66.3%). Despite this equitable distribution at the lower end, Generation Z holds the highest income share at the top 1% (7.8%) among individuals aged 18–27, surpassing Millennials

(6.6%), Generation X (7.0%), and Baby Boomers (5.6%). This pattern underscores a widening gap between the highest earners and the rest of Generation Z, particularly when focusing on younger cohorts.

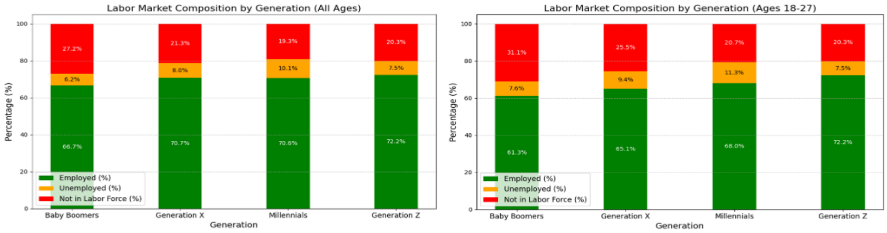


Figure 6: Stacked bar charts illustrating the labor market composition by generation for the entire population (left) and ages 18–27 (right) in the United States.

Moreover, as generations progress, a clear trend emerges: the income share held by the top 10% (excluding the top 1%) shrinks, while the top 1% consolidates an increasingly larger share of total income. For example, the top 10% share decreases from 27.1% for Baby Boomers to 25.6% for Generation Z, while the top 1% share simultaneously increases. This suggests that individuals from lower income groups tend to move into higher income brackets over time, leading to a redistribution of income towards the upper tiers. While this upward mobility may appear positive, it ultimately widens income inequality by concentrating more wealth within the top 1%, creating structural barriers to equitable wealth distribution.

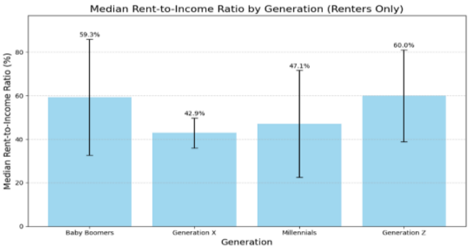


Figure 7: Bar chart illustrating the median rent-to-income ratio for different generations of renters in the United States.

These findings highlight how the concentration of wealth within the top 1% exacerbates economic inequality, disproportionately affecting younger cohorts like Generation Z. This trend compounds their financial challenges and limits their ability to achieve long-term economic stability, underscoring the systemic nature of these disparities.

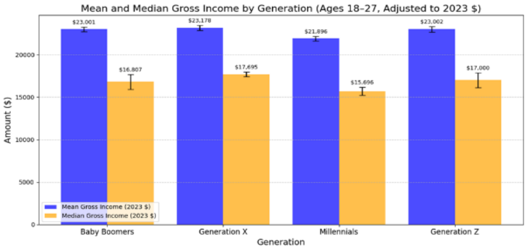


Figure 8: Bar chart showing the mean and median gross income (in 2023 dollars) by generation for ages 18–27 in the United States.

The narrative that members of Generation Z are 'lazy' or 'don't want to work' is contradicted by their higher employment rates compared to those of other generations, as shown in Figure 6, both overall and within the 18-27 age range. At ages 18-27, Generation Z leads with an employment rate of 72.2%, surpassing Millennials (68.0%), Generation X (65.1%), and Baby Boomers (61.3%) at the same life stage, highlighting their active participation in the workforce. Despite their workforce participation, Generation Z continues to face financial challenges and bears the greatest median rent-to-income burden (60.0%), surpassing Millennials (47.1%), Generation X (42.9%), and Baby Boomers (59.3%), as demonstrated in Figure 7, reflecting worsening economic pressures for younger generations. This is particularly concerning given that Generation Z earns a comparable mean income to individuals in the same age range across older generations, as shown in Figure 8. However, despite earning similar incomes, Generation Z suffers from much higher rent-to-income ratios and housing affordability challenges, which significantly impact on their financial stability and ability to build wealth.

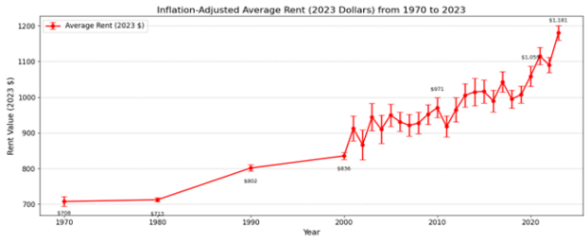


Figure 9: Line chart illustrating inflation-adjusted average rent (in 2023 dollars) from 1970 to 2023 across the United States.

The broader systemic issue of rising housing costs is evident in inflation-adjusted average rent, which has increased steadily from $708 in 1970 to $1,181 in 2023—a 66% rise in real terms, as shown in Figure 9. This sharp rise, particularly after 2000, has compounded the financial strain on younger generations, making it increasingly difficult for them to achieve economic security. These findings underscore the urgent need for policy interventions to address rising rents and ensure equitable access to affordable housing.

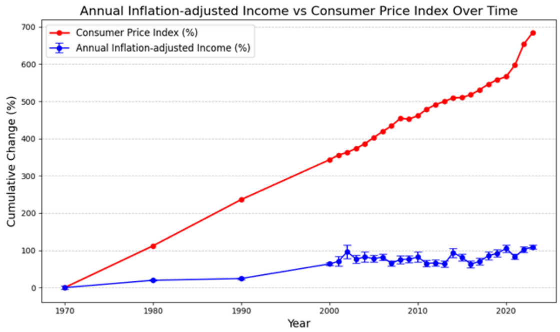


Figure 10: Annual inflation-adjusted income vs. Consumer Price Index (CPI) cumulative change over time (1970–2023). The CPI data were retrieved using the Federal Reserve Economic Data API (U.S. Bureau of Labor Statistics, 2025).

To further understand the economic pressures faced by different generations, it is crucial to examine the drastic disparity between inflation and income growth over time. As shown in Figure 10, the Consumer Price Index (CPI), obtained from the Federal Reserve Economic Data (U.S. Bureau of Labor Statistics, 2025), has surged by 700% over the past five decades, indicating a sharp rise in the cost of living, while inflation-adjusted income has increased by only 100%. This imbalance highlights the shrinking ability of wages to keep pace with rising costs, which disproportionately affects younger generations. To put this into perspective, in 1970, the average cost of a week's worth of groceries for a family was approximately $30 (inflation-adjusted). A worker earning $15 per hour in 2023-equivalent dollars could cover this cost with just 2 hours of work. Today, that same basket of groceries would cost around $210, but a worker's hourly income would have risen to only $30—requiring 7 hours of work. This stark increase highlights the disproportionate growth of living expenses relative to wages, leaving less financial flexibility for savings, investments, or other essential needs. While this issue disproportionately affects younger generations, particularly Generation Z, who face additional economic barriers such as stagnant wages and skyrocketing housing costs, the trend impacts all generations. If this trajectory continues unchecked, Generation Alpha is likely to face even greater financial challenges than Generation Z, further exacerbating systemic financial insecurity and severely limiting opportunities for upward mobility.

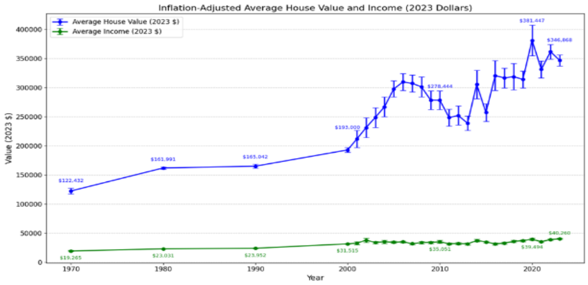


Figure 11: Line chart illustrating inflation-adjusted average house value and income (in 2023 dollars) from 1970 to 2023 in the United States.

Moreover, as highlighted in Figure 11, the disparity between house values and household incomes has widened significantly over time, with house costs increasing at a much faster rate than income. Since 1970, household income has increased by approximately 100%, while the average house value has surged by nearly 180%, exacerbating the affordability gap. In 2023, the average house value was approximately $350,000, while the average household income for two earners was $80,000, making the maximum affordable mortgage $240,000—$110,000 short of the average house price. This stark imbalance underscores the escalating affordability crisis, requiring households to increase their income by approximately 150% to afford a home, an unfeasible goal to achieve on a large scale without systemic changes. This affordability crisis disproportionately impacts Generation Z and severely limits their ability to achieve homeownership. As homeownership is one of the most effective ways for households—especially low-income households—to build wealth over time (Harvard Joint Center for Housing Studies, 2006), the inability of Generation Z to access homeownership not only affects their current financial stability but also undermines their prospects for upward mobility and long-term wealth accumulation, perpetuating systemic inequality across generations.

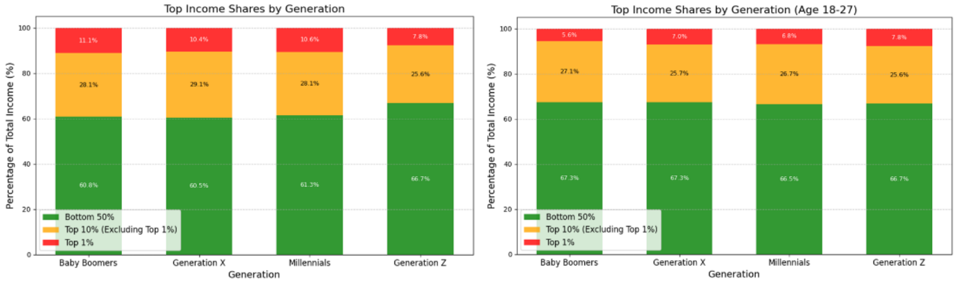


Figure 12: Stacked bar charts illustrating the distribution of income among different generations (all population vs ages 18-27) in the United States.

While these findings highlight the immediate challenges facing Generation Z, an interesting observation emerges when examining how income distribution evolves over time. As generations age, the gap between the top earners and bottom earners widens significantly. For individuals aged 18–27, income distribution is relatively balanced, with the bottom 50% holding 67.3% of total income for Baby Boomers and Generation X. However, this share declines to 60.8% and 60.5% for these same generations across all ages, reflecting a growing disparity. The top 1% share also increases notably, from 5.6% for Baby Boomers (aged 18–27) to 11.1% for all ages (Figure 12).

These trends are further illustrated by key inequality metrics. For example, the P90/P10 ratio for Millennials increases from 17.7 at ages 18–27 to 18.8 across all ages, illustrating how income inequality grows as the cohort ages (Figure 5). Similarly, Gini indices increase from 0.46–0.47 in younger groups to 0.51–0.52 for older ones, reflecting a clear progression of inequality (Figure 4). Together, these measures highlight how wealth and income become increasingly concentrated among top earners as generations progress.

This observation makes sense given that, at younger ages, individuals tend to start from relatively similar socioeconomic positions as they enter adulthood, often with limited income and wealth. Over time, as people make different life choices, pursue varied career paths, and encounter diverse opportunities or challenges, their financial trajectories naturally diverge. Some may benefit from lucrative careers, advantageous networks, or entrepreneurial success, while others might face barriers that limit their upward mobility. These individual trajectories are further compounded by systemic factors—such as unequal access to resources, education, and opportunities—which exacerbate disparities over time. Collectively, these dynamics likely contribute to the persistent and growing inequality observed across generations.

6 Conclusion

This study provides key insights into the socioeconomic disparities faced by different generations, particularly Generation Z, who experience significant financial challenges. Classification models highlighted distinct patterns in socioeconomic features that differentiate generations, with the Explainable Boosting Machine achieving the highest accuracy. Regression analysis further demonstrated the model's ability to capture temporal trends, with predicted birth years aligning closely with actual values, particularly at the generational level.

Economic metrics revealed that Generation Z faces the highest median rent-to-income burden (60.0%) and the greatest barriers to homeownership, with a $110,000 gap between affordable mortgages and average house prices. Despite higher workforce participation (72.2%) compared to previous generations at similar life stages, these efforts are undermined by disproportionate income disparities, as evidenced by their high P90/P10 ratio (19.9). While their income distribution is more equitable at the lower end, systemic issues like unaffordable housing and stagnant wage growth hinder financial stability and wealth accumulation.

The findings highlight the value of machine learning in uncovering generational trends and systemic inequalities. While this study focuses on quantitative analysis, future research would benefit from integrating qualitative insights, such as psychological, behavioral, and cultural perspectives, to enrich the contextual understanding of the data and provide a more holistic view of generational experiences. Future work should also investigate targeted policy interventions to address the economic challenges facing Generation Z and promote equity across generations.

References

Aljishi, A., Marjani, M., Latifi, A., & Shamir, L. (2025). GenZ-Economic-Challenges [Data set and source code]. GitHub. https://github.com/MatinMarjani/GenZ-Economic-Challenges

Autor, D. H., Katz, L. F., & Krueger, A. B. (1998). Computing inequality: Have computers changed the labor market? *The Quarterly Journal of Economics*, 113(4), 1169–1213.

Bianchi, S. M. (2000). Maternal employment and time with children: Dramatic change or surprising continuity? Demography, 37(4), 401–414. https://doi.org/10.1353/dem.2000.0001

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Bound, J., & Turner, S. (2007). Cohort crowding: How resources affect collegiate attainment. *Journal of Public Economics*, 91(5–6), 877–899.

Charles, K. K., & Hurst, E. (2003). The correlation of wealth across generations. *Journal of Political Economy*, 111(6), 1155–1182.

Chetty, R., Grusky, D., Hell, M., Hendren, N., Manduca, R., & Narang, J. (2017). The fading American dream: Trends in absolute income mobility since 1940. Science, 356(6336), 398–406.

Cover, T. M., & Thomas, J. A. (2006). Elements of information theory (2nd ed.). Wiley.

Elder, G. H. (2018). Children of the Great Depression: Social change in life experience (25th anniversary ed.). Routledge.

Fan, C., Xu, J., Natarajan, B. Y., & Mostafavi, A. (2023). Interpretable machine learning learns complex interactions of urban features to understand socio-economic inequality. *Computer-Aided Civil and Infrastructure Engineering*, 38(14), 2013–2029.

Gallipoli, G., Low, H., & Mitra, A. (2020). Consumption and income inequality across generations. Centre for Economic Policy Research.

Green, R. K., & Lee, H. (2016). Age, demographics, and the demand for housing, revisited. Regional Science and Urban Economics, 61, 86–98.

Gregory, V. (2023). Generational gaps in income and homeownership. Economic Synopses, (15). https://doi.org/10.20955/es.2023.15

Gini, C. (1997). Concentration and dependency ratios. Rivista di Politica Economica, 87, 769–792.

Harvard Joint Center for Housing Studies. (2006). The state of the nation's housing 2006. Harvard Joint Center for Housing Studies.

Ipsos. (2023). We need to talk about generations: Understanding generations. https://www.ipsos.com/en/we-need-talk-about-generations-understanding-generations

Lev, T. A. (2021). Generation Z: Characteristics and challenges to entering the world of work. *Cross-Cultural Management Journal*, 23(1), 107–115. https://doi.org/10.22381/CCMJ2320216

Mishel, L., & Bivens, J. (2021). Identifying the policy levers generating wage suppression and wage inequality. Economic Policy Institute, 13.

Palfrey, J., & Gasser, U. (2011). Born digital: Understanding the first generation of digital natives. Basic Books. https://books.google.com/books?id=wWTI-DbeA7gC

Patterson, J. T. (1996). Grand Expectations: The United States, 1945-1974. Oxford University Press.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.

Ruggles, S., Flood, S., Sobek, M., Backman, D., Chen, A., Cooper, G., Richards, S., Rogers, R., & Schouweiler, M. (2024). PUMS USA: Version 15.0 [Data set]. IPUMS. https://usa.ipums.org/usa/

Twenge, J. M. (2023). Generations: the real differences between Gen Z, Millennials, Gen X, Boomers, and Silents—and what they mean for America's future. Simon & Schuster.

U.S. Bureau of Labor Statistics. (2025). Consumer Price Index for all urban consumers: All items in U.S. city average (CPIAUCSL) [Data set]. Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/CPIAUCSL

Western, B., & Rosenfeld, J. (2011). Unions, norms, and the rise in US wage inequality. American Sociological Review, 76(4), 513-537.

The Impact of Scaling Techniques on Breast Cancer Prediction Algorithms

[1]Oluwaseyi Ezekiel Olorunshola, [2*]Okeh Dominic Ebuka, [3]Adeniran Kolade Ademuwagun, [4]Fatimah Adamu-Fika and [5]Muhammad Abdullahi Kabir

[1, 2, 3]Computer Science Department, Air Force Institute of Technology, Kaduna, Nigeria
[4, 5]Cyber Security Department, Air Force Institute of Technology, Kaduna, Nigeria

email: [1]seyisola25@yahoo.com, [2*]okehoroko2019@gmail.com, [3]kademuwagun@gmail.com, [4]f.adamu-fika@afit.edu.ng, [5]abdullahimk82@gmail.com

*Corresponding author

Received: 10 December 2024 | Accepted: 24 April 2025 | Early access: 04 May 2025

Abstract - *Breast cancer develops when the genetic material of breast cells undergoes mutations, causing the cells to grow uncontrollably and form tumors. Efforts however have been made to combat it by developing machine learning models to help clinicians with early detection. This study investigates the impact of scaling techniques on the performance of algorithms used for breast cancer prediction. Two scaling approaches were compared with models utilizing the raw, unscaled data. The result revealed that the different scaling techniques had minimal effect on the prediction performance after Hyperparameter tuning. This suggests that for the specific dataset and algorithms used, potential sources of bias were analyzed and the classifiers adapted their internal parameters to compensate for the difference in feature scaling. The model's performance was evaluated using four metrics which are Accuracy, Recall, Precision, and F1-score through the 5-fold cross-validation. The results of this study showed that the Random Forest an ensemble model outperformed all other individual classifier after hyperparameter tuning was performed, it had an Accuracy value of 0.9578, a Recall value of 0.9297, a Precision of 0.9571, and an F1-score of 0.9425.*

Keywords: Breast cancer, benign, malignant, hyperparameter tuning, ensemble.

1 Introduction

Breast cancer is a type of tumor that occurs in the tissues of the breast. It is the most common type of cancer found in women around the world (Fatima et al., 2020). It is the most recognized global malignancy and the leading cause of cancer deaths. Despite this, undergraduate and postgraduate exposure to breast cancer is limited, impacting the ability of clinicians to accurately recognize, assess, and refer appropriate patients (Katsura et al., 2022). It is also a fact that most breast cancer cases are discovered late (Mahesh et al., 2022). The early symptoms of breast cancer may not be apparent, but it commonly presents itself as a lump in the breast and is usually painless (Huang et al., 2024). It has been discovered that 90% of breast masses are benign, such as fibroadenomas, cysts, and fibrocystic change (WHO, 2021), and although extremely common, breast pain in isolation without other signs is rarely a presentation of breast cancer (Fonseca et al., 2019). It is highly curable when they are diagnosed early before they metastasize (Harbeck et al., 2019). The diagnosis of breast cancer is time consuming due to the limited availability of diagnostic systems such as dynamic MRI, X-rays, etc. (Das et al., 2024), but recent research indicates that early detection of the disease can lead to a positive prognosis and a higher survival rate. The integration of machine learning (ML) algorithms in breast cancer prediction holds promising potential for improving accuracy. The effectiveness of these algorithms can however be significantly impacted by data preprocessing techniques, particularly scaling. This important preprocessing step ensures that all features are on a similar scale, which can potentially enhance model convergence and interpretability.

This research study focuses on investigating the impact of scaling techniques on the performance of breast cancer prediction algorithms. The study involves the analysis of the performance of five different classifiers using two distinct scaling techniques. A comprehensive comparison is made between the scaled and unscaled models to discern the impact of scaling. Additionally, the study further explores the effect of hyperparameterization on the scaled and unscaled models in order to better understand the relationship between scaling techniques and model performance. Following the initial analysis, a comprehensive examination of the scaled and unscaled models is carried out, integrating hyperparameter tuning to refine the models. As a result of this thorough analysis, a stacking ensemble ML model tailored for breast cancer diagnosis is developed. The ensemble model is constructed using the two best performing classifiers after hyperparameter tuning with the most effective scaling technique identified during the study. The results of this research demonstrated that employing specific feature engineering techniques can significantly enhance the overall performance of the ensemble model while also helping to mitigate potential biases. It is evident that the choice of scaling technique plays a critical role in influencing the model's capacity to capture relevant information from the data and make unbiased predictions across various subgroups. The objective of this research is two-fold: firstly, to analyze ML algorithms that produced the best result and secondly to build a superior ML model by combining the best two analyzed methods using the stacking ensemble method that can predict breast cancer using the Wisconsin Breast Cancer Diagnostic (WBCD) dataset.

The study includes implementing and evaluating different ML algorithms and ultimately creating an ensemble model to enhance prediction accuracy. The dataset used in the study is the WBCD from the University of California Irvine (UCI) repository and was preprocessed, analyzed and prepared for training and testing. The remaining part of this paper is arranged as follows; Section 2 contains a literature review while Section 3 contains the methodology. Results are analyzed and discussed in Section 4 while Section 5 concludes the paper.

2 Literature Review

Ahsan et al. (2021) carried out an experiment using data preprocessing steps like feature reduction, data conversion, and data scaling to create a standard dataset. This was important in reducing inaccuracy in final prediction. In the research, eleven ML algorithms which were Logistic Regression (LR), Linear Discriminant Analysis, K-Nearest Neighbors (KNN), Classification and Regression Trees , Naive Bayes (NB), Support Vector Machine (SVM), XGBoost, Random Forest (RF), Gradient Boost, AdaBoost, Extra Tree Classifier were analyzed under six different data scaling methods which were Normalization (NR), Standard scale (SS), MinMax (MM), MaxAbs (MA), Robust Scaler (RS), and Quantile Transformer (QT). The result showed that Classification and Regression Tree, along with RS or QT, outperforms all other ML algorithms with 100% accuracy, 100% precision, 99% recall, and 100% F1 score.

Ambarwari et al. (2020) demonstrated that data scaling techniques like MinMax normalization and standardization significantly impact data analysis. The research utilized machine learning algorithms such as KNN, Naïve Bayes, ANN, and SVM with an RBF kernel. The findings showed that Naïve Bayes maintained the most consistent performance without data scaling, while KNN was more stable than both SVM and ANN. Nevertheless, their computational results indicated that combining MinMax scaling with SVM yielded the best overall performance.

Shahriyari et al. (2019) demonstrated that normalization significantly influences the performance of various machine learning classifiers. Their study involved twelve different ML algorithms, including several commonly used in heart disease prediction, and applied multiple normalization techniques. The results highlighted a strong relationship between the choice of normalization method and the effectiveness of the algorithms. Among the eleven supervised models, SVM achieved the highest accuracy at 78%. However, Naïve Bayes stood out by offering the best overall performance in terms of both accuracy and the shortest training time.

Balabaeva et al. (2020) explored the impact of various scaling techniques on heart failure patient datasets. Their research employed advanced machine learning algorithms including XGBoost, Logistic Regression, Decision Trees, and Random Forest, alongside scaling methods such as Standard Scaler, MinMaxScaler, MaxAbsScaler, RobustScaler, and QuantileTransformer. The results indicated that Random Forest performed better when combined with StandardScaler and RobustScaler. In contrast, the performance of the Decision Tree algorithm remained unaffected by the choice of scaling method.

Singh and Singh (2020) carried out an analysis to investigate the impact of fourteen data normalization methods on classification performance while also considering the full feature set, feature selection, and feature weighting. Also, a modified Ant Lion optimization that searches feature subsets and the best feature weights along with the parameter of Nearest Neighbour Classifier was presented in the research. The Experiments were performed on 21 publicly available real and synthetic datasets, and results were analyzed based on accuracy, the percentage of feature reduced, and runtime. From the results, it was observed that no single method outperformed the others. Therefore, a set of the best and the worst methods combining the normalization procedure and empirical analysis of results was suggested. After this, it was observed that the better performers were the z-Score and Pareto Scaling for the full feature set and feature selection, and tanh and its variant for feature weighting. The Mean Centered, Variable Stability Scaling and Median and Median Absolute Deviation methods along with un-normalized data were the worst performers.

Yang et al. (2021) proposed a prediction model for breast cancer recurrence based on clinical nominal and numeric features. In the study, the data used consisted of 1,061 patients from the Breast Cancer Registry from Shin Kong Wu HoSu Memorial Hospital between 2011 and 2016, in which 37 records were denoted as breast cancer recurrence. The approach used consisted of three stages. First, data pre-processing and feature selection techniques to consolidate the dataset was carried out. Among all features, six features were identified for further processing in the following stages. Next, resampling techniques were applied to resolve the issue of class imbalance. Finally, the construction of the two classifiers, AdaBoost and cost-sensitive learning, to predict the risk of recurrence and carrying out the performance evaluation in three-fold cross-validation. By applying the AdaBoost method, an accuracy of 0.973 and sensitivity of 0.675 was achieved. By combining the AdaBoost and cost-sensitive method of the proposed model, a reasonable accuracy of 0.468 and substantially high sensitivity of 0.947 which guarantee almost no false dismissal was achieved.

Elsadig et al. (2023) selected eight classification algorithms that had been used to predict breast cancer to be under investigation. These classifiers include single and ensemble classifiers. A trusted dataset has been enhanced by applying five different feature selection methods to pick up only weighted features and neglect others. Accordingly, a dataset of only 17 features has been developed, SVM was ranked at the top by obtaining an accuracy of 97.7% with classification errors of 0.029 False Negative (FN), and 0.019 False Positive (FP). Therefore, it was noteworthy that SVM was the best classifier and it outperformed even the stack classier.

Using the WBCD dataset, Strelcenia and Prakoonwit (2023) presented an effective feature engineering method to extract and modify features from data and the effects it has on different classifiers. The feature was used to compare six popular ML models for classification. The models compared were LR, RF, DT, K-NN, MLP, and XGBoost. The results showed that the DT model, when applied to the proposed feature engineering, was the best performing, achieving an average accuracy of 98.64%.

Jaiswal et al. (2023) proposed an improved version of the XGBoost ensemble algorithm called I-XGBoost. The study focused on enhancing identification accuracy through three crucial phases: data pre-treatment, feature extraction, and target role. The performance evaluations used the WBCD and compared the results with various classification techniques, including precision, recall, f1-score, and accuracy, as well as ML algorithms such as SVM, LR, K-NN, NB, DT, RF, AdaBoost, and XGBoost. The results indicated that I-XGBoost achieved an impressive accuracy score of 98.24%, while the LR classifier reached an accuracy score of 97%.

Laghmati et al. (2023) presented a supervised ML Computer Aided Design system for breast cancer classification based on feature selection, PCA, grid search for hyperparameter tuning, and cross-validation. The proposed system draws on seven ML classifiers ANN, K-NN, SVM, DT, RF, XGboost, and Adaboost. Two ensemble models were developed by concatenating the prediction of each ML model using majority voting and stacking with LR S-LR for the final prediction. The performance of the system was evaluated by computing various evaluation metrics, mainly accuracy, specificity, precision, recall, Matthews Correlation Coefficient, Jaccard, and F1-score. Wisconsin and Mass mammography datasets were used. The results indicated that the XGboost model achieved the highest recall of over 96% for the Mammographic Mass dataset, while for the WBCD, both the AdaBoost and the S-LR models outperformed the others with a Recall of 95.35%. The stacking with LR ensemble model obtained the highest accuracies of 93.37% for the Mammographic Mass dataset and 97.37% for the WBCD.

Omondiagbe et al. (2019) investigated SVM (using radial basis kernel), ANNs, and NB using WBCD Dataset to integrate these ML techniques with feature selection/feature extraction methods and compared their performances to identify the most suitable approach. The paper proposed a hybrid approach for breast cancer diagnosis by reducing the high dimensionality of features using linear discriminant analysis (LDA) and then applying the new reduced feature dataset to SVM. The proposed approach obtained an accuracy of 98.82%, a sensitivity of 98.41%, a specificity of 99.07%, and an area under the receiver operating characteristic curve of 0.9994.

Chaurasia and Pal (2021c) developed a stack-based ensemble techniques and feature selection methods for the comprehensive performance of the algorithm and comparative analysis of breast cancer datasets with reduced attributes. In the article, the SVM, K-NN, NB and perceptron were the four ML algorithms combined to make the new model, called blending (stacking). Finally, LR was used to predict the stacked model. It was significant that the sub-models produced different results that were not correlated predictions. The stacking technique was best when all the sub-models were skilfully combined. The article used the five-feature selection technique because it affected the model's overall performance.

2.1 Research Gap

While existing studies have extensively explored the effects of different scaling techniques on ML algorithms, there has not been any notable gap in understanding the interaction between hyperparameter tuning and scaling techniques**.** Therefore, in this research, after comparing the different scaling techniques, the various algorithms were hyperparameterized to observe if hyperparameter tuning can mitigate or eliminate the differences in performance caused by different scaling methods, such as MaxAbsScaler and StandardScaler.

3 Methodology

The research design, environment, and dataset are described in this Section. And also, the algorithm and performance metrics are also examined. The Python Jupyter Notebook was used to analyze the datasets in order to determine the best performing classifier of the 5 classifiers against the 4 performance metrics for breast cancer prediction. Jupyter Notebook was used for analysis because Python programming language possesses power and flexibility for building and deploying advanced ML models for breast cancer analysis. It also offers advanced techniques, scalability, integration, deployment and sharing.

3.1 Research Design

This research utilizes quantitative research methodology to conduct a comparative analysis of various ML algorithms, assessing their performance using specific metrics to predict breast cancer mortality.

3.2 Dataset Description

The WBCD dataset used in this study was sourced from the UCI Repository. It consists of clinical and demographic features of breast cancer patients. The dataset comprises features extracted from digitized Fine Needle Aspirate biopsies images. This multivariate dataset consists of 569 instances and 33 features which includes ID number, Diagnosis (M/B), Radius (mean), Texture (mean), Perimeter (mean), Area (mean), Smoothness (mean), Compactness (mean), Concavity (mean), Concave points (mean), Symmetry (mean), Fractal dimension (mean), Radius (standard error), Texture (standard error), Perimeter (standard error), Area (standard error), Smoothness (standard error), Compactness (standard error), Concavity (standard error), Concave points (standard error), Symmetry (standard error), Fractal dimension (standard error), Radius (worst), Texture (worst), Perimeter (worst), Area (worst), Smoothness (worst), Compactness (worst), Concavity (worst), ab. Concave points (worst), ac. Symmetry (worst), ad. Fractal dimension (worst), and it has 0 mismatches and 0 missing values.

Table 1 shows the feature of the dataset.

Table 1: Dataset snippet

ID	Diagnosis	Radius_Mean	Texture_Mean	Perimeter_Mean	Area_Mean	Smoothness_Mean
842302	M	17.99	10.38	122.8	1001	0.1184
842517	M	20.57	17.77	132.9	1326	0.08474
84300903	M	19.69	21.25	130	1203	0.1096
84348301	M	11.42	20.38	77.58	386.1	0.1425
84358402	M	20.29	14.34	135.1	1297	0.1003
843786	M	12.45	15.7	82.57	477.1	0.1278

3.3 Data Preprocessing

The data preprocessing phase involves processing the data to identify and remove any unintended and unnamed columns that might exist due to formatting issues, thereby ensuring that the dataset is clean and structured correctly. It is vital for the subsequent steps in the model-building process. After the dataset is loaded and cleaned, the next step is to convert the target variable, 'diagnosis,' from categorical to numerical values. This conversion is necessary to make the data compatible with ML algorithms. Specifically, the diagnosis labels 'M'(malignant) and 'B' (benign) were mapped to 1 and 0, respectively. Following the conversion, the data was split into features and target variables.

3.4 Data Scaling Method

Data normalization is an activity in data preprocessing that changes the attribute value according to a common scale or range to improve the performance of a ML algorithm. There are different types of techniques for data normalization, but this research is only limited to two types. The StandardScaler and the MaxAbsScaler. Applying standard normalization ensures that for each feature, the mean is 0 and the variance is 1, resulting to all features being on the same scale. However, this normalization does not guarantee obtaining any specific minimum and maximum feature values. The StandardScaler was selected because it standardizes features by removing the mean and scaling to unit variance. This is particularly beneficial for SVM and MLP, which are sensitive to features scales and require normalized input for optimal performance. MaxAbsScaler was chosen because it is useful when working with sparse data or when faster, simpler transformation is needed. It scales each feature by its maximum absolute value and keeps the sign of the data, which preserves sparsity and is more computaionally efficient. NB, RF, and DT are relatively scale invariant, meaning they are not significantly afftected by feature magnitudes. However, scaling was still applied for uniformity and to observe whether indirect effects (eg on feature interaction) would influence performance. The use of both scaling techniques allowed for a comparative analysis of their influence across both scale sensitive and scale insensitive algorithms, providing deeper insight into their practical implications in breast cancer prediction.

To optimize the performance of the classification algorithms, hyperparameter tuning was performed using grid search and cross validation. This approach systematically explores a predefined set of hyperparameter values and evaluate model performance using 5-fold cross validation to ensure robustness and to avoid overfitting. For each algorithm, the best combination of hyperparameters was selected based on the higest cross validation score.

The standardScaler normalization is determined by the formula:

$$z = (x - \mu)/\sigma \tag{1}$$

Where:

i. z is the standardized value
ii. x is the original value of the feature
iii. μ is the mean of the feature in the training set
iv. σ is the standard deviation of the feature in the training set.

The MaxAbsScaler scales the data by setting the maximum absolute value of each feature to 1. It analyzes the training data and finds the absolute maximum value for each feature. The MaxAbsScaler is determined by the formula:

$$X_{Scaled} = X/|X\_\text{max}| \tag{2}$$

where:

i. X_{Scaled} is the scaled value
ii. X is the original value of the feature
iii. $|X\_\text{max}|$ is the absolute value of the maximum value of the feature in the training set.

The framework in Figure 1 shows the workflow from dataset input, through normalization techniques, to output produced.

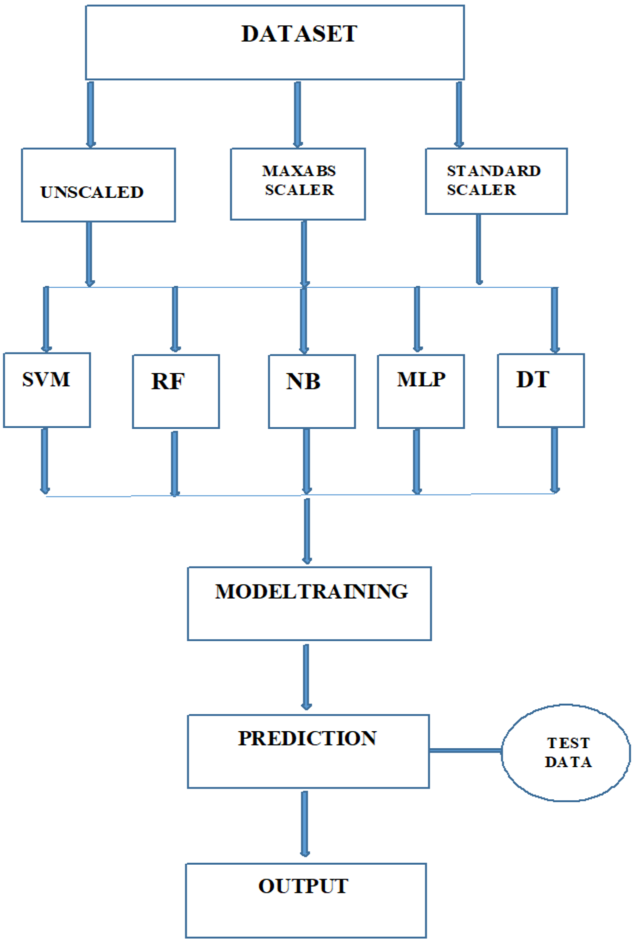


Figure 1: The framework of the proposed Breast Cancer Classification Model.

4 Result and Discussion

This Section discusses and analyzes the values of the performance metrics obtained on each classifier after analyzing it on the different scaling techniques. The analysis carried out in this research investigated the influence of scaling techniques on the performance of the algorithms for breast cancer prediction. MaxAbsScaler and StandardScaler were the two different scaling approaches used. Their impact was compared to the performance of the raw, unscaled data. The results revealed that the different scaling techniques produced slight variations in the prediction performance when compared to the unscaled models. This indicates that scaling has some influence on the algorithms. Interestingly, after hyperparameter tuning, all the performance metrics yielded the same results for both scaled and unscaled predictions. This suggests that the hyperparameter tuning process potentially compensated for the lack of explicit scaling. It can also be noted that after hyperprameter tuning, the scaling techniques became minimal because the models adapt their internal parameters to compensate for the difference in feature scaling. This tuning allows models to maintain strong performance across scaled and unscaled datasets, particularly when regularization, learning rates and kernel parameters are optimally adjusted. The specific algorithms used and the distribution of the data itself also contributed to this observation. It is important to note also that, the sensitivity of different algorithms might be inherently more robust to feature scale variations than others. Also, hyperparameter tuning can sometimes adjust internal model parameters in a way that mimic the effect of scaling and this explains why the performance metrics became similar after tuning.

4.1 Analysis of Results before Hyperparameter Tuning

After conducting a thorough analysis of the models, it became clear that the MaxAbsScaler demonstrated the most optimal performance across all performance metrics in the SVM model. Following closely behind, the StandardScaler also exhibited strong performance in the SVM model. The RF and DT models delivered identical results across all metrics in both the scaled and unscaled techniques. However, it is worth noting that the NB model produced similar results in the scaled technique, while the unscaled technique resulted in higher values across the metrics. For the MLP model, the StandardScaler achieved the highest score across the metrics, with the

MaxAbsScaler closely trailing behind in performance. The analysis revealed that different algorithms might have varying levels of sensitivity to scaling. Some algorithms might be more robust to feature scales than others. Figure 2 shows the performance of the different scaling technique and the unscaled technique across the different classifiers, and the performance of the classifiers were evaluated based on four different metrics.

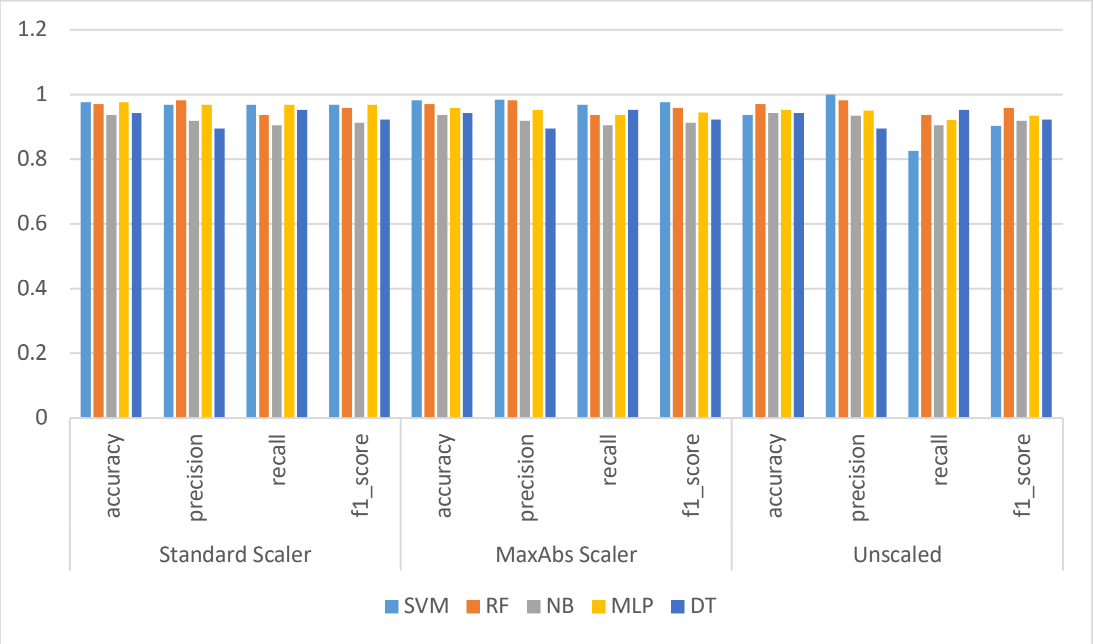


Figure 2: Analysis of the different models before Hyperparameter Tuning

The following Tables show the result of the classifiers after it has been evaluated on the four different metrics parameters. It was observed from the results that the SVM classifier had a value of 1.0 for the precision for the unscaled technique, while the StandardScaler and the MaxAbsScaler had value of 0.968 and 0.984 respectively. The two scaling techniques normalize features, but some of SVM parameters are scale sensitive. The unscaled data is not always at disadvantage when compared with scaled data especially when features are roughly on the same scale, so even though scaling is generally recommended, the result shows that it is not always superior.

Table 2: Result comparison of unscaled algorithms before hyperparameter tuning

Model	Accuracy	Precision	Recall	F1_score
SVM	0.936	1	0.825	0.904
RF	0.971	0.983	0.937	0.959
NB	0.942	0.934	0.905	0.919
MLP	0.953	0.951	0.921	0.935
DT	0.942	0.896	0.952	0.923

Table 3: Result comparison of standard scaled algorithms before hyperparameter tuning

Model	Accuracy	Precision	Recall	F1_score
SVM	0.977	0.968	0.968	0.968
RF	0.971	0.983	0.937	0.959
NB	0.936	0.919	0.905	0.912
MLP	0.977	0.968	0.968	0.968
DT	0.942	0.896	0.952	0.923

Table 4: Result comparison of MaxAbs scaled algorithms before hyperparameter tuning

Model	Accuracy	Precision	Recall	F1_score
SVM	0.982	0.984	0.968	0.976
RF	0.971	0.983	0.937	0.959
NB	0.936	0.919	0.905	0.912
MLP	0.959	0.952	0.937	0.944
DT	0.942	0.896	0.952	0.923

4.2 Analysis of Results after Hyperparameter Tuning

Following hyperparameter tuning, the classification models developed using both scaling techniques (MaxAbsScaler and StandardScaler) and the model trained without any scaling produced virtually identical results. This suggests that the tuning process effectively adjusted internal parameters to account for differences in feature scale. The performance metrics for the classifiers were as follows:

SVM had an Accuracy of 0.9508, Precision value of 0.9429, Recall of 0.9249, and F1-score of 0.9333. SVM demonstrated balanced performance across all metrics, showing its effectiveness in separating classes, even without scaled input, once properly tuned. RF had Accuracy of 0.9578, Precision of 0.9571, Recall value of 0.9297, and F1-score of 0.9425. RF outperformed all other models due to its ensemble approach, robustness to scale variation, and strong generalization capability. NB had 0.9385 for Accuracy, 0.9467 Precision, 0.8870 for Recall, and F1-score of 0.9148. NB achieved high precision but lower recall, indicating a tendency to avoid false positives at the cost of some false negatives. MLP had Accuracy of 0.9314, Precision of 0.9205, Recall of 0.8961, and F1-score of 0.9065. MLP performed well overall, though slightly impacted by scale-sensitive behavior, especially in recall. DT had an Accuracy of 0.9403, Precision of 0.9311, Recall of 0.9111, and F1-score of 0.9184. DT showed reliable and interpretable performance with minimal sensitivity to feature scaling. Overall, the similarity in results confirms that hyperparameter tuning reduced the dependency on scaling techniques. RF being an ensemble of multiple DTs, consistently delivered the highest performance due to its robustness and capacity to reduce overfitting. RF being an ensemble model produced the highest results among all the techniques because of its robustness, scale invariance and ability to generalize well. The classifier further benefitted from combining the strength of the multiple DT.

Figure 3 shows a diagrammatical chart representation of the analysis of the different classifiers after the performance of hyperparameter tuning.

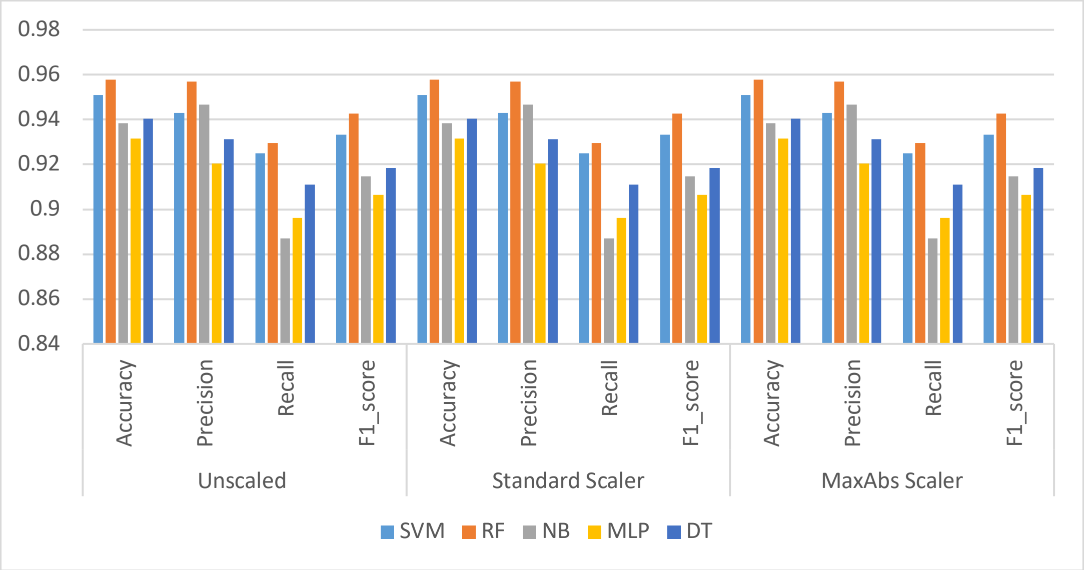


Figure 3: Analysis of the different models after Hyperparameter Tuning

The following Tables show the result of the performance of the different classifiers after hyperparameter tuning, and they are arranged as follows, Unscaled, StandardScaler, and MaxAbsScaler. The results are evaluated based on four metric parameters which are accuracy, precision, recall, and f1_score.

Table 5: Result comparison of unscaled algorithms after hyperparameter tuning

Model	Accuracy	Precision	Recall	F1_score
SVM	0.9508	0.9429	0.9249	0.9333
R	0.9578	0.9571	0.9297	0.9425
NB	0.9385	0.9467	0.8870	0.9148
MLP	0.9314	0.9205	0.8961	0.9065
DT	0.9403	0.9311	0.9111	0.9184

Table 6: Result comparison of standard scaled algorithms after hyperparameter tuning

Model	Accuracy	Precision	Recall	F1_score
SVM	0.9508	0.9429	0.9249	0.9333
RF	0.9578	0.9571	0.9297	0.9425
NB	0.9385	0.9467	0.8870	0.9148
MLP	0.9314	0.9205	0.8961	0.9065
DT	0.9403	0.9311	0.9111	0.9184

Table 7: Result comparison of MaxAbs Scaled Algorithms after hyperparameter tuning

Model	Accuracy	Precision	Recall	F1_score
SVM	0.9508	0.9429	0.9249	0.9333
RF	0.9578	0.9571	0.9297	0.9425
NB	0.9385	0.9467	0.8870	0.9148
MLP	0.9314	0.9205	0.8961	0.9065
DT	0.9403	0.9311	0.9111	0.9184

4.3 Comparative Analysis of Scaling and Classifiers Performance

Table 8 presents a comparative analysis of the impact of various scaling techniques on ML algorithms. While some previous studies reported higher performance metrics, the findings of this study demonstrate that scaling had minimal effect after hyperparameter tuning. This suggests that tuning effectively mitigates scale sensitivity across most models. Notably, the classifiers in this research achieved high yet realistic performance scores, indicating reliable model behavior on real-world data. In contrast, the near-perfect results reported in some studies may point to potential issues such as overfitting, data leakage, or insufficient model validation.

Table 8: Performance comparison of scaling techniques used

References	Techniques Used	Algorithm	Accuracy
Balabaeva et al. (2020)	StandardScaler, MaxAbsScaler, Unscaled, MinMaxScaler, RobustScaler	XG boost	100%
Ahsan et al. (2021)	RobustScaler, and Quantile Transformer	Classification and Regression Tree	100%
Shahriyari et al. (2019)	Normalization (Technique not Specified)	SVM	78%
Proposed method	MaxAbsScaler	SVM before Hyperparameter Tuning	98%
	StandardScaler, MaxAbsScaler, Unscaled	RF after Hyperparameter Tuning	96%

5 Conclusions

The incidence of breast cancer has caused significant concern due to its aggressive nature and its impact on countless patients, particularly women. Fortunately, recent research and studies have focused on early detection of this global malignancy. The emergence of ML has played a significant role in detecting and identifying the disease at an early stage. This has resulted in the development of ML models that when integrated into systems will help clinicians to detect breast cancer and determine its type. This study highlights the importance of evaluating the impact of data preprocessing techniques like scaling. While scaling might not always lead to substantial performance gain, it has a valuable practice to ensure model robustness and interpretability, especially when dealing with future dataset or algorithms with potentially higher sensitivity to feature scaling. In this research, it was discovered that hyperparameter tuning might have the ability to adjust internal model parameters in a way that compensates for the lack of scaling, thereby making different scaling techniques to have little or no influence on predictions made after hyperparameter tuning. The methodical approach of data cleaning, preprocessing, model training, evaluation, and hyperparameter tuning combined, resulted in a robust and accurate predictive tool. Future work could explore the inclusion of additional features and further optimization of the model to enhance its predictive capabilities.

References

Ahsan, M., Mahmud, M., Saha, P., Gupta, K., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. Technologies, 9(3), 52. https://doi.org/10.3390/technologies9030052

Ambarwari, A.; Adrian, Q.J. & Herdiyeni, Y. (2020) Analysis of the Effect of Data Scaling on the Performance of the Machine Learning Algorithm for Plant Identification. *Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi)* 2020, 4, 117–122.

Balabaeva, K. & Kovalchuk, S. (2019). Comparison of Temporal and Non-Temporal Features Effect on Machine Learning Models Quality and Interpretability for Chronic Heart Failure Patients. Procedia Computer Science. 2019, 156, 87–96.

Chaurasiya, S., & Rajak, R. (2022). Comparative analysis of machine learning algorithms in breast cancer classification. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-1772158/v1

Das, A. K., Biswas, S. K., Mandal, A., Bhattacharya, A., & Sanyal, S. (2024). Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP). *Expert Systems with Applications*, 242, 122673. https://doi.org/10.1016/j.eswa.2023.122673

Elsadig, M. A., Altigani, A., & Elshoush, H. T. (2023). Breast cancer detection using machine learning approaches: a comparative study. *International Journal of Electrical and Computer Engineering*, 13(1), 736. https://doi.org/10.11591/ijece.v13i1.pp736-745

Fatima, N., Liu, L., Hong, S., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8, 150360–150376. https://doi.org/10.1109/access.2020.3016715

Fonseca, M. M., Lamb, L. R., Verma, R., Ogunkinle, O., & Seely, J. M. (2019). Breast pain and cancer: should we continue to work-up isolated breast pain? Breast Cancer Research and Treatment, 177(3), 619–627. https://doi.org/10.1007/s10549-019-05354-1

Halim, K. N. A., Jaya, A. S. M., & Fadzil, A. F. A. (2020). Data Pre-Processing Algorithm for Neural Network Binary Classification model in Bank Tele-Marketing. *International Journal of Innovative Technology and Exploring Engineering*, 9(3), 272–277 https://doi.org/10.35940/ijitee.c8472.019320

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., & Cardoso, F. (2019). Breast cancer. Nature Reviews. Disease Primers, 5(1). https://doi.org/10.1038/s41572-019-0111-2

Huang, Y., Zeng, P., & Zhong, C. (2024). Classifying breast cancer subtypes on multi-omics data via sparse canonical correlation analysis and deep learning. *BMC Bioinformatics*, 25(1). https://doi.org/10.1186/s12859-024-05749-y

Jaiswal, V., Suman, P., & Bisen, D. (2023). An improved ensembling techniques for prediction of breast cancer tissues. *Multimedia Tools and Applications*, 83(11), 31975–32000. https://doi.org/10.1007/s11042-023-16949-8

Katsura, C., Ogunmwonyi, I., Kankam, H. K., & Saha, S. (2022). Breast cancer: presentation, investigation and management. *British Journal of Hospital Medicine*, 83(2), 1–7. https://doi.org/10.12968/hmed.2021.0459

Laghmati, S., Hamida, S., Hicham, K., Cherradi, B., & Tmiri, A. (2023). An improved breast cancer disease prediction system using ML and PCA. *Multimedia Tools and Applications*, 83(11), 33785–33821. https://doi.org/10.1007/s11042-023-16874-w

Mahesh, T. R., Kumar, V. V., Vivek, V., Raghunath, K. M. K., & Madhuri, G. S. (2022). Early predictive model for breast cancer classification using blended ensemble learning. *International Journal of System Assurance Engineering and Management*, 15(1), 188–197. https://doi.org/10.1007/s13198-022-01696-0

Omondiagbe, D. A., Veeramani, S., & Sidhu, A. S. (2019a). Machine learning classification techniques for breast cancer diagnosis. IOP Conference Series: Materials Science and Engineering, 495, 012033. https://doi.org/10.1088/1757-899x/495/1/012033

Polyakova, M. V., & Krylov, V. N. (2022). Data normalization methods to improve the quality of classification in the breast cancer diagnostic system. *Applied Aspects of Information Technologies*, 5(1), 55–63. https://doi.org/10.15276/aait.05.2022.5

Shahriyari, L. (2019). Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ datasets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Briefings in Bioinformatics*. 2019, 20, 985–994.

Sharma, A., Goyal, D., & Mohana, R. (2024). An ensemble learning-based framework for breast cancer prediction. *Decision Analytics Journal*, 10, 100372. https://doi.org/10.1016/j.dajour.2023.100372

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Strelcenia, E., & Prakoonwit, S. (2023). Effective feature engineering and Classification of breast cancer diagnosis: a Comparative study. *BioMedInformatics,* 3(3), 616–631. https://doi.org/10.3390/biomedinformatics3030042

World Health Organization. Breast cancer. 2021. https://www.who.int/news-room/fact- sheets/detail/breast-cancer

Yang, P., Wu, W., Wu, C., Shih, Y., Hsieh, C., & Hsu, J. (2021). Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning. Open Medicine, 16(1), 754–768. https://doi.org/10.1515/med-2021-0282

Fuel Consumption Prediction of Vehicles Using Machine Learning Algorithm

[1]Oladejo Olubusayo, [2]Ozoh Patrick, [3*] Ibrahim Musibau, [4]Adigun Adepeju, [5]Oyinloye Olufunke, [6]Dimple Ariyo, [7]Ojo Oluwafolake and [8]Abanikannda Mutahir

[1]Department of Physics, Osun State University, Osogbo, Nigeria
[2,3,4,5,6,7,8]Faculty of Computer Science and Information Technology, Osun State University, Osogbo, Nigeria

email: [3*]kunle_ibrahim2001@yahoo.com

Corresponding author

Received: 15 November 2024 | Accepted: 24 April 2025 | Early access: 19 May 2025

Abstract - *The major objective of this study revolves around accurately estimating the distance a car can travel in kilometres per Litter of fuel consumed. This study develops a precise machine-learning algorithm for predicting vehicle petroleum consumption. This encompasses adapting distinct machine-learning techniques, evaluating their performance, selecting the most optimal model, and validating its real-world applicability. The dataset used for this study includes attributes such as Miles per gallon (Mpg), acceleration, horsepower, displacement, cylinder count, and car model. The implementation strategy entails comprehensive data pre-processing and employing well-established machine learning techniques: Random Forest, Decision Tree, and Linear Regression. The Python programming environment is applied for coding and data manipulation. Model performance assessment uses the Mean Squared Error (MSE) metric. The findings show the performance of the Random Forest algorithm as having the lowest MSE value of 0.008806 among the assessed models. In conclusion, the proficiency of the Random Forest algorithm. in predicting fuel consumption will open avenues for informed decision-making and resource optimization within the automotive sector.*

Keywords: Tracking, estimation, machine learning, model performance, decision making, optimization.

1 Introduction

John McCarthy coined the term machine learning in 1955. He defined it as a branch of engineering. However, it includes experience and quick responses to questions (Gupta et al., 2024). Machine learning can also be defined as that which enables computers to learn (Samuel et al., 2022). This research uses the most effective method for predicting vehicle fuel consumption. It is important to understand the factors that affect fuel consumption and models that can accurately predict fuel consumption (Ashqar et al., 2024). The automobile industry is one of the fastest-growing industries in the world. With the increasing number of vehicles on the road, there is also an increase in fuel consumption. The understanding of factors that affect the models that can accurately predict fuel consumption.

This study consists of five sections. Section one contains the background and contribution to knowledge. Section two includes a review of fundamental concepts and related works. Section three showcases the presentation of the methodology and the techniques, technologies, and tools to be used. Section four presents the results obtained. Section five includes the conclusions and recommendations of this study.

The contributions are

(1) To adapt Random Forest, Decision Tree, and Linear Regression techniques to prediction
(2) Evolve algorithm that accurately predicts fuel consumption in vehicles
(3) Validate the prediction model with the best performance.

2 Literature review

This section provides an in-depth review of the relevant literature about the current study. The review encompasses various aspects, including machine learning techniques in predicting fuel consumption, the methodologies adopted, and an overview of pertinent works in the field. Fuel consumption prediction is vital for optimizing energy use and minimizing environmental impact in transportation and energy systems. It helps us understand consumption patterns, make efficient energy decisions, and develop strategies. Recent progress, driven by data-driven methods and technology, has improved fuel consumption prediction. These predictive models, from single vehicles to entire cities, consider factors like vehicle type, driving conditions, and weather (Su et al., 2023).

Machine learning drives these models, analysing past and real-time data for accurate predictions. This has wide-ranging benefits, reducing costs, and carbon footprints, and enhancing sustainability. The progress in this field also contributes to our understanding of energy dynamics and supports the shift to sustainable energy. By combining machine learning and data analysis, researchers are revolutionizing fuel management for a greener future (Smith, 2022). Chen et al. (2021) introduced machine learning for predicting fuel consumption for spark-ignition engine vehicles based on real-world driving conditions. A set of driving data was collected from an on-board diagnostic (OBD) system, pre-processed data, and the most relevant features using a correlation-based feature selection method. Then, three machine learning algorithms were used to predict fuel consumption: random forest, support vector regression, and multilayer perceptron. The models were evaluated using three metrics: MAE, RMSE, and coefficient of determination. The results showed that the random forest model outperformed the other two models, and the feature importance analysis identified engine load, vehicle speed, and engine coolant temperature as the most critical features for fuel consumption prediction. The authors concluded that their machine learning approach is an effective method for fuel consumption prediction of spark-ignition engine vehicles, which can lead to economic and environmental benefits.

The purpose of Xu et al. (2021) is to develop a general research guideline on structural sciences with the current researchers to have greater knowledge to understand the importance of the basic sciences. Yang et al. (2024) improve the accuracy of model prediction and describe the basis for controlling consumption. Manivannan (2024) applies machine learning to produce a smart energy system for electrical vehicles using optimization methods to reduce costs and time. Katyare et al. (2024) investigate issues in forecasting different levels of systems, with a focus on accuracy. as accurate predictions are important for improving efficiency.

The literature review of the different prediction models is given as follows.

2.1 ANN

Li et al. (2020) demonstrated that machine-learning models predict urban buses. They collected data from a real-world driving cycle and applied four models: decision tree, RF technique, and artificial networks (ANN). They investigated the performance of the accuracy. The three models identified the most critical factors. The study demonstrated that machine learning models can effectively predict urban buses optimize the performance of urban buses, reduce fuel consumption, and minimize carbon emissions.

Tang et al. (2022) developed a prediction model for a heavy-duty truck using artificial neural networks (ANNs). The authors collected data on various driving conditions, including vehicle speed, acceleration, engine load, and other parameters, from an on-board diagnostic (OBD) system installed in the truck. The collected data was pre-processed, and the features were selected using a correlation-based feature selection method. The authors then developed an ANN model for the truck's features. The performance was evaluated using two metrics: mean absolute error (MAE) and coefficient of determination (R^2). The accuracy in predicting fuel consumption has an MAE of 0.3 km/L and R^2 of 0.96. The authors concluded that their ANN-based fuel consumption prediction model can be used to optimize the control strategies of heavy-duty trucks and improve their fuel efficiency, which can have significant economic and environmental benefits.

2.2 Random forest

Wu et al. (2024) state that the random forest method for monitoring malware is an improvement in the application of machine learning. The technique consists of obtaining different sets of datasets. The evaluation for the estimation of accuracy is greater than previous techniques. Noviyanti et al. (2023) provide an increase in the reliability of fast monitoring of diabetes using the Random Forest algorithm model. The study was done with the aid of data collection, data pre-processing, modelling, and evaluation.

2.3 Decision tree

Blockeel et al. (2023) present the importance of decision trees in machine learning. The study investigates changes in the study of decision trees over time. It provides the strong characteristics and shortcomings of decision trees. Costa and Pedreira (2023) present an in-depth of the major progress made in decision tree research, in terms of training data, modeling, and interpretation.

2.4 Linear regression models

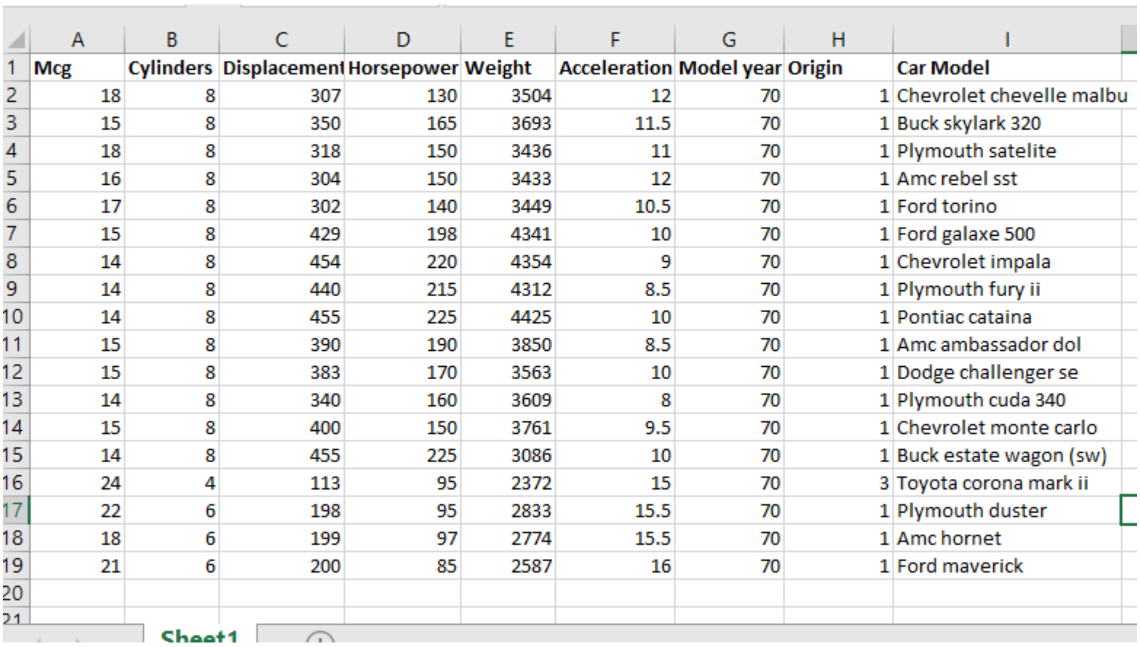
To eradicate imbalanced datasets in regression, a regression method is provided together with the Gaussian technique (Wen et al., 2024). These methods involve the gradient-boosting decision tree (GBDT), random forest (RF), and extreme gradient boosting (XGBoost) models. As a result, the GBDT model possesses the highest accuracy with R2 = 0.95, MAPE = 29.8%, and evaluation results (n = 3214, R2 = 0.84, MAPE = 38.8%). Aziz et al. (2024) present a model to monitor information from the data sets. The proposed technique is useful for reducing crime. The methods involve the use of several regression models that are developed depending on the regression techniques, decision tree regression (DTR), simple linear regression (SLR), and support vector regression (SVR). The study was pre-processed using MySQL and R programming.

3 Methodology

This section discusses the study methods used in predicting fuel consumption using machine learning. It provides details of data collection, data pre-processing, and the models used.

3.1 Data collection

This section outlines the methodology adopted to achieve the research goals and objectives. The initial phase involves data collection, with the dataset sourced from Kaggle, providing comprehensive information about fuel consumption. This data set encompasses attributes such as mpg) acceleration, horsepower, displacement, cylinder count, weight, and car model. The data size used consists of 18 vehicles of different models. The dataset contains 9 columns while the target variable is the MPG (miles per gallon) as the target variable in research work represents the primary focus of modeling efforts. For training purposes, the pre-processed dataset is divided into the experimental vehicle and the vehicle of other models. These are converted into numerical values: 0 for experimental cars and 1 for different vehicle models. This study uses 70% of the data: training, and 30%: testing. The fuel consumption dataset is shown in Figure 1.



	A	B	C	D	E	F	G	H	I
1	Mcg	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model year	Origin	Car Model
2	18	8	307	130	3504	12	70	1	Chevrolet chevelle malbu
3	15	8	350	165	3693	11.5	70	1	Buck skylark 320
4	18	8	318	150	3436	11	70	1	Plymouth satelite
5	16	8	304	150	3433	12	70	1	Amc rebel sst
6	17	8	302	140	3449	10.5	70	1	Ford torino
7	15	8	429	198	4341	10	70	1	Ford galaxe 500
8	14	8	454	220	4354	9	70	1	Chevrolet impala
9	14	8	440	215	4312	8.5	70	1	Plymouth fury ii
10	14	8	455	225	4425	10	70	1	Pontiac cataina
11	15	8	390	190	3850	8.5	70	1	Amc ambassador dol
12	15	8	383	170	3563	10	70	1	Dodge challenger se
13	14	8	340	160	3609	8	70	1	Plymouth cuda 340
14	15	8	400	150	3761	9.5	70	1	Chevrolet monte carlo
15	14	8	455	225	3086	10	70	1	Buck estate wagon (sw)
16	24	4	113	95	2372	15	70	3	Toyota corona mark ii
17	22	6	198	95	2833	15.5	70	1	Plymouth duster
18	18	6	199	97	2774	15.5	70	1	Amc hornet
19	21	6	200	85	2587	16	70	1	Ford maverick
20									
21									

Figure 1: Fuel consumption dataset

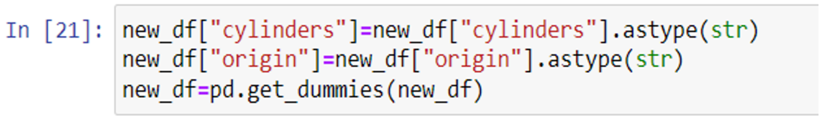
3.2 Data pre-processing

This step involves cleaning and filtering to achieve further analysis. The technique used for the pre-processing of the data is the One Hot encoding technique. One-hot encoding is a technique used in data pre-processing to convert categorical variables into a binary matrix format. It's commonly applied in machine learning and data analysis when dealing with categorical data that cannot be directly used by algorithms that expect numerical input. In one-hot encoding:

1. Each unique category in the categorical variable.

2. For each observation (row) in the dataset, only one element in the binary vector is 1, indicating the presence of that category, while the other elements are 0, indicating absence.

For the above dataset, one hot encoding was used to convert the cylinder attribute from being a categorical variable to Binary format (0s and 1s). One-hot encoding enables machine learning algorithms to effectively work with categorical data, which are often nominal or ordinal. It prevents the algorithm from assuming any ordinal relationship between categories and treats each category as a distinct and independent feature to capture categorical information without introducing unintended biases. The data pre-processing techniques performed on the cylinder attributes are shown in Figure 2.



```
In [21]: new_df["cylinders"]=new_df["cylinders"].astype(str)
         new_df["origin"]=new_df["origin"].astype(str)
         new_df=pd.get_dummies(new_df)
```

Figure 2: A snippet of the encoding technique performed on the cylinder attribute

3.3 Model description

The models used in this study are described in this section. This involves the process of choosing the most appropriate machine learning model architecture it involves selecting a model that is well-suited to the data and provides accurate and reliable predictions.

3.3.1 Random forest

Random Forest is a powerful supervised learning technique for regression and classification tasks. The classifier is an ensemble algorithm that builds multiple decision trees and combines them to produce a more effective classifier. The Random Forest algorithm pseudocode is shown in the following steps.

```
Step 1: Select randomly M features from the feature set

Step 2: For each x in M
        a. Calculate the Information Gain
           Gain (t, x) = E (t) – E (t, x)
           $E(t) = \sum c \ - p\$ \log 2\ p\$$
           $E\ (t, x) = \sum c \in K\ P(c)E(c)$
           Where E(t) is the entropy of the two classes, E(t,x) is the entropy of feature x
        b. Select the node d
        c. Split the node into sub-nodes
        d. Repeat steps a, b, and c to construct the tree until it reaches the minimum number of samples to split.

Step 3: Repeat steps 1 and 2 N for times to build a forest of N trees
```

The structure of the vehicle fuel consumption prediction based on the random forest is given in Figure 3.

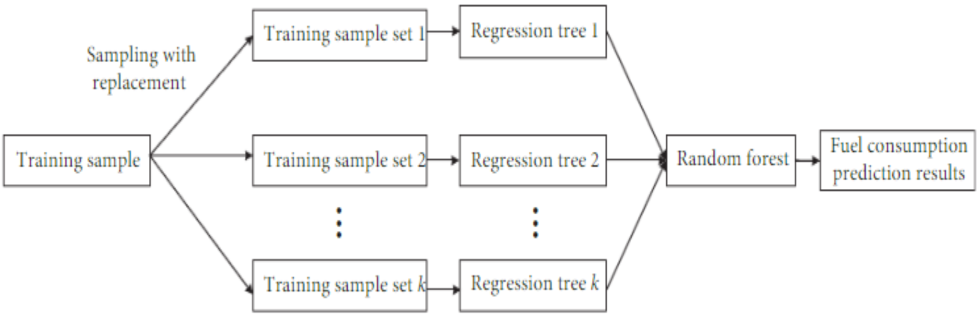


Figure 3: The structure of a vehicle fuel consumption prediction based on the random forest

3.3.2 Linear regression

Linear regression, a form of supervised machine learning, establishes a linear connection between a dependent variable and one or more independent features as shown below:

Given

$y_i \in Y$ $(i = 1, 2, \cdots, n)$ are labels to data (Supervised learning)
$x_i \in X$ $(i = 1, 2, \cdots, n)$ are the input independent training data (univariate – one input variable/parameter)
$\hat{y}_i \in \hat{Y}$ $(i = 1, 2, \cdots, n)$ are the predicted values, with $\hat{Y} = \theta_1 + \theta_2 X$ where $\hat{y_\imath} = \theta_1 + \theta_2 x_i$ and θ_1, θ_2 are the intercept and slope, respectively.

3.3.3 The DT algorithm

The steps for a decision tree algorithm are given as follows.

1. Start.
2. For every iteration, compute the Entropy(H) and Information gain (IG) of this attribute.
3. Select the attribute with the smallest Entropy or Largest Information gain.
4. Get S.
5. The algorithm begins processing.

3.4 Model evaluation

This section delves into the techniques employed to assess the performance of the algorithms utilized for the prediction model. In the context of this project, the evaluation method is the Mean Square Error (MSE). This is given as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (mpg_i - \widehat{mpg_\imath})^2$$

where n represents the total number of instances, mpg_i is the actual "mpg" value for the *i*-th instance and $\widehat{mpg_\imath}$ is the predicted "mpg" value for the *i*-th instance.

4 Results and discussion

This section presents the results of the implementation performed in this study and its evaluation metrics employed to validate the result, and performance of the model. It also provides a detailed discussion of the results and findings of the model.

4.1 Model implementation

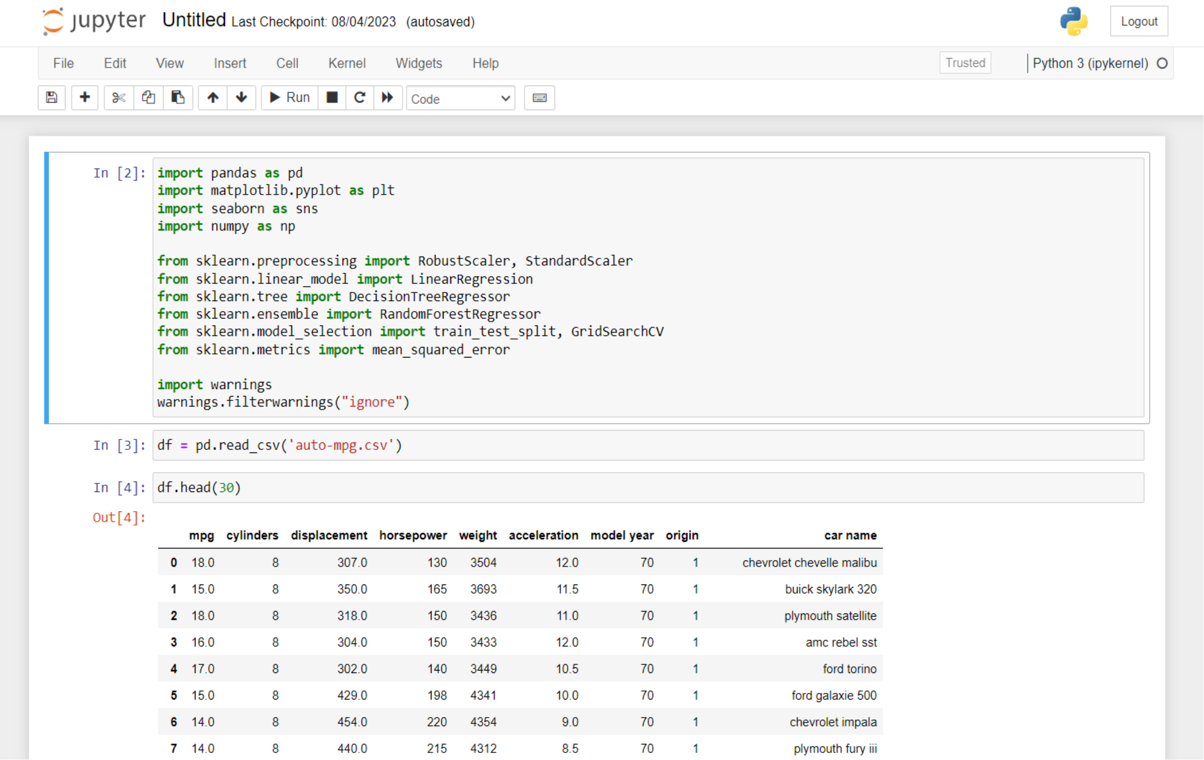
In this study, the model was implemented using hardware and software tools.

4.1.1 Hardware tools

The hardware tool used for this study is a computer system running Windows operating system it serves as the host for the other hardware tools this research requires one computer system.

4.1.2 Implementation of the study

The software used for implementing this study is the Python programming language. Python is good for data mining, analysis, and Machine Learning. The Jupyter Notebook is the version of the Python programming language used. It is an online intelligent computational environment-based software for creating notebook documents. This software allows users to create and share documents with live code, equations, visuals, and narrative text with the open-source Jupyter Notebook program. Figure 4 shows the comparison between the sample data and the corresponding predicted values.

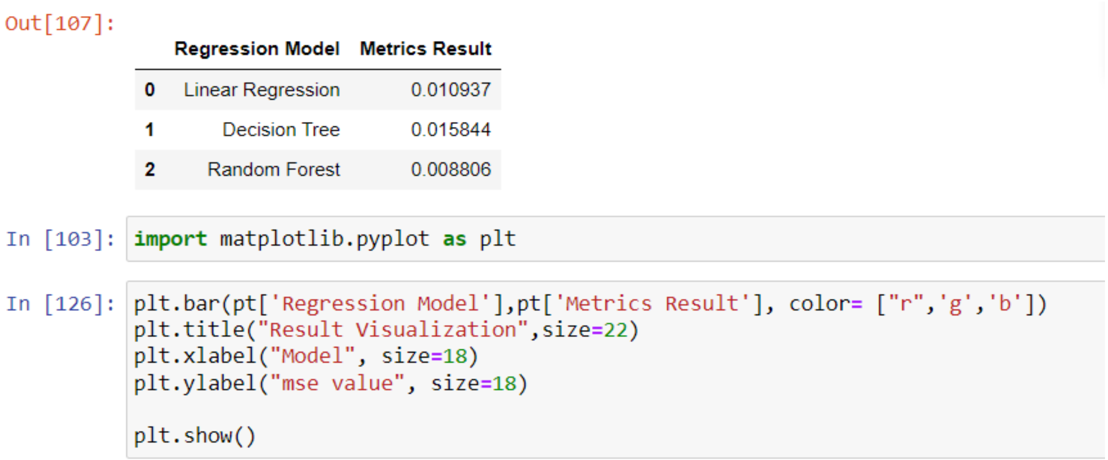


```
In [2]: import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        import numpy as np

        from sklearn.preprocessing import RobustScaler, StandardScaler
        from sklearn.linear_model import LinearRegression
        from sklearn.tree import DecisionTreeRegressor
        from sklearn.ensemble import RandomForestRegressor
        from sklearn.model_selection import train_test_split, GridSearchCV
        from sklearn.metrics import mean_squared_error

        import warnings
        warnings.filterwarnings("ignore")

In [3]: df = pd.read_csv('auto-mpg.csv')

In [4]: df.head(30)

Out[4]:
```

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18.0	8	307.0	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15.0	8	350.0	165	3693	11.5	70	1	buick skylark 320
2	18.0	8	318.0	150	3436	11.0	70	1	plymouth satellite
3	16.0	8	304.0	150	3433	12.0	70	1	amc rebel sst
4	17.0	8	302.0	140	3449	10.5	70	1	ford torino
5	15.0	8	429.0	198	4341	10.0	70	1	ford galaxie 500
6	14.0	8	454.0	220	4354	9.0	70	1	chevrolet impala
7	14.0	8	440.0	215	4312	8.5	70	1	plymouth fury iii

Figure 4: Results using Jupyter Notebook environment

For the three models that were used for this study, Random Forest has the least MSE (mean square error) value. Figure 5 shows the MSE value for the three models. Random Forest has the least (MSE) value 0.008806, Linear Regression: 0.010937, and Decision Tree: 0.015844 respectively, and the lower value indicates a better fit, hence making the Random Forest the most efficient model. Furthermore, during the development of the model, a comparison was made between a sample with the corresponding predicted data. This comparison served as a validation technique to assess the performance of the trained model. The agreement between the predicted and actual data provided insights into the accuracy and reliability of the model's prediction. This is shown in Figure 5. Figure 5 shows the visualizations for linear regression, decision trees, and the Random Forest technique. The figure shows that the Random Forest technique has the lowest matric value when the three methods are compared. This indicates that Random Forest is the most efficient method.



Out[107]:

	Regression Model	Metrics Result
0	Linear Regression	0.010937
1	Decision Tree	0.015844
2	Random Forest	0.008806

```
In [103]: import matplotlib.pyplot as plt
```

```
In [126]: plt.bar(pt['Regression Model'],pt['Metrics Result'], color= ["r",'g','b'])
          plt.title("Result Visualization",size=22)
          plt.xlabel("Model", size=18)
          plt.ylabel("mse value", size=18)

          plt.show()
```

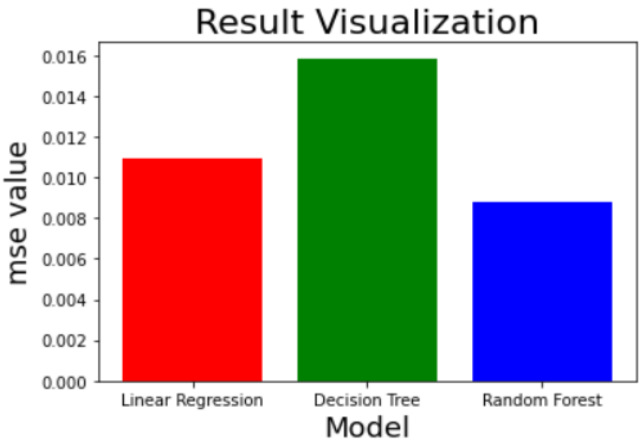


Figure 5: Result visualization of the metric value for the three model

4.2 Evaluation of results

According to Figures 4 and 5, the Random Forest technique is the best-performing technique. To determine its accuracy, it is compared with Eyring et al. (2024) and Asnicar et al. (2024). The three methods were tested on 23 datasets. Table 1 gives the actual data and their respective estimates using Eyring et al. (2024), Asnicar et al. (2024), and the Random Forest technique.

Table 1: Actual data and estimates for techniques

Eyring et al. (2024)		**Asnicar et al. (2024)**		**Random Forest**	
Actual data	**Estimates**	**Actual data**	**Estimates**	**Actual data**	**Estimates**
39.59	46.11	39.59	43.56	39.59	44.18
43.60	48.43	43.60	46.77	43.60	47.91
40.74	44.67	40.74	43.14	40.74	43.94
42.75	48.95	42.75	46.32	42.75	47.01
39.18	46.21	39.18	43.34	39.18	44.22
34.79	40.66	34.79	37.21	34.79	38.54
39.68	45.79	39.68	43.11	39.68	44.32
40.84	44.41	40.84	42.32	40.84	42.99
42.03	45.76	42.03	43.01	42.03	43.87
41.78	46.46	41.78	43.32	41.78	44.58
42.94	46.42	42.94	43.11	42.94	44.36
38.12	43.68	38.12	39.92	38.12	40.93
35.91	40.43	35.91	37.88	35.91	38.03

40.95	45.74	40.95	42.43	40.95	43.46
41.56	47.32	41.56	44.47	41.56	45.84
42.34	45.46	42.34	43.13	42.34	43.79
42.64	46.34	42.64	43.73	42.64	44.74
42.15	46.85	42.15	43.58	42.15	44.36
36.90	40.85	36.90	37.41	36.90	38.58
35.76	39.47	35.76	36.74	35.76	37.74
41.58	45.27	41.58	42.82	41.58	43.43
42.15	46.11	42.15	43.15	42.15	44.85
40.92	44.32	40.92	41.89		

The results of the comparisons are shown in Table 2. Table 2 indicates that RMSE for Eyring et al. (2024), Random Forest, and Asnicar et al. (2024) are 0.873, 0.596, and 0.704, respectively. The MAPE for the Random Forest is 0.921%, which is the smallest compared with Eyring et al. (2024), and Asnicar et al. (2024) having MAPE values of 1.957%, and 1.199%, respectively.

Table 2: Evaluation of techniques

	Eyring et al. (2024)	**Random Forest**	**Asnicar et al. (2024)**
RMSE	0.873	0.596	0.704
MAPE (%)	1.957	0.921	1.199

The output for the RMSE and MAPE of Eyring et al. (2024), Random Forest, and Asnicar et al. (2024) indicate that the Random Forest is the most accurate for estimating data. The RMSE and MAPE values of the Random Forest are the smallest evaluated compared to Eyring et al. (2024) and Asnicar et al. (2024). A critical study of past studies depicted in Section 2 shows that the Random Forest technique is the most widely used machine learning technique. Section 2 is organized to show the strengths and weaknesses of each method.

5 Conclusion

The study aims to develop a robust and dependable model capable of accurately predicting fuel consumption. The models' performance was assessed using the Mean Squared Error (MSE) metric. This metric was chosen as it effectively measures the accuracy of predictions, with lower values indicating better performance. Upon evaluating the models, it was observed that the Random Forest algorithm exhibited the least MSE value (0.008806), signifying superior predictive capabilities. The Linear Regression algorithm had an MSE of 0.010937, while the Decision Tree algorithm had an MSE of 0.015844. This outcome underscored the Random Forest algorithm's dominance in accurately predicting fuel consumption. These findings underscore the Random Forest's potential to revolutionize fuel consumption prediction, a crucial aspect of vehicle management and resource allocation. The model developed here offers a reliable tool for vehicle fuel consumption. The future research for this study will consider having diverse samples to validate the model's robustness and universality.

References

Ashqar, H. I., Obaid, M., Jaber, A., Ashqar, R., Khanfar, N. O., & Elhenawy, M. (2024). Incorporating driving behavior into vehicle fuel consumption prediction: methodology development and testing. *Discover Sustainability*, 5(1), 344.

Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L., & Segata, N. (2024). Machine learning for microbiologists. *Nature Reviews Microbiology*, 22(4), 191-205.

Aziz, R. M., Sharma, P., & Hussain, A. (2024). Machine learning algorithms for crime prediction under Indian penal code. *Annals of data Science*, 11(1), 379-410.

Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). Decision trees: from efficient prediction to responsible AI. *Frontiers in Artificial Intelligence*, 6, 1124553.

Chen, J., Li, K., Zhang, Z., Li, K., & Yu, P. S. (2021). A survey on applications of artificial intelligence in fighting against COVID-19. *ACM Computing Surveys (CSUR),* 54(8), 1-32.

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. Artificial Intelligence Review, 56(5), 4765-4800.

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., & Zanna, L. (2024). Pushing the frontiers in climate modelling and analysis with machine learning. *Nature Climate Change*, 14(9), 916-928.

Gupta, A. K., Pal, G. K., Rajput, K., & Bhatnagar, S. (2024, March). Analysis of Machine Learning Techniques for Fault Detection in 3D Printing. In 2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 1032-1037). IEEE.

Katyare, P., Joshi, S., & Kulkarni, M. (2024). Utilizing Machine Learning Approach to Forecast Fuel Consumption of Backhoe Loader Equipment. *International Journal of Advanced Computer Science & Applications*, 15(5).

Liu, T., Lin, L., Bi, X., Tian, L., Yang, K., Liu, J., & Pan, F. (2019). In situ quantification of interphasial chemistry in Li-ion battery. *Nature nanotechnology*, 14(1), 50-56.

Manivannan, R. (2024). Research on IoT-based hybrid electrical vehicles energy management systems using machine learning-based algorithm. *Sustainable Computing: Informatics and Systems*, 41, 100943.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. AI magazine, 27(4), 12-12.

Noviyanti, C. N., & Alamsyah, A. (2024). Early Detection of Diabetes Using Random Forest Algorithm. *Journal of Information System Exploration and Research*, 2(1).

Samuel, J., Kashyap, R., Samuel, Y., & Pelaez, A. (2022). Adaptive cognitive fit: Artificial intelligence augmented management of information facets and representations. *International journal of information management*, 65, 102505.

Su, M., Su, Z., Cao, S., Park, K. S., & Bae, S. H. (2023). Fuel Consumption Prediction and Optimization Model for Pure Car/Truck Transport Ships. *Journal of Marine Science and Engineering*, 11(6), 1231.

Tang, X., Zhou, H., Wang, F., Wang, W., & Lin, X. (2022). Longevity-conscious energy management strategy of fuel cell hybrid electric Vehicle Based on deep reinforcement learning. Energy, 238, 121593.

Wen, Z., Wang, Q., Ma, Y., Jacinthe, P. A., Liu, G., Li, S., & Song, K. (2024). Remote estimates of suspended particulate matter in global lakes using machine learning models*.* *International Soil and Water Conservation Research*, 12(1), 200-216.

Wu, Y. C., & Chang, Y. L. (2024). Ransomware detection on Linux using machine learning with random forest algorithm. Authorea Preprints.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., & Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4).

Yang, H., Sun, Z., Han, P., & Ma, M. (2024). Data-driven prediction of ship fuel oil consumption based on machine learning models considering meteorological factors. Proceedings of the Institution of Mechanical Engineers, Part M: *Journal of Engineering for the Maritime Environment*, 238(3), 483-502.