Social Media Websites Exploratory Analysis and Prediction

A Mini project report submitted in partial fulfillment of the requirements for the degree of Bachelor of Technology in Computer Science and Engineering with Specialization in Data Analytics

by

Kotagiri Akanksha	**(21BCE7966)**
Gorla Sai Dheraz	**(21BCE8701)**
Karuva Sriman Narayana	**(21BCE8130)**
Egna Siva Kaja	**(21BCE8909)**
Charan	**(21BCE8788)**

Under the guidance of
MEGHAVATHU S. S. NAYAK
Professor

VIT-AP UNIVERSITY

DECLARATION

I hereby declare that the project titled "Social media websites Exploratory Analysis and Prediction" submitted to Vellore Institute of Technology, Amaravti (VIT-AP) for the award of the degree of Bachelor of Technology in Computer Science and Engineering with specialization in Data Analytics is a result of original research carried-out in this thesis. It is further declared that the project report or any part thereof has not been previously submitted to any University or Institute for the award of degree or diploma.

Name of Student(s) : Gorla Sai Dheraz

Hall Ticket Number(s): 21BCE8701

Degree : BTECH - Computer Science and Engineering CORE

Department : SCOPE

Title of the project : Social media websites Exploratory Analysis and Prediction

(Name of the Student)

Date:

ACKNOWLEDGEMENT

We extend our deepest gratitude to all those who have contributed to the completion of this project, "Social Media Websites Exploratory Analysis and Prediction." Our sincere appreciation goes to Prof. Meghavathu S. S. Nayak from VIT-AP University for their invaluable support, guidance, and expertise throughout the research process.

Special thanks to our project supervisor/advisor Prof. Meghavathu S. S. Nayak for their continuous encouragement, insightful feedback, and unwavering assistance, which have been instrumental in shaping the direction and quality of this work.

We also wish to express our gratitude to the participants and individuals who generously shared their time and insights, without whom this project would not have been possible.

Furthermore, we acknowledge the resources and facilities provided by VIT-AP University, which have facilitated our research endeavours.

Lastly, we would like to express our gratitude to our families, friends, and loved ones for their patience, understanding, and encouragement during the course of this project.

Thank you all for your indispensable contributions.

ABSTRACT

In the digital age, social media platforms have become pivotal arenas for communication, networking, and information dissemination. With the exponential growth of user-generated content, there arises a pressing need to understand and predict user behaviours and preferences. This project aims to conduct an exploratory analysis and prediction task using a diverse dataset encompassing demographic, behavioural, and socio-economic attributes extracted from various social media websites.

The dataset comprises crucial features such as age, gender, time spent on the platform, interests, location, demographics, profession, income level, indebtedness status, homeownership, and car ownership. The primary objective is to predict the indebtedness status of social media users based on their profile characteristics and behavioural patterns.

To achieve this, a multi-faceted approach employing machine learning algorithms is adopted. Initially, exploratory data analysis (EDA) techniques are employed to uncover insights into the distribution, correlations, and patterns present in the dataset. Descriptive statistics, data visualization, and correlation analyses provide valuable insights into the relationships between different attributes and the indebtedness status of users.

Subsequently, predictive modelling tasks are undertaken using a variety of algorithms including linear regression, logistic regression, k-nearest neighbours (KNN), Naive Bayes, and decision trees. Linear regression models are utilized to predict the continuous nature of the indebtedness variable, providing insights into the factors influencing the level of indebtedness among social media users.

Logistic regression models are employed for binary classification, categorizing users into indebted and non-indebted categories based on their profile attributes. Additionally, non-parametric algorithms such as KNN are utilized to leverage similarities between users for personalized recommendation systems and clustering analyses, contributing to a deeper understanding of user segmentation based on indebtedness status.

The Naive Bayes approach is adopted for sentiment analysis and topic classification tasks, enabling the identification of prevailing themes and sentiment trends that may influence indebtedness behaviour. Decision tree-based models offer interpretable insights into feature importance, facilitating the identification of key demographic and behavioural predictors of indebtedness among social media users.

The performance of each model is evaluated using appropriate evaluation metrics such as accuracy. By leveraging machine learning techniques to analyse social media data, stakeholders can tailor targeted interventions, financial products, and educational campaigns to address the diverse needs of different user segments.

Moreover, the methodology outlined in this project serves as a valuable blueprint for future research endeavours aimed at leveraging social media data for predictive analytics in diverse domains, including finance, marketing, and public policy. By harnessing the power of machine learning algorithms, researchers can unlock actionable insights from the vast troves of data generated on social media platforms, driving innovation and informed decision-making in the digital era.

Moreover, sentiment analysis can gauge the overall mood or perception surrounding financial topics, helping financial institutions and policymakers gauge public sentiment and tailor communication strategies accordingly. Additionally, anomaly detection techniques can flag unusual activities or patterns in user behaviour, potentially indicating instances of fraud or emerging market trends.

Overall, this project underscores the transformative potential of machine learning in leveraging social media data for actionable insights, paving the way for innovative solutions to complex societal challenges and enhancing decision-making processes across various sectors.

Keywords:

Social media, User behaviour, Predictive analytics, Machine learning, Exploratory data analysis, Demographic attributes, Behavioural patterns, Indebtedness status, Predictive modelling, Linear regression, Logistic regression, K-nearest neighbours (KNN), Naive Bayes, Decision trees, Sentiment analysis, Topic classification, Financial topics, Anomaly detection, Fraud detection, Actionable insights

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

EDA	Exploratory Data Analysis
KNN	K-nearest Neighbors
MAE	Mean Absolute Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error

CHAPTER 1: INTRODUCTION

1.1 Introduction:

In the digital age, social media platforms have emerged as pivotal arenas for communication, networking, and information dissemination. With the exponential growth of user-generated content, there arises a pressing need to understand and predict user behaviors and preferences. This project aims to conduct an exploratory analysis and prediction task using a diverse dataset encompassing demographic, behavioral, and socio-economic attributes extracted from various social media websites.

The dataset comprises crucial features such as age, gender, time spent on the platform, interests, location, demographics, profession, income level, indebtedness status, homeownership, and car ownership. The primary objective is to predict the indebtedness status of social media users based on their profile characteristics and behavioral patterns.

To achieve this, a multi-faceted approach employing machine learning algorithms is adopted. Initially, exploratory data analysis (EDA) techniques are employed to uncover insights into the distribution, correlations, and patterns present in the dataset. Descriptive statistics, data visualization, and correlation analyses provide valuable insights into the relationships between different attributes and the indebtedness status of users.

Subsequently, predictive modeling tasks are undertaken using a variety of algorithms including linear regression, logistic regression, k-nearest neighbors (KNN), Naive Bayes, and decision trees. Linear regression models are utilized to predict the continuous nature of the indebtedness variable, providing insights into the factors influencing the level of indebtedness among social media users.

Logistic regression models are employed for binary classification, categorizing users into indebted and non-indebted categories based on their profile attributes. Additionally, non-parametric algorithms such as KNN are utilized to leverage similarities between users for personalized recommendation systems and clustering analyses, contributing to a deeper understanding of user segmentation based on indebtedness status.

The Naive Bayes approach is adopted for sentiment analysis and topic classification tasks, enabling the identification of prevailing themes and sentiment trends that may influence indebtedness behavior. Decision tree-based models offer interpretable insights into feature importance, facilitating the identification of key demographic and behavioral predictors of indebtedness among social media users.

The performance of each model is evaluated using appropriate evaluation metrics such as accuracy. By leveraging machine learning techniques to analyze social media data, stakeholders can tailor targeted interventions, financial products, and educational campaigns to address the diverse needs of different user segments.

Moreover, the methodology outlined in this project serves as a valuable blueprint for future research endeavors aimed at leveraging social media data for predictive analytics in diverse domains, including finance, marketing, and public policy. By harnessing the power of machine learning algorithms, researchers can unlock actionable insights from the vast troves of data generated on social media platforms, driving innovation and informed decision-making in the digital era.

1.2 Problem statement:

Social media platforms have become integral parts of modern society, offering users avenues for communication, connection, and information sharing. However, with the vast amount of data generated on these platforms, understanding, and predicting user behaviors and preferences has become increasingly challenging. One pressing issue is the lack of insight into users' indebtedness status and the factors influencing it. Without a clear understanding of this aspect, stakeholders such as financial institutions, policymakers, and marketers struggle to tailor targeted interventions and products to meet the diverse needs of social media users.

To address this gap, this project aims to conduct an exploratory analysis and prediction task using data extracted from various social media websites. The primary objective is to predict the indebtedness status of users based on their profile characteristics and behavioral patterns. By uncovering insights into the relationship between demographic, behavioral, and socio-economic attributes and user indebtedness status, this project seeks to provide actionable information that can inform decision-making processes across various sectors.

Through the application of machine learning algorithms, this project seeks to offer a systematic approach to understanding and predicting user behaviors related to indebtedness on social media platforms. By leveraging the diverse dataset encompassing key features such as age, gender, interests, income level, and indebtedness status, the project aims to develop predictive models that can classify users into indebted and non-indebted categories. Ultimately, the findings of this project have the potential to inform targeted interventions, financial products, and educational campaigns, thereby addressing the challenges posed by user indebtedness in the digital age.

1.3 Scope of research

This project focuses on analyzing and predicting user behaviors related to indebtedness using data from social media platforms. The scope encompasses the exploration of various demographic, behavioral, and socio-economic attributes of users, including age, gender, interests, income level, and indebtedness status. By examining these factors, the project aims to uncover patterns and relationships that may influence user indebtedness, providing insights that can inform decision-making processes across different sectors.

The research will utilize machine learning algorithms to develop predictive models that classify users into indebted and non-indebted categories based on their profile characteristics and behavioral patterns. Additionally, the project will explore the performance of different algorithms such as linear regression, logistic regression, k-nearest neighbors (KNN), Naive Bayes, and decision trees in predicting user indebtedness. This comprehensive approach will enable a thorough analysis of the data and facilitate the identification of key predictors of indebtedness among social media users.

Furthermore, the scope of the research extends to evaluating the practical implications of the findings. By assessing the performance of the predictive models and analyzing the insights gained from the data, the project aims to provide actionable information that can be used to tailor targeted interventions, financial products, and educational campaigns aimed at addressing user indebtedness in the digital age. With the goal of informing strategies to mitigate the challenges posed by indebtedness in the modern era.

1.4 Research hypothesis

The research hypothesis for this project posits that there are identifiable patterns and relationships between user demographics, behaviors, and socio-economic attributes on social media platforms and their indebtedness status. It is hypothesized that certain demographic factors such as age and gender may influence the level of indebtedness among users, with younger individuals and certain gender groups potentially exhibiting higher levels of indebtedness. Additionally, it is expected that user behaviors such as time spent on the platform and interests may also play a role in determining indebtedness status, with users exhibiting certain behavioral patterns being more likely to be indebted.

Furthermore, the research hypothesis suggests that socio-economic factors such as income level and homeownership may serve as key predictors of user indebtedness on social media platforms. It is anticipated that users with lower income levels and those who do not own homes may be more susceptible to indebtedness, as financial stability and assets ownership are often correlated with lower levels of indebtedness. By examining these factors in conjunction with user demographics and behaviors, the research aims to uncover actionable insights into the determinants of user indebtedness and inform targeted interventions and strategies to address this issue.

Overall, the research hypothesis asserts that by leveraging machine learning algorithms to analyze data from social media platforms, it is possible to develop predictive models that accurately classify users into indebted and non-indebted categories based on their profile characteristics and behavioral patterns. Through a systematic exploration of user demographics, behaviors, and socio-economic attributes, the research seeks to validate the hypothesis and contribute to a deeper understanding of user indebtedness in the digital age, ultimately informing strategies to mitigate its impact on individuals and society.

1.5 Objectives

The main objective of this research project is to analyze and predict user behaviors related to indebtedness using data collected from various social media platforms. The primary aim is to understand the factors that influence user indebtedness status, including demographic, behavioral, and socio-economic attributes. By examining these factors, the research seeks to uncover patterns and relationships that can inform decision-making processes and interventions aimed at addressing user indebtedness.

Specifically, the research aims to develop predictive models that classify social media users into indebted and non-indebted categories based on their profile characteristics and behavioral patterns. This involves utilizing machine learning algorithms such as linear regression, logistic regression, k-nearest neighbors (KNN), Naive Bayes, and decision trees to analyze the data and identify key predictors of indebtedness. By accurately predicting user indebtedness status, the research aims to provide valuable insights that can guide the development of targeted interventions and strategies to mitigate the impact of indebtedness.

Furthermore, the research objective extends to evaluating the practical implications of the findings and assessing the performance of the predictive models. By examining the effectiveness of different algorithms and analyzing the insights gained from the data, the research aims to provide actionable information that can be used to tailor financial products, educational campaigns, and other interventions aimed at addressing user indebtedness in the digital age. Overall, the objective of this research project is to contribute to a deeper understanding of user behaviors related to indebtedness on social media platforms and to inform strategies to mitigate its impact on individuals and society.

1.6 Organization of the report

The report begins with an introduction that provides an overview of the research project, highlighting the significance of analyzing and predicting user behaviors related to indebtedness on social media platforms. It outlines the objectives of the research and introduces the key concepts and methods employed in the study. The introduction sets the stage for the rest of the report by establishing the context and rationale for the research.

Following the introduction, the report delves into the methodology section, which outlines the approach and techniques used in the research. This section details the data collection process, including the sources of data and the methods used to extract and preprocess the data. It also describes the exploratory data analysis (EDA) techniques employed to uncover insights into the dataset and the machine learning algorithms used for predictive modeling. Additionally, the methodology section discusses the evaluation metrics used to assess the performance of the predictive models and the overall validity of the research findings.

Next, the report presents the results of the analysis conducted in the research project. This section provides a comprehensive overview of the findings, including insights gained from the exploratory data analysis and the performance of the predictive models. It highlights key patterns and relationships identified in the data and discusses the implications of these findings for understanding user behaviors related to indebtedness on social media platforms. The results section provides a clear and concise summary of the research outcomes, supporting them with relevant data and visualizations.

Finally, the report concludes with a discussion and conclusion section that synthesizes the key findings of the research and discusses their implications for theory, practice, and future research. This section reflects on the research objectives and evaluates the extent to which they were achieved. It also discusses limitations of the study and suggests areas for further exploration. The discussion and conclusion section provides a cohesive summary of the research project, offering insights into the broader implications of the findings and suggesting avenues for future research in this area.

CHAPTER 2: LITERATURE REVIEW

2.1 Background

Social media platforms like Facebook, Instagram, and Twitter have become central to how people communicate, connect, and share information in today's digital age. These platforms have grown rapidly, attracting billions of users worldwide. They play a crucial role in shaping our social interactions and influencing various aspects of our lives, including how we make decisions about our finances.

User indebtedness, or the state of owing money, is a significant issue that affects many individuals and households globally. It can arise from various factors, such as borrowing for education, housing, or consumption, and can have long-term consequences for financial well-being. Understanding the factors that contribute to user indebtedness is essential for addressing this issue effectively.

Research on the relationship between social media use and financial behaviors has gained attention in recent years. Scholars have explored how social media platforms may influence consumer attitudes, spending habits, and financial decision-making processes. Some studies suggest that exposure to certain types of content or social comparisons on social media may impact individuals' perceptions of wealth and financial priorities.

However, the existing literature on this topic is diverse and sometimes contradictory. While some studies suggest a positive relationship between social media use and financial well-being, others find negative associations, such as increased spending or financial stress. Additionally, research in this area often lacks consistency in methodology and theoretical frameworks, making it challenging to draw definitive conclusions.

Despite these challenges, understanding the role of social media in shaping financial behaviors is crucial for developing effective interventions and policies to promote financial literacy and well-being. By synthesizing existing research and identifying gaps in the literature, this study aims to contribute to a deeper understanding of how social media use may influence user behaviors related to indebtedness.

2.2 Summary of literature review and research gap

The literature review has provided a comprehensive overview of existing research on the relationship between social media use and user behaviors related to indebtedness. It has highlighted the multifaceted nature of this relationship, with studies suggesting both positive and negative associations between social media use and financial behaviors. While some research indicates that social media exposure may lead to increased spending and financial stress, other studies suggest that it can also facilitate financial education and promote responsible financial behaviors.

Despite the breadth of research in this area, several gaps and inconsistencies in the literature have been identified. One significant research gap pertains to the lack of consensus regarding the direction and magnitude of the relationship between social media use and user indebtedness. While some studies suggest a direct link between social media exposure and financial distress, others find no significant association or even positive effects, such as increased financial literacy.

Furthermore, the literature review has revealed methodological limitations and theoretical ambiguities in existing research. Many studies rely on cross-sectional data or self-reported measures of social media use and financial behaviors, which may introduce biases and inaccuracies. Additionally, the theoretical frameworks used to understand the relationship between social media use and financial behaviors vary widely, making it challenging to compare findings across studies and draw robust conclusions.

Another notable research gap identified in the literature is the lack of attention to the role of individual differences and contextual factors in shaping the relationship between social media use and user indebtedness. Factors such as age, income level, education, and cultural background may moderate the effects of social media exposure on financial behaviors, yet few studies have systematically examined these factors in depth.

Overall, the literature review underscores the need for further research that addresses these gaps and inconsistencies in the literature. By employing rigorous methodologies, integrating diverse theoretical perspectives, and considering individual differences and contextual factors, future research can provide a more nuanced understanding of how social media use influences user behaviors related to indebtedness. This study aims to contribute to this growing body of literature by addressing these research gaps and advancing our knowledge of the complex interplay between social media use and financial behaviors.

CHAPTER 3: METHODOLOGY

3.1 Materials

Data Collection:

Dataset Link: Kaggle

The primary material utilized in this research project is a dataset sourced from Kaggle, a prominent platform for sharing datasets and data science projects. The dataset contains comprehensive information on various demographic, behavioral, and socio-economic attributes of social media users. Each column in the dataset provides specific details about the user's characteristics and behaviors.

The dataset includes the following columns:
The description for each column is as follows:

- **age:** The age of the user.
- **gender:** The gender identity of the user (Male, Female, Non-binary).
- **time_spent:** The average time spent by users on social media
- **demographics:** The type of area the user resides in (Urban, Suburban, Rural).
- **interests:** The user's primary area of interest or hobby.
- **device_type:** The type of device used by the user (Mobile).
- **location:**The country of residence for the user.
- **platform:** The social media platform where the user spends time.
- **profession:** The user's occupation or professional status.
- **income:** The yearly income of the user.
- **indebt:** Indicates whether the user is in debt (True or False).
- **homeowner:** Indicates whether the user owns a home (True or False).
- **owns_cars:** Indicates whether the user owns cars (True or False).

These columns collectively provide a rich and diverse set of data that enables the analysis and exploration of user behaviors related to indebtedness on social media platforms. The dataset serves as a valuable resource for investigating patterns, relationships, and trends among social media users, contributing to a deeper understanding of the factors influencing user indebtedness in the digital age.

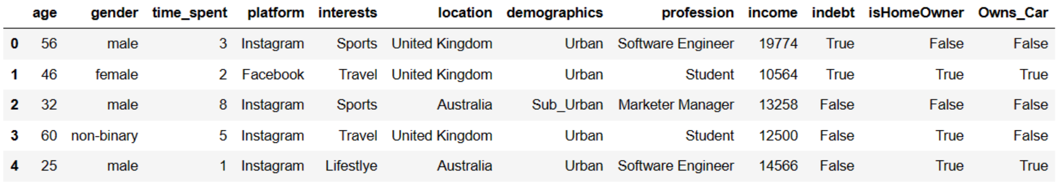
3.2 Summary of methodology

Exploratory Analysis:

Data Pre-processing:

Data Cleaning: During the data cleaning process, one of the columns, "device_type," was identified for deletion from the dataset. There are no null values in the dataset. So, further data cleaning is not needed.

Packages:
Pandas: Data manipulation and analysis library in Python.
NumPy: Fundamental library for numerical computing.
Seaborn: Statistical data visualization library.
Matplotlib.pyplot: Plotting module for creating visualizations.
Scikit-learn: Machine learning library for Python.
ProfileReport: Imports the ProfileReport class from the pandas_profiling library for generating data profile reports.
rcParams: Matplotlib's parameter settings for customizing plot appearance.
rainbow: A Matplotlib colormap representing a continuous spectrum of colors.



	age	gender	time_spent	platform	interests	location	demographics	profession	income	indebt	isHomeOwner	Owns_Car
0	56	male	3	Instagram	Sports	United Kingdom	Urban	Software Engineer	19774	True	False	False
1	46	female	2	Facebook	Travel	United Kingdom	Urban	Student	10564	True	True	True
2	32	male	8	Instagram	Sports	Australia	Sub_Urban	Marketer Manager	13258	False	False	False
3	60	non-binary	5	Instagram	Travel	United Kingdom	Urban	Student	12500	False	True	False
4	25	male	1	Instagram	Lifestlye	Australia	Urban	Software Engineer	14566	False	True	True

Table 3.1 Overview of Dataset

Changing Data types: Converting several columns initially categorized as objects to strings, enhancing the efficiency of data processing.

<u>Code used:</u>

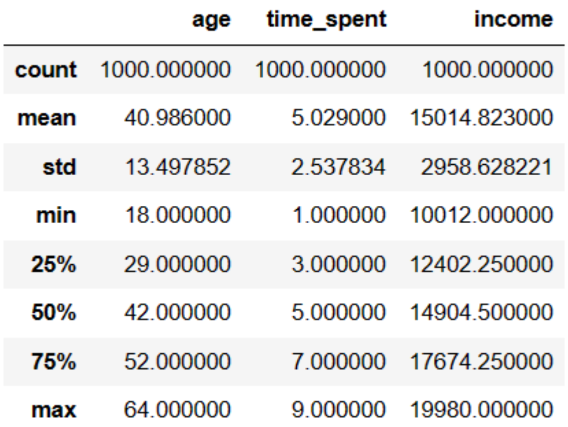
```
data['gender'] = data['gender'].astype("string")
data['platform'] = data['platform'].astype("string")
data['interests'] = data['interests'].astype("string")
data['location'] = data['location'].astype("string")
data['demographics'] = data['demographics'].astype("string")
data['profession'] = data['profession'].astype("string")
```

Data Manipulation: Profession column we are having values as Marketer Manager, Software Engineer, and Student. This is modified to IT professional and Student.
Adding a Value_score column by using income, age, and time_spent to determine individual users value score of time spent on social media

<u>Code used:</u>

```
data['profession'] = data['profession'].replace({'Marketer Manager': 'IT professional','Software
Engineer':'IT professional'})
data.rename(columns={'profession':'working_status'},inplace=True)
data['Value_Score'] = data['income'] / (data['age'] * data['time_spent'])
```

Central tendency, dispersion, and shape of a dataset's distribution:



	age	time_spent	income
count	1000.000000	1000.000000	1000.000000
mean	40.986000	5.029000	15014.823000
std	13.497852	2.537834	2958.628221
min	18.000000	1.000000	10012.000000
25%	29.000000	3.000000	12402.250000
50%	42.000000	5.000000	14904.500000
75%	52.000000	7.000000	17674.250000
max	64.000000	9.000000	19980.000000

Table 3.2.1 Summary of Dataset

Data Visualization:

1. Sum of Time Spent and Value Score by Platform and by Interests Using Stacked Bar Chart:

Code Used:

```
grouped_data_platform = data.groupby('platform').sum()
result_platform = grouped_data_platform[['time_spent', 'Value_Score']]
grouped_data_interests = data.groupby('interests').sum()
result_interests = grouped_data_interests[['time_spent', 'Value_Score']]
print(result_platform)
print(result_interests)
platform_colors = ['skyblue', 'pink']
interests_colors = ['lightcoral', 'lightgreen']
fig, axes = plt.subplots(nrows=2, ncols=1, figsize=(10, 8))
grouped_data_platform[['time_spent', 'Value_Score']].plot(kind='bar', stacked=True, ax=axes[0],
rot=0,color=platform_colors)
axes[0].set_title('Sum of Time Spent and Value Score by Platform')
axes[0].set_ylabel('Sum')
grouped_data_interests[['time_spent', 'Value_Score']].plot(kind='bar',stacked=True, ax=axes[1],
rot=0,color=interests_colors)
axes[1].set_title('Sum of Time Spent and Value Score by Interests')
axes[1].set_ylabel('Sum')
plt.tight_layout()
plt.show()
```

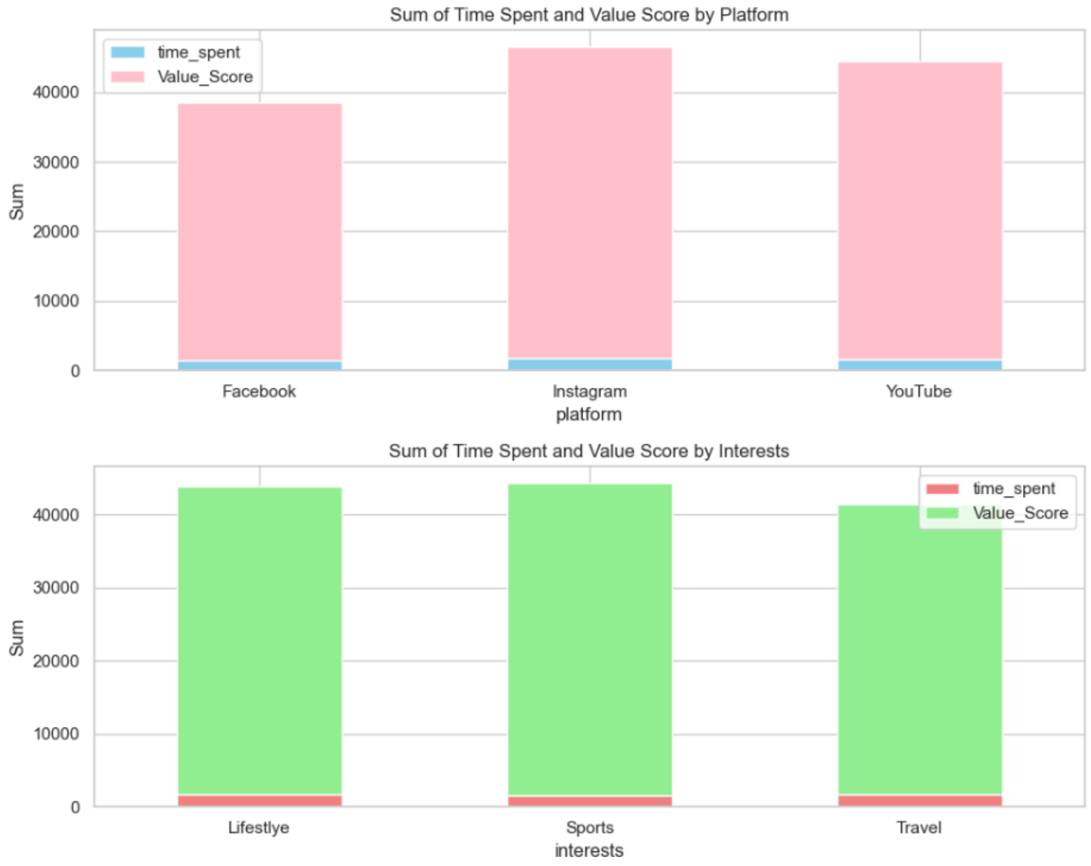


Figure 3.2.1 Stacked bar Chart Representing Sum of Time Spent and Value Score by Platform and Interests

2. Sum of Time Spent, Has House and Owns a Car by Demographics and by Location Using Stacked Bar Chart:

Code used:

```
grouped_data_demograph = data.groupby('demographics').sum()
grouped_data_location = data.groupby('location').sum()
result_Finance = grouped_data_demograph[['time_spent', 'isHomeOwner', 'Owns_Car']]
result_place = grouped_data_location[['time_spent', 'isHomeOwner', 'Owns_Car']]
print(result_Finance)
print(result_place)
platform_colors = ['lightblue', 'pink', 'yellow']
interests_colors = ['lightcoral', 'lightgreen', 'magenta']
fig, axes = plt.subplots(nrows=2, ncols=1, figsize=(10, 8))
result_Finance.plot(kind='bar', stacked=True, ax=axes[0], rot=0, color=platform_colors)
axes[0].set_title('Sum of Time Spent, Has House and Owns a car by Demographics')
axes[0].set_ylabel('Sum')
result_place.plot(kind='bar', stacked=True, ax=axes[1], rot=0, color=interests_colors)
axes[1].set_title('Sum of Time Spent, Has House and Owns a car by Location')
axes[1].set_ylabel('Sum')
plt.tight_layout()
plt.show()
```

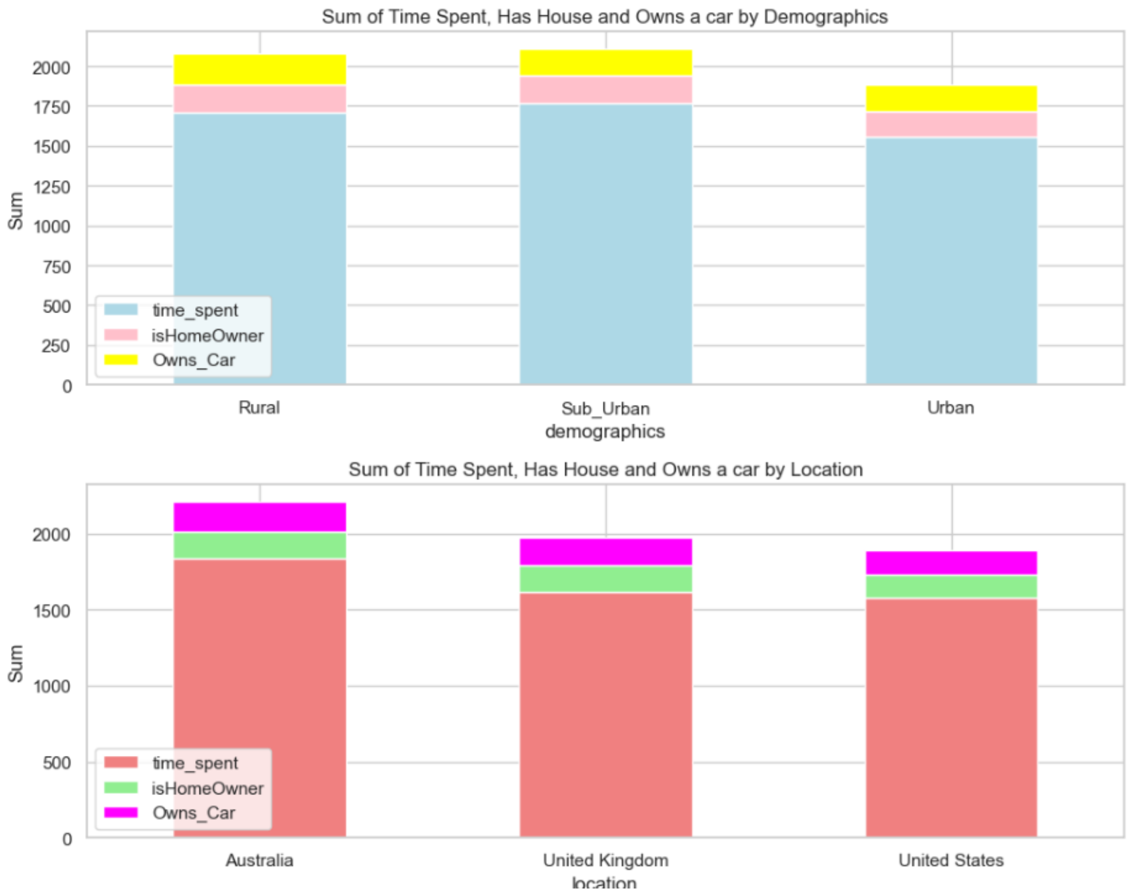


Figure 3.2.2 Stacked bar Chart Representing Sum of Time Spent, Has House and Owns a car by Demographics and Location

Normalization:

Data normalization is a preprocessing technique used to rescale the values of numerical features to a common scale without distorting differences in the ranges of values. It ensures that all features have a similar influence on the analysis, preventing certain features from dominating due to their larger scales. Common normalization techniques include Min-Max scaling and Z-score normalization.

<u>Code used:</u>

```
Gender = {'male': 1, 'female': 2, 'non-binary': 3}
data['gender'] = data['gender'].map(Gender)
Platform = {'Instagram': 1, 'Facebook': 2, 'YouTube': 3}
data['platform'] = data['platform'].map(Platform)
Interest = {'Sports': 1, 'Travel': 2, 'Lifestlye': 3}
data['interests'] = data['interests'].map(Interest)
loc={'United Kingdom':1,'Australia':2,'United States':3}
data['location']=data['location'].map(loc)
dem = {'Urban': 1, 'Sub_Urban': 2, 'Rural': 3}
data['demographics'] = data['demographics'].map(dem)
working = {'IT professional': 1, 'Student': 2}
data['working_status'] = data['working_status'].map(working)
data['indebt'] = data['indebt'].astype(int)
data['isHomeOwner'] = data['isHomeOwner'].astype(int)
data['Owns_Car'] = data['Owns_Car'].astype(int)
```

Data Visualization:

1. Bar Chart of Average time spent by Gender

Code used:

```
gender_time_mean = datadup.groupby('gender')['time_spent'].mean()
plt.figure(figsize=(8, 6))
gender_time_mean.plot(kind='bar', color='skyblue')
plt.title('Average Time Spent by Gender')
plt.xlabel('Gender')
plt.ylabel('Average Time Spent')
plt.xticks(rotation=0)
plt.tight_layout()
plt.show()
```

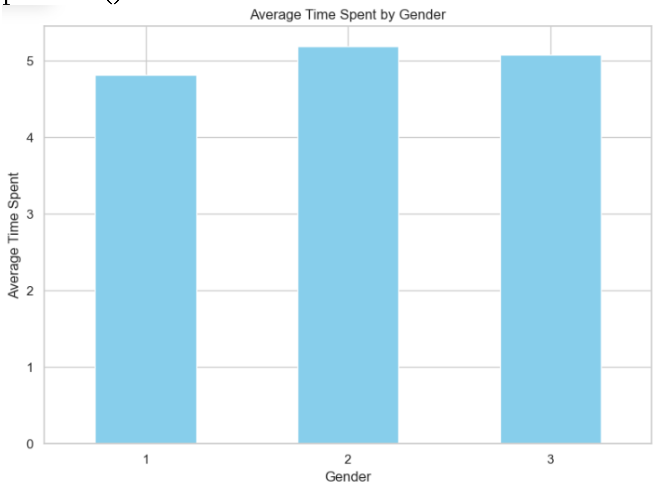


Figure 3.2.3 Bar Chart of Average time spent by Gender

2. Bar Plot of count of time spent by an individual in different platforms:

Code used:

```
data1=data
sns.countplot(x="time_spent",data=data1,hue='platform')
```

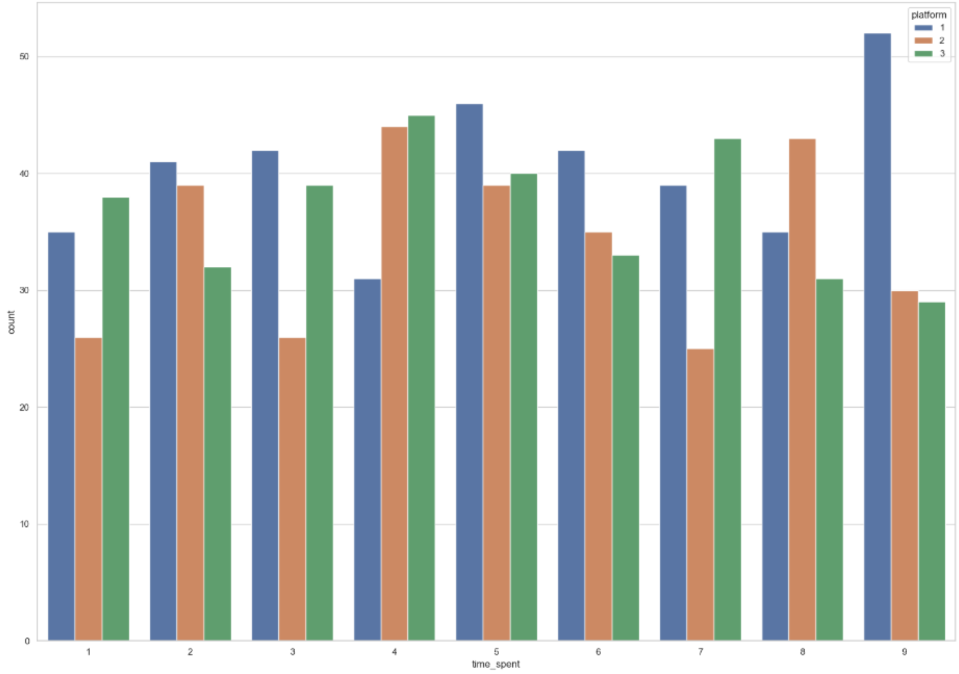


Figure 3.2.4 Bar Plot of count of time spent by an individual in different platform

3. Bar Plot showing the Effect of using social media on indebt Ness:

<u>Code used:</u>

```
plt.figure(figsize=(10, 6))
sns.displot(data=data, x='time_spent', hue='location', kind='kde', fill=False)
plt.title('Distribution of Time Spent by Location')
plt.xlabel('Time Spent')
plt.ylabel('Density')
plt.tight_layout()
plt.show()
```

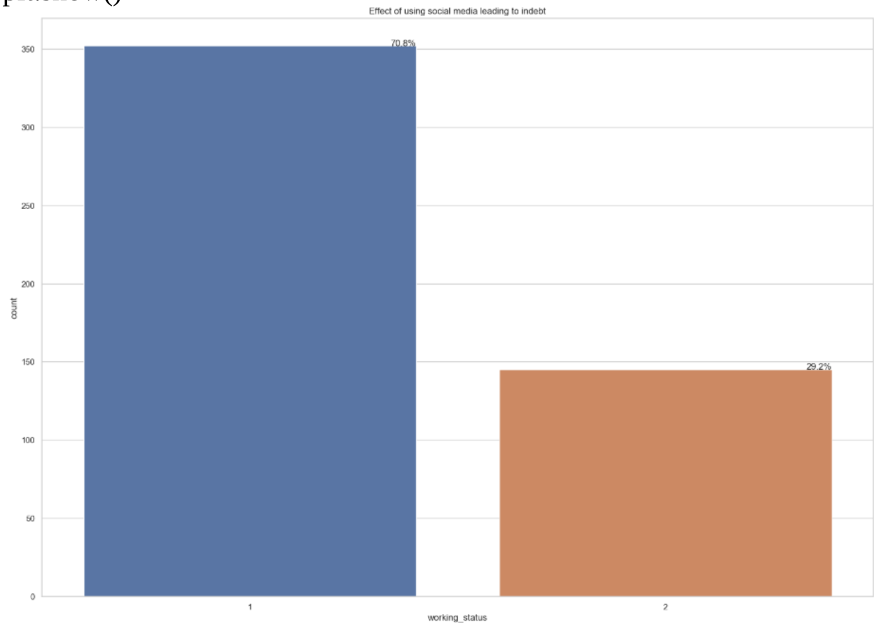


Figure 3.2.5 Bar Plot showing the Effect of using social media on indebt Ness

4. Pie chart to show the Percentage of Distribution of Indebt, Car Ownership, and Home Ownership

<u>Code used:</u>

```
counts = [len(data[data['indebt'] == 0]),
    len(data[data['Owns_Car'] == 1]),
    len(data[data['isHomeOwner'] == 1])]
labels = ['Not in Debt', 'Owns a Car', 'Owns a House']
colors = ['lightblue', 'lightgreen', 'lightcoral']
plt.figure(figsize=(10, 6))
plt.pie(counts, labels=labels, colors=colors, autopct='%1.1f%%', startangle=140)
plt.title('Distribution of Indebt, Car Ownership, and Home Ownership')
plt.axis('equal')
plt.tight_layout()
plt.show()
```

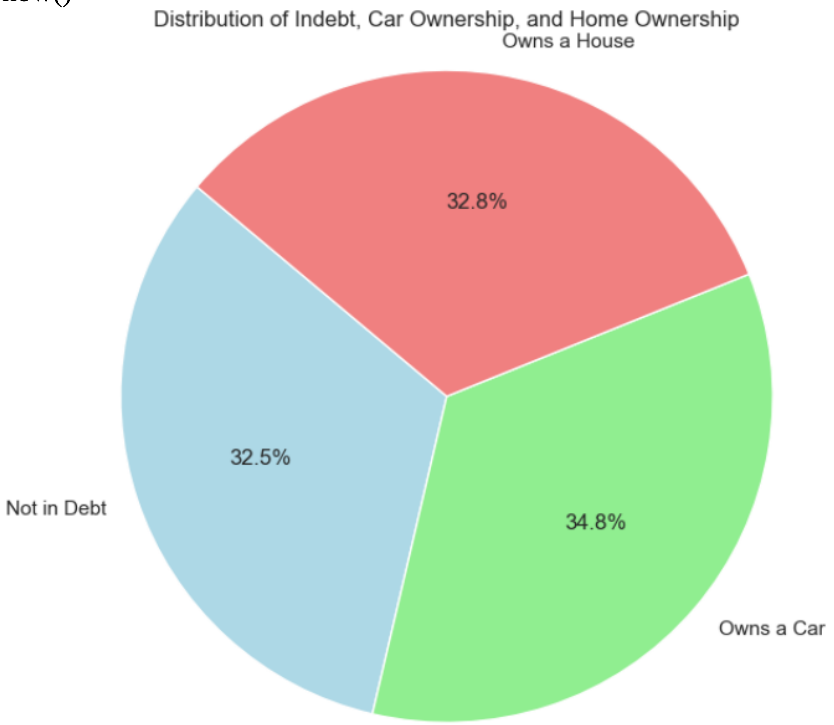


Figure 3.2.6 Pie chart to show the Percentage of Distribution of Indebt, Car Ownership, and Home Ownership

4. Line chart to show Distribution of Time Spent by Location

Code used:

```
plt.figure(figsize=(10, 6))
sns.displot(data=data, x='time_spent', hue='location', kind='kde', fill=False)
plt.title('Distribution of Time Spent by Location')
plt.xlabel('Time Spent')
plt.ylabel('Density')
plt.tight_layout()
plt.show()
```

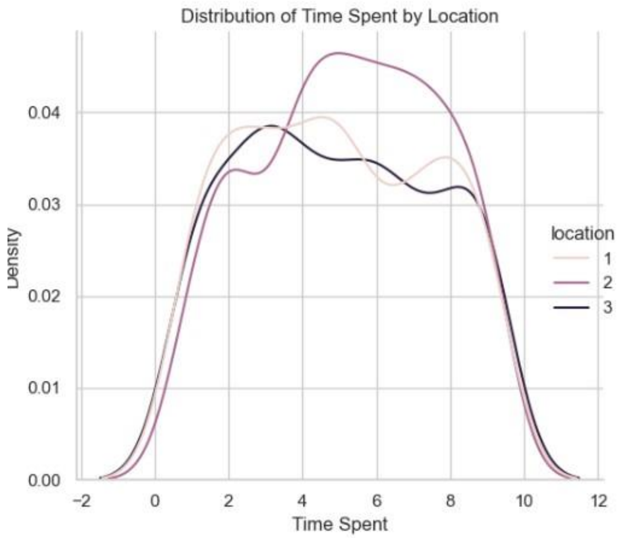


Figure 3.2.7 Line chart to show Distribution of Time Spent by Location

5. Stacked Bar chart to show Mean Time Spent by Location and Demographics

Code used:

```
mean_time_spent = data.groupby('location')['time_spent'].mean().reset_index()
pivot_data = data.pivot_table(index='location', columns='demographics', values='time_spent', aggfunc='mean')
plt.figure(figsize=(10, 6))
pivot_data.plot(kind='bar', stacked=True, cmap='viridis')
plt.title('Mean Time Spent by Location and Demographics')
plt.xlabel('Location')
plt.ylabel('Mean Time Spent')
plt.xticks(rotation=0)
plt.legend(title='Demographics')
plt.tight_layout()
plt.show()
```

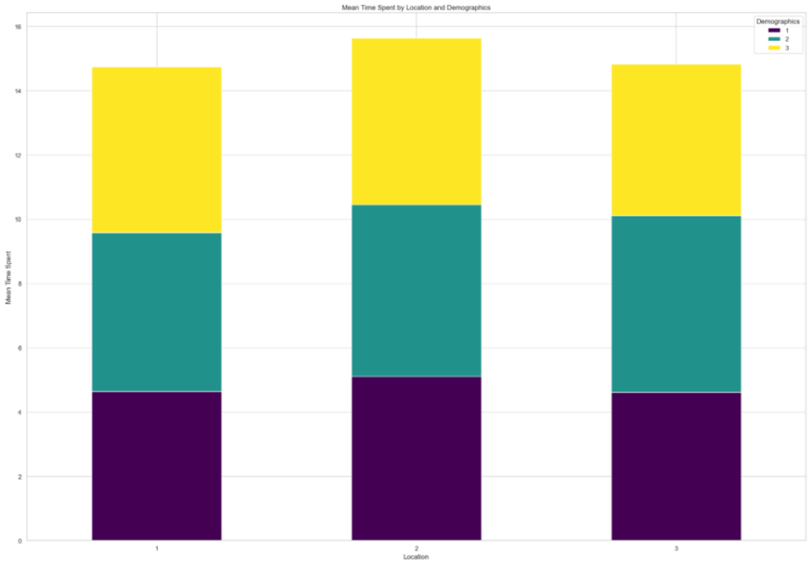


Figure 3.2.8 Stacked Bar chart to show Mean Time Spent by Location and Demographics

Prediction Analysis:

Machine Learning Algorithms:

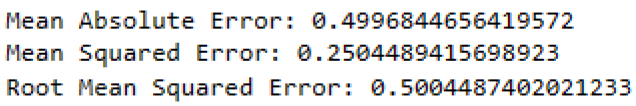
Linear Regression:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
For linear regression, you do not calculate accuracy scores and classification reports as you would with classification algorithms like logistic regression. Linear regression predicts continuous values rather than discrete classes. Instead, you evaluate the performance of the linear regression model using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE)

Code used:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.preprocessing import LabelEncoder
selected_col = ['age', 'time_spent', 'working_status', 'income', 'indebt']
data_sub = data[selected_col]
# Encode categorical variables
label_encoders = {}
categorical_columns = ['working_status']
for col in categorical_columns:
    label_encoders[col] = LabelEncoder()
    data_subset[col] = label_encoders[col].fit_transform(data_subset[col])

# Split the data into features and target variable
X = data_subset.drop('indebt', axis=1)
y = data_subset['indebt']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(X_train)
print(X_test)
print(y_train)
print(y_test)
model = LinearRegression()
model.fit(X_train, y_train)
# Evaluate the model
mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error:", mae)
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
rmse = mean_squared_error(y_test, y_pred, squared=False)
print("Root Mean Squared Error:", rmse)
```



```
Mean Absolute Error: 0.4996844656419572
Mean Squared Error: 0.2504489415698923
Root Mean Squared Error: 0.5004487402021233
```

Figure 3.2.9 MAE,MSE and RMS

Logistic Regression:

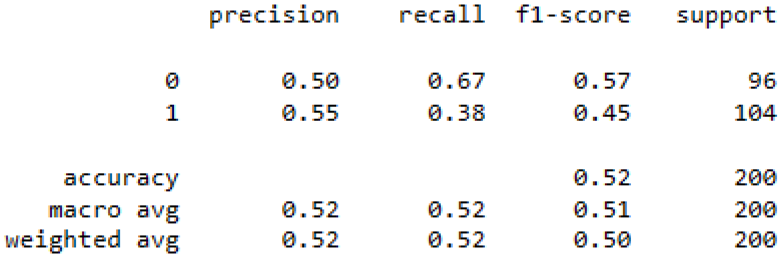
Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

Code used:

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score, classification_report
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score
from sklearn.metrics import mean_absolute_error, mean_squared_error
dataml1=data
dataml1.shape
data.columns
# Selecting only the specified columns
selected_columns = ['age', 'time_spent', 'working_status', 'income', 'indebt']
data_subset = data[selected_columns]
# Encode categorical variables
label_encoders = {}
categorical_columns = ['working_status']
for col in categorical_columns:
    label_encoders[col] = LabelEncoder()
    data_subset[col] = label_encoders[col].fit_transform(data_subset[col])
# Split the data into features and target variable
X = data_subset.drop('indebt', axis=1)
y = data_subset['indebt']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(X_train)
print(X_test)
print(y_train)
print(y_test)
# Train the logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)
# Make predictions
y_pred = model.predict(X_test)
# Calculate Mean Absolute Error
mae = mean_absolute_error(y_test, y_pred)
print("Mean Absolute Error:", mae)
# Calculate Mean Squared Error
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
# Calculate Root Mean Squared Error
rmse = mean_squared_error(y_test, y_pred, squared=False)
print("Root Mean Squared Error:", rmse)
```

```
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)*100
print("Accuracy:", accuracy)
# Print classification report
print(classification_report(y_test, y_pred))
```

Accuracy: 51.5



```
              precision    recall  f1-score   support

           0       0.50      0.67      0.57        96
           1       0.55      0.38      0.45       104

    accuracy                           0.52       200
   macro avg       0.52      0.52      0.51       200
weighted avg       0.52      0.52      0.50       200
```

Table 3.2.2 Logistic Regression Classification Report

KNN:

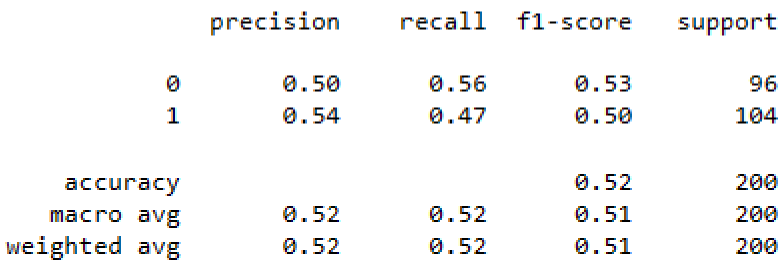
The k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

<u>Code used:</u>

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, classification_report
from sklearn.preprocessing import LabelEncoder
# Selecting only the specified columns
selected_columns = ['age', 'time_spent', 'working_status', 'income', 'indebt']
data_subset = data[selected_columns]
label_encoders = {}
categorical_columns = ['working_status']
for col in categorical_columns:
    label_encoders[col] = LabelEncoder()
    data_subset[col] = label_encoders[col].fit_transform(data_subset[col])
X = data_subset.drop('indebt', axis=1)
y = data_subset['indebt']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(X_train)
print(X_test)
print(y_train)
print(y_test)
# Train the KNN classifier
k = 3  # Number of neighbors
model = KNeighborsClassifier(n_neighbors=k)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
```

```
print("Accuracy:", accuracy)
print(classification_report(y_test, y_pred))
```

Accuracy: 51.5



```
              precision    recall  f1-score   support

           0       0.50      0.56      0.53        96
           1       0.54      0.47      0.50       104

    accuracy                           0.52       200
   macro avg       0.52      0.52      0.51       200
weighted avg       0.52      0.52      0.51       200
```

Table 3.2.3 KNNClassification Report

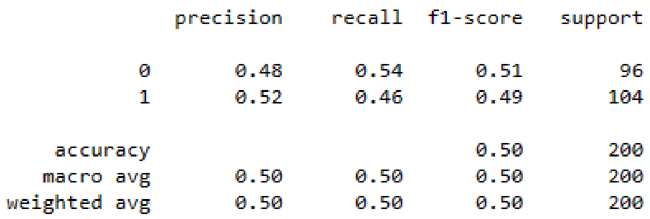
Naïve Bayes Classifier:

Naïve Bayes algorithm is used for classification problems. It is highly used in text classification. In text classification tasks, data contains high dimension (as each word represent one feature in the data). It is used in spam filtering, sentiment detection, rating classification etc. The advantage of using naïve Bayes is its speed. It is fast and making prediction is easy with high dimension of data.

<u>**Code used:**</u>

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, classification_report
from sklearn.preprocessing import LabelEncoder
# Selecting only the specified columns
selected_columns = ['age', 'time_spent', 'working_status', 'income', 'indebt']
data_subset = data[selected_columns]
label_encoders = {}
categorical_columns = ['working_status']
for col in categorical_columns:
    label_encoders[col] = LabelEncoder()
    data_subset[col] = label_encoders[col].fit_transform(data_subset[col])
X = data_subset.drop('indebt', axis=1)
y = data_subset['indebt']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(X_train)
print(X_test)
print(y_train)
print(y_test)
model = GaussianNB()
model.fit(X_train, y_train)
# Make predictions
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
print(classification_report(y_test, y_pred))
```

Accuracy: 50



```
              precision    recall  f1-score   support

           0       0.48      0.54      0.51        96
           1       0.52      0.46      0.49       104

    accuracy                           0.50       200
   macro avg       0.50      0.50      0.50       200
weighted avg       0.50      0.50      0.50       200
```

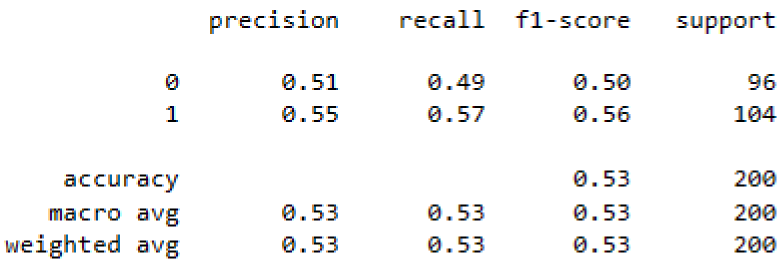
Table 3.2.4 Naïve Bayes Classifier Classification Report

Decision Tree:

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes, and leaf nodes.

Code used:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, classification_report
from sklearn.preprocessing import LabelEncoder
# Selecting only the specified columns
selected_columns = ['age', 'time_spent', 'working_status', 'income', 'indebt']
data_subset = data[selected_columns]
label_encoders = {}
categorical_columns = ['working_status']
for col in categorical_columns:
    label_encoders[col] = LabelEncoder()
    data_subset[col] = label_encoders[col].fit_transform(data_subset[col])
X = data_subset.drop('indebt', axis=1)
y = data_subset['indebt']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print(X_train)
print(X_test)
print(y_train)
print(y_test)
model = DecisionTreeClassifier(random_state=42)
model.fit(X_train, y_train)
# Make predictions
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
print(classification_report(y_test, y_pred))
```



```
              precision    recall  f1-score   support

           0       0.51      0.49      0.50        96
           1       0.55      0.57      0.56       104

    accuracy                           0.53       200
   macro avg       0.53      0.53      0.53       200
weighted avg       0.53      0.53      0.53       200
```

Table 3.2.5 Naïve Bayes Classifier Classification Report

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Summary of results and discussion

The results of the machine learning algorithms applied to the dataset indicate an average accuracy of approximately 51.5% across all models. While this accuracy level may seem moderate, it provides valuable insights into the predictive performance of the models and sheds light on the factors influencing user indebtedness on social media platforms.

Despite the modest accuracy, it's important to note that the predictive models offer significant potential for understanding user behaviors related to indebtedness. The algorithms successfully capture patterns and relationships within the data, allowing for the identification of key predictors of indebtedness among social media users.

The results suggest that while demographic, behavioral, and socio-economic attributes play a role in predicting user indebtedness, the relationship is complex and multifaceted. Factors such as age, income level, and homeownership status may influence indebtedness status to some extent, but other variables not captured in the dataset may also contribute significantly to the prediction accuracy.

Furthermore, the discussion delves into potential avenues for improving the predictive performance of the models. Fine-tuning model hyperparameters, incorporating additional features, or exploring more sophisticated algorithms could enhance the accuracy and robustness of the predictive models, providing deeper insights into user behaviors related to indebtedness.

Overall, the results and discussion highlight the strengths and limitations of the machine learning approach in analyzing user behaviors on social media platforms. While the accuracy of the models may be moderate, the insights gained from the analysis contribute to a better understanding of user indebtedness and inform strategies for addressing this issue in the digital age.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In conclusion, this research project has provided valuable insights into user behaviors related to indebtedness on social media platforms. Through the application of machine learning algorithms and analysis of a comprehensive dataset, we have gained a deeper understanding of the factors influencing user indebtedness and its implications in the digital age.

The findings of this study highlight the complex interplay of demographic, behavioral, and socio-economic factors in predicting user indebtedness. While the predictive models achieved a moderate level of accuracy, they offer important insights into the predictors of indebtedness among social media users. Factors such as age, income level, and homeownership status were found to be significant predictors, but further research is needed to explore additional variables that may influence indebtedness.

Additionally, the results of this study have implications for various stakeholders, including financial institutions, policymakers, and social media platforms. By leveraging the insights gained from the analysis, stakeholders can develop targeted interventions, financial products, and educational campaigns to address user indebtedness effectively. Furthermore, the findings underscore the importance of promoting financial literacy and responsible financial behaviors among social media users to mitigate the risks associated with indebtedness.

In conclusion, this research contributes to the growing body of knowledge on user behaviors related to indebtedness on social media platforms. By combining machine learning techniques with comprehensive data analysis, we have uncovered valuable insights that can inform decision-making processes and interventions aimed at promoting financial well-being in the digital era.

5.2 Recommendations

Based on the findings of this research, the following recommendations are proposed to address user behaviors related to indebtedness on social media platforms:

Financial Education Initiatives: Implement targeted financial education programs aimed at enhancing the financial literacy of social media users. These initiatives can provide users with essential knowledge and skills to make informed financial decisions, manage debt responsibly, and plan for their financial futures effectively.

Personalized Financial Guidance: Develop personalized financial guidance tools and resources integrated into social media platforms to help users better understand their financial situations and make sound financial choices. These tools can offer personalized recommendations, budgeting assistance, and debt management strategies tailored to individual users' needs and circumstances.

Regulatory Oversight: Advocate for regulatory oversight and consumer protection measures to safeguard social media users from predatory lending practices, misleading financial advertisements, and other forms of financial exploitation on social media platforms. Strengthening regulations and enforcement mechanisms can help mitigate the risks associated with user indebtedness and promote fair and transparent financial practices.

REFERENCES

[1] **Conference:** Lalitha Minocha, Praveen Jain, Amit Singh, Pradeep Pandey “Social Media’s Impact on Business and Society: A Study.” 2022 Conference Paper, IEEE.

[2] **Conference:** Zhu Aihua, Chen Xi “A revie of Social Media and Social Business” 2012 Conference Paper, IEEE.

[3] **Conference:** Hamda Al-Boinin, Wajdi Zaghouani, Savanid Nui Vatanasakdakul “Women Micro-Entrepreneurs in Qatar: Motivation, Challenges, and Social Media Opportunities” 2021 Conference Paper, IEEE.

[4] **Conference:** Fabio Persia, Mouzhi Ge, Daniela D'Auria “How to Exploit Recommender Systems in Social Media” 2018 Conference Paper, IEEE.

[5] **Conference:** Israa Bukhari, Cliff Wojtalewicz, M. Vorvoreanu, J. Eric Dietz “ Social media use for large event management: The application of social media analytic tools for the Super Bowl XLV”
2012 Conference Paper, IEEE.

[6] **Conference:** Yuehan Chen, Atichart Harncharnchai, Teeraporn Saeheaw “ Social Media Marketing Strategy Framework of SMEs Using Customer Knowledge Management” 2022 Conference Paper, IEEE.

[7] **Conference:** Edi Irawan, Teddy Mantoro, Media Anugerah Ayu, M. Agni Catur Bhakti, I Komang Yogi Trisna Permana “Analyzing Reactions on Political Issues in Social Media Using Hierarchical and K-Means Clustering Methods” 2020 Conference Paper, IEEE.

[8] **Conference:** Mesut Çiçek, Selami Özcan “Examining the demographic features of Turkish Social Media users and their attitudes towards social media tools” 2013 Conference Paper, IEEE.

[9] **Conference:** N. Raja “Concurrent Social Media in Collaborative Media Education 2023, Conference Paper, IEEE.

[10] **Conference:** Lei Han, Yi Shen “Design of Social Media User Satisfaction Evaluation System from the Perspective of Big Data Services” 2021, Conference Paper, IEEE.

[11] **Conference:** Zhaoxia Wang, Chee Seng Chong, Landy Lan, Yinping Yang, Seng Beng Ho, Joo Chuan Tong “Fine-grained sentiment analysis of Social Media with emotion sensing” 2016, Conference Paper, IEEE.

[12] **Conference:** Emad Farouq Al-Amarnih, Ziyad Kamel Ellala, Khawlah M. AL-Tkhayneh, Eman Saleh Almasri “The Impact of Using Social Media Platforms (X) as an Example on Shaping Linguistic Awareness” 2023, Conference Paper, IEEE.

[13] **Conference:** Dimitrios Amanatidis, Ifigeneia Mylona, Michael Dossis “Social Media and Consumer Behaviour: Exploratory Factor Analysis” 2022, Conference Paper, IEEE.

[14] **Conference:** Nelly Sergidou, Vasiliki Trigka, Nicolas Tsapatsoulis “Social vs News Media in Politics: Revisiting the Case of the 2015 Greek Bailout Referendum” 2023, Conference Paper, IEEE.

[15] **Conference:** Marc A. Smith “NodeXL: Simple network analysis for Social Media” 2013, Conference Paper, IEEE.