# Social Media Vs Productivity

## Course: Python Essentials

**FACULTY**

Dr. Monika Vyas

**PRESENTED BY**

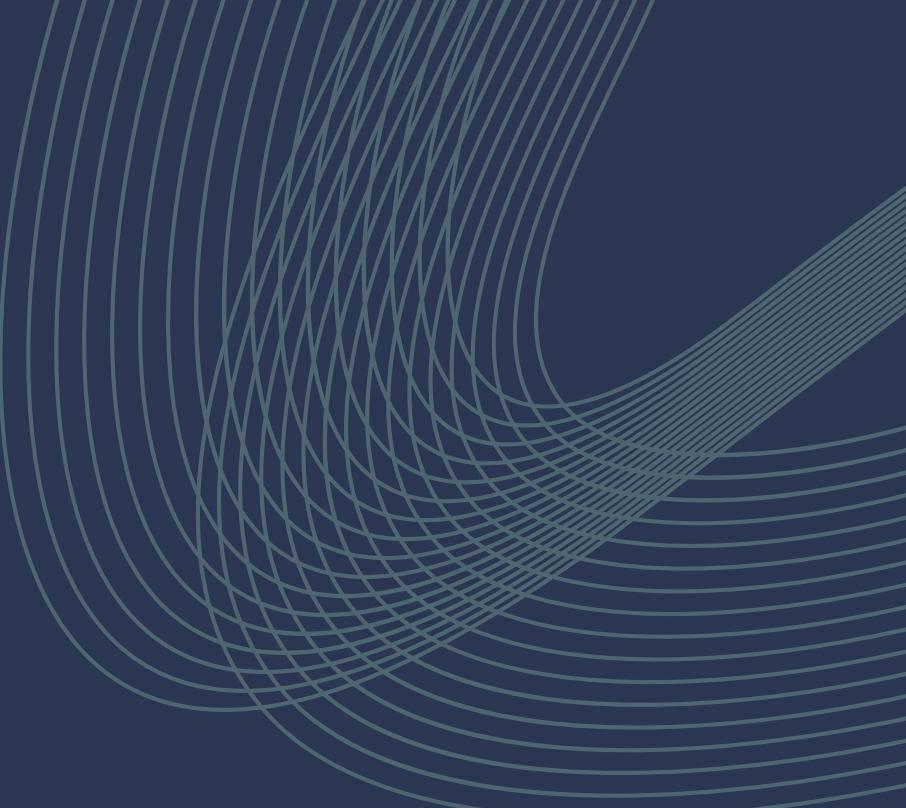Busani Naga Jahnavi

# TABLE OF CONTENTS

# Abstract:

In today's world, social media is a part of almost everyone's daily routine. People use it for communication, entertainment and information. However, when the usage becomes too high, it may reduce a person's productivity. In this project, I analyse a public Kaggle dataset to study the relationship between daily social media usage and self-reported productivity.

Using Python in Google Colab, I load the dataset, clean it, perform simple exploratory data analysis and build a basic Linear Regression model. The project does not try to create a perfect prediction system. Instead, it aims to show a complete, easy-to-understand data science workflow using a real dataset and to get a rough idea of how social media time may be connected with productivity.

# Introduction:

Social media platforms such as Instagram, WhatsApp, Facebook, YouTube and Snapchat are widely used by students and working professionals. They help us stay connected and updated, but they can also distract us from important tasks. Many people feel that when they spend too much time on social media, their productivity reduces.

Instead of just relying on opinions, this project uses actual data to explore the relationship between social media usage and productivity. The main idea is to see whether the number of hours spent on social media per day has any visible effect on a person's productivity score. To do this, I use Python in Google Colab, along with basic data science steps such as data loading, cleaning, visualisation and simple machine learning.

# Dataset Details:

   The dataset used in this project is taken from Kaggle and is related to social media usage and productivity.
- Source: Kaggle
- Dataset name: Social Media vs Productivity
- File format: CSV

Some of the important columns in the dataset are:
- age – age of the person
- gender – gender of the person
- daily_social_media_time – number of hours spent on social media per day
- actual_productivity_score – self-reported productivity score
- social_platform_preference – most used social media platform
- number_of_notifications – number of notifications received

The dataset was downloaded as a ZIP file from Kaggle, extracted on the local machine and then uploaded to Google Drive. In Google Colab, the CSV file was read into a pandas DataFrame using the read_csv() function.

# Tools and Technologies:

The following tools and technologies were used in this project:

- Programming language: Python
- Environment: Google Colab
- Storage: Google Drive

Python libraries:

- pandas – for reading the CSV file and handling tabular data
- matplotlib – for creating graphs and visualisations
- scikit-learn – for building and evaluating the Linear Regression model

# Methodology:

The project was completed using the following steps:

1. Data Loading
   - Mounted Google Drive in Google Colab.
   - Provided the path of the CSV file stored in Drive.
   - Loaded the dataset into a pandas DataFrame using the read_csv() function.
2. Data Cleaning
   - Checked the shape of the dataset and the names of all columns.
   - Calculated the number of missing values in each column.
   - Removed rows where daily_social_media_time or the productivity score column were missing, so that the analysis and model are based on complete records.
3. Exploratory Data Analysis (EDA)
   - Viewed the first few rows of the dataset using head().
   - Used describe() to get summary statistics such as mean and standard deviation.
   - Plotted a histogram of daily_social_media_time to see how social media time is distributed.
   - Plotted a scatter plot of daily_social_media_time versus productivity score to observe their relationship.
   - Plotted a bar chart showing average productivity for different social_platform_preference values.
4. Model Building
   - Chose daily_social_media_time as the input feature (X).
   - Used the productivity score column as the target (y).
   - Split the data into training and testing sets.
   - Trained a Linear Regression model from scikit-learn on the training data.
5. Model Evaluation
   - Predicted the productivity scores for the test set.
   - Evaluated the model using Mean Absolute Error (MAE) and $R^2$ score.
   - Printed the values of MAE and $R^2$ to understand how well the model performed.
6. Interpretation
   - Combined the information from the graphs and the evaluation metrics.
   - Wrote a conclusion about how strongly social media time is related to productivity and whether this single feature is enough to make accurate predictions.

# Results and Discussion:

The exploratory data analysis gives some useful observations:

• The histogram of daily_social_media_time shows that most users spend a few hours per day on social media, with some users spending very high or very low time.
• The scatter plot of daily_social_media_time versus productivity score shows a wide spread of points. There is no very strong or clear straight-line pattern between the two variables. This suggests that other factors may also play an important role in determining productivity.
• The bar chart of average productivity by social_platform_preference shows only small differences in productivity between users of different platforms. Platform choice alone does not appear to have a very big effect.

   After training the Linear Regression model with daily_social_media_time as the only input feature, the model produced the following results:
• Mean Absolute Error (MAE): 1.621
• $R^2$ score: approximately 0.00

An MAE of 1.621 means that, on average, the model's predictions are off from the real productivity scores by about 1.6 units. An $R^2$ value close to zero means that the model is not able to explain much of the variation in productivity using only social media time. In simple words, just the number of hours spent on social media is not enough to accurately predict productivity for the people in this dataset.

These results match the visual patterns seen in the scatter plot, where points are scattered and do not lie closely along a line.

# Conclusion:

This project used Python and Google Colab to analyse the "Social Media vs Productivity" dataset from Kaggle. The goal was to study how daily social media usage is related to self-reported productivity and to build a simple predictive model.

The analysis showed that while social media usage is an interesting factor, it alone does not strongly determine productivity. The Linear Regression model gave a Mean Absolute Error of about 1.621 and an $R^2$ score near zero. This means that the model cannot accurately predict productivity using only daily social media time. The scatter plots and bar charts also support this conclusion, as they do not show a strong relationship.

Even though the model is not highly accurate, the project successfully demonstrates a complete data science pipeline: loading a real dataset, cleaning it, exploring it using visualisations, building a simple model and interpreting the results. It also highlights that real-life problems are often affected by many different variables, not just one.

# Future Enhancements:

The project can be improved in several ways:

• Add more input features such as sleep duration, study or work hours, stress level and time spent on other applications.
• Try more powerful machine learning models like Decision Trees, Random Forests or Gradient Boosting which may capture complex patterns better.
• Perform feature selection and correlation analysis to identify which factors have the strongest influence on productivity.
• Develop a simple web or mobile interface where a user can enter their daily habits and get an estimated productivity score along with suggestions to improve it.

# References:

1. Kaggle, "Social Media vs Productivity" dataset.
2. Scikit-learn documentation, https://scikit-learn.org
3. Pandas documentation, https://pandas.pydata.org
4. Matplotlib documentation, https://matplotlib.org