Graphical user interface

Description automatically generated with low confidence

**Predicting Students’ Grades from Their Activities**

By

Mohammad Messbah Uddin

Leeds Beckett University

School of Built Environment, Engineering, and Computing

Abstract

What students do on a daily basis throughout their academic year has an impact on their grades. Their daily activities include the time they are studying, spending time with friends, the apps they are using, the time they spend using their phones, texts, calls, etc. Their mental condition also plays a significant role in their grades. In this exercise, I have worked with some of these above-mentioned data, collected from students for ten weeks to find their relationships among themselves, and tried to predict their grades based on those data. The finding confirms that there are some correlations between the data. This report discusses the algorithm used, findings, limitations, and how to make the result more accurate.

Contents

[Abstract 1](#_Toc73059194)

[1. Introduction 3](#_Toc73059195)

[1.1 Aim and research objectives 3](#_Toc73059196)

[2. Literature Review 4](#_Toc73059197)

[2.1 CO2 emission from Mobile phones 4](#_Toc73059198)

[2.2 Smartphones vs. Grades 4](#_Toc73059199)

[2.3 Loneliness vs. Academic Performance 4](#_Toc73059200)

[3. Methodology 4](#_Toc73059201)

[3.1 Data Collection 4](#_Toc73059202)

[3.2 Methodology 4](#_Toc73059203)

[3.2.1 Data Selection and Cleaning 5](#_Toc73059204)

[3.2.2 Data Transformation 6](#_Toc73059205)

[4. Finding and Discussion 7](#_Toc73059206)

[4.1 Descriptive Analysis 7](#_Toc73059207)

[4.2 Inferential Analysis 11](#_Toc73059208)

[4.3 Regression Analysis 12](#_Toc73059209)

[4.3.1 Linear Regression 12](#_Toc73059210)

[4.3.2 Multiple Variable Linear Regression 13](#_Toc73059211)

[4.3.3 Polynomial Regression 14](#_Toc73059212)

[5. Conclusion and Future work 17](#_Toc73059213)

[References 17](#_Toc73059214)

[Bibliography 17](#_Toc73059215)

[Appendix 18](#_Toc73059216)

Table of Figures

[Figure 1: KDD Steps (Costagliola, et al., 2009) 5](#_Toc73059217)

[Figure 2: App Usage Data 6](#_Toc73059218)

[Figure 3: Call Log Data Overview 6](#_Toc73059219)

[Figure 4: Columns in survey 7](#_Toc73059220)

[Figure 5: Analysis of students' call duration 8](#_Toc73059221)

[Figure 6: CO2 Emission by Students 8](#_Toc73059222)

[Figure 7: Cumulative CO2 Emission by Students 9](#_Toc73059223)

[Figure 8: CO2 Emission from SMS 9](#_Toc73059224)

[Figure 9: Processed Dataset's Descriptive analysis 10](#_Toc73059225)

[Figure 10: Overview of the processed dataset 10](#_Toc73059226)

[Figure 11: Scatterplot Map of the Final Dataset 11](#_Toc73059227)

[Figure 12: Pearson's Correlation Graph 12](#_Toc73059228)

[Figure 13: Linear Regression Analysis 13](#_Toc73059229)

[Figure 14: Multiple Variable Linear Regression Model Summary 13](#_Toc73059230)

[Figure 15: Multiple Variable Linear Regression Result 14](#_Toc73059231)

[Figure 16: Grades Vs. Phone Usage(Order=1, underfit) 15](#_Toc73059232)

[Figure 17: Grades Vs. Phone Usage (Order=5, perfect fit) 15](#_Toc73059233)

[Figure 18: Polynomial Regression plot 15](#_Toc73059234)

[Figure 19:Grades Vs. Loneliness Scale(Order=1) 16](#_Toc73059235)

[Figure 20:Grades Vs. Loneliness Scale(Order=5) 16](#_Toc73059236)

[Figure 21: Grades vs. Loneliness Polynomial Regression Plot 16](#_Toc73059237)

[Figure 22: Confidence Scale Vs. Loneliness Scale(Order=1) 17](#_Toc73059238)

[Figure 23: Polynomial Regression of Confidence Vs. Loneliness 17](#_Toc73059239)

Tables

[Table 1: Survey Response Conversion 7](#_Toc73059253)

[Table 2: Application Usage Data Conversion 7](#_Toc73059254)

[Table 3: Correlation coefficient r values and P values of mostly correlated columns 12](#_Toc73059255)

# 1. Introduction

With the advent of science and technology, students’ dependency on phones has been increasing a lot lately, which has some pros and cons. One major con is the impact on the environment. According to (Reboxed.co, 2021), an hour of mobile phone use is responsible for 63kg of CO2 emission a year. Besides, many people believe there is a negative correlation between the time students spend on their phones and their academic performance. This belief may be proven untrue in some cases since the purpose of using mobile phones may vary, especially during the current pandemic, when the classes have moved online. Even before that, some students used their smartphones to improve their academic and personal skills via different educational applications. The problem does not lie on the phone, rather on the way it is being used. This experiment aims to prove the previous statement by data.

## 1.1 Aim and research objectives

To be more precise, this study aims to find how students’ daily activities and their mental condition affect their academic results. The following objectives will assist in reaching the aim:

1. RO1: Level 1 Descriptive statistical analysis includes:

* Analyzing relevant studies.
* Determining CO2 emission by the students via different ways of phone use.
* Getting a clear idea about the dataset analyzing and plotting some data available.

2. RO2: Level 2 Inferential Statistical Analysis includes:

* Finding correlations between some key attributes and decide what activities affect students’ grades most.

3. RO3: Level 3, Machine learning technique includes:

* Creating several regression models to better understand and predict students’ grades as per their activities and mental conditions.

These objectives have been created to support the aim, and this will help to conclude how much impact students’ daily activities on phones have on their grades.

# 2. Literature Review

## 2.1 CO2 emission from Mobile phones

The use of mobile phones has evolved a lot from merely sending texts, making calls to browsing the internet, watching movies, playing high-end video games, banking, etc. While this extensive use comes as a boon, it also comes with some drawbacks. These drawbacks include high Green House Gas (GHG) emissions and addiction to certain types of applications that waste a vast amount of time. (Belkhir, 2018) predicted that CO2 emission from smartphones would jump to 125 megatons per year from 17 megatons. This is an alarming 730% growth. Although most of these emissions are caused by smartphone production, emission due to use is not negligible. An individual uses more than five gigabytes of data by their phones per month, emitting at least 15kg of CO2 (Julian Pscheid, 2020).

## 2.2 Smartphones vs. Grades

Smartphones have made information more accessible and faster. However, (Felisoni & Godoi, 2018) conducted a study that suggests, “It is more likely that a student who uses less his or her cellphone will have a higher grade than the one who uses more, given an equal performance in the college's entrance exam and same belief to self regulate their own studying.” Many other similar studies have been conducted which infers that there is a negative correlation between students’ smartphone use and their academic performance (Morphitou, 2014), (Kibona & and Mgaya, 2015), (Yi, et al., 2016).

## 2.3 Loneliness vs. Academic Performance

Loneliness plays a role in students' academic performance. An author (Bek, 2017) found a significant connection between students’ loneliness and academic performance. She says, “students who feel lonely and isolated tend to spend their time idly and therefore do not thrive in academic environments.” Following a study, (Pervez, 2018) opposes Bek,s findings mentioning, “There is a definite but low and positive overall correlation exists between Loneliness and Academic Achievement.”

# 3. Methodology

## 3.1 Data Collection

A dataset called “StudentLife” is used in this study. The dataset contains 49 students’ phone data collected for ten weeks and some survey data (Rui, et al., 2014). The dataset includes information such as when a student made a phone call, when he ended it, when a text message is sent, time a phone was charged, how many times a particular application is opened, how long a phone was charged, surveys about their mental conditions, etc.

## 3.2 Methodology

The methodology used in this work is Knowledge discovery in databases (KDD) which includes steps like Data selection and cleaning, Data transformation, Data mining/ Pattern Discovery, and Evaluation.

Diagram

Description automatically generated

Figure 1: KDD Steps (Costagliola, et al., 2009)

### 3.2.1 Data Selection and Cleaning

The data set is enormous, divided into many CSV files, and many data did not go with my research goal. I chose to work with app\_usage, call\_log, grades, phone\_charge, phone\_lock, survey, and dropped all other data.

App\_usage contains 49 students’ CSV files, each file has ten columns, and the row number is variable. From there, I extracted only timestamp and package activity data like in figure 2. The preprocessing also includes getting the number of times an application is opened.

Table

Description automatically generated

Figure 2: App Usage Data

Call\_log has only twenty students’ phone call information. Each file contains many null values which needed cleaning (Figure 3).

A screenshot of a computer

Description automatically generated with medium confidence

Figure 3: Call Log Data Overview

Data in this figure contains one student’s information only. User ID and call duration are extracted from this file, and all the students’ call logs’ data are merged for further processing.

Phone\_Charge contains two columns named ‘start’ and ‘end,’ which denotes timestamps when the phone is plugged in and out of the charger. Phone\_lock includes the same type of columns indicating locking and unlocking time.

The survey data includes 20 columns as in figure 4. Unnecessary columns and rows were dropped. Again Grades dataset does not have data for all students. I took survey data into account for only those students whose grades are available.

Text, letter

Description automatically generated

Figure 4: Columns in survey

### 3.2.2 Data Transformation

Timestamp data, categorical data, etc., needed some transformation for the analysis. The conversion of survey responses is shown in table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Survey Response Conversion | | | |
| Do you feel lonely? | | Confidence Level? | |
| Never | 1 | Very weak | 1 |
| Rarely | 2 | Weak | 2 |
| Sometimes | 3 | Fair | 3 |
| Often | 4 | Strong | 4 |
| Always | 5 | Very strong | 5 |

Table 1: Survey Response Conversion

I needed time spent on using particular applications by each student, but that data is not achievable from the dataset. The dataset only includes how many times a user opens an application. Since the number of applications is very high, I categorized the applications into three sectors: Entertainment, Educational, Health, and Fitness. Then the data is transformed again like in table 2.

|  |  |  |  |
| --- | --- | --- | --- |
| Application Usage | | | |
| Entertainment Applications Opened | | Education and Fitness Applications Opened | |
| > 4000 times | 5 | > 2500 times | 5 |
| >3000 & <4000 | 4 | >1500 & <2500 | 4 |
| >1500 & <3000 | 3 | >1000 & <1500 | 3 |
| >500 & <1500 | 2 | >200 & <1000 | 2 |
| <500 | 1 | <200 | 1 |

Table 2: Application Usage Data Conversion

To get how long a student used his phone, timestamp data needed to be converted into standard time. Phone usage is calculated by the formula below:

Time in locked Modedate = date

Phone Usage = 24 – Time in locked Mode.

# 4. Finding and Discussion

## 4.1 Descriptive Analysis

For the descriptive part of the study, 19 users' call duration is combined in a file, and the descriptive result is shown in figure 5. Figures 6 to 8 show how much carbon is emitted by each user from their phone calls and text messages, respectively. However, Carbon emission data is calculated by multiplying call duration by 0.06 and SMS by 0.014 since one minute of phone call produces 0.06kg of CO2 and an SMS produces 0.014gCO2 (Sanchez, et al., 2018). The finding from descriptive analysis includes that all the students combined emitted 604.938KG CO2 in 10 weeks by making phone calls and sending SMS only.

Text

Description automatically generated

Figure 5: Analysis of students' call duration

Chart, histogram

Description automatically generated

Figure 6: CO2 Emission by Students

Chart, line chart

Description automatically generated

Figure 7: Cumulative CO2 Emission by Students

Chart, histogram

Description automatically generated

Figure 8: CO2 Emission from SMS

A new dataset is created, taking and merging information from separated datasets containing Students User ID, Phone\_Usage, Time spent on Educational, Entertainment and Fitness Application and grades. A brief analysis of the dataset can be seen in figure 9.

Table

Description automatically generated

Figure 9: Processed Dataset's Descriptive analysis

Graphical user interface, chart

Description automatically generated with medium confidence

Figure 10: Overview of the processed dataset

A picture containing sky, shoji, train, outdoor

Description automatically generated

Figure 11: Scatterplot Map of the Final Dataset

Figures 10 and 11 give a lot of information about the dataset prepared for the main part of the report. Most of the students did not spend much time using entertainment applications, but they were meticulous about their health and fitness. This figure also suggests that they did not use educational applications that much either.

## 4.2 Inferential Analysis

To understand the behaviors of the dataset, Pearson's correlation method is applied. It helped to conduct further research by showing which columns will yield better results in the next level. Figure 11 shows that loneliness is negatively but highly correlated with confidence; phone usage also has an adverse impact on students’ grades. However, One finding here contradicts other researchers' results. Most researchers found that loneliness is negatively correlated with academic results, but it is positively correlated here. Students in this dataset probably got more time to study because they had fewer friends to spend time with.

Chart, treemap chart

Description automatically generated

Figure 12: Pearson's Correlation Graph

|  |
| --- |
|  |
| Text  Description automatically generated with medium confidence |
|  |
| A picture containing chart  Description automatically generated |
|  |

Table 3: Correlation coefficient r values and P values of mostly correlated columns

## 4.3 Regression Analysis

### 4.3.1 Linear Regression

Linear regression algorithm is applied Phone Usage Vs. Grades data (Figure 12).

Chart, scatter chart

Description automatically generated

Figure 13: Linear Regression Analysis

### 4.3.2 Multiple Variable Linear Regression

For multiple variable linear regression analysis, data is modelled as shown below.

*data\_x = table[['Phone\_Usage','Time\_Spent\_On\_Entertainment\_App',Time\_Spent\_On\_Educational\_App',*

*'Health\_Fitness\_App', 'Loneliness\_Scale', 'Confidence\_Scale']]*

*data\_y = table[['Grades']]*

The model’s summary and the result are shown in Figures 12 and 13.

*Table

Description automatically generated*

Figure 14: Multiple Variable Linear Regression Model Summary

Chart

Description automatically generated

Figure 15: Multiple Variable Linear Regression Result

The graph in figure 15 indicates that students who often felt lonely (Red) got better grades than students who sometimes (Green) or rarely (Yellow) felt lonely. However, one student can be considered an outlier in this dataset since he never felt lonely but did get good grades.

### 4.3.3 Polynomial Regression

Pair-wise polynomial regression is applied in all the columns that showed some sort of correlation.

In the case of Phone Usage Vs. Grades,

Mean Absolute Error, MAE is 0.24834590693235192

Mean Squared Error, MSE is 0.10360913276440142

Root Mean Square Error is 0.3218837255351712

Coefficient of determination or R squared value is

0.19162076856043453

The outcome is shown in figures 16 and 17.

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated  Figure 16: Grades Vs. Phone Usage(Order=1, underfit) | Chart  Description automatically generated  Figure 17: Grades Vs. Phone Usage (Order=5, perfect fit) |

Chart, scatter chart

Description automatically generated

Figure 18: Polynomial Regression plot

In case of the Loneliness scale Vs. Grades,

Mean Absolute Error, MAE is 0.2538246642246651

Mean Squared Error, MSE is 0.10080663606023609

Root Mean Square Error is 0.31750060796829366

Coefficient of determination or R squared value is

0.21348640985459144

Result:

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated  Figure 19:Grades Vs. Loneliness Scale(Order=1) | Chart  Description automatically generated  Figure 20:Grades Vs. Loneliness Scale(Order=5) |

Chart, line chart

Description automatically generated

Figure 21: Grades vs. Loneliness Polynomial Regression Plot

In case of the Confidence scale Vs. Loneliness Scale,

Mean Absolute Error, MAE is 0.36076516076516074

Mean Squared Error, MSE is 0.2102971102971103

Root Mean Square Error is 0.45858162882643944

Coefficient of determination or R squared value is

0.5427096393539348

Result:

|  |  |
| --- | --- |
| Chart, scatter chart  Description automatically generated  Figure 22: Confidence Scale Vs. Loneliness Scale(Order=1) | Chart, line chart  Description automatically generated  Figure 23: Polynomial Regression of Confidence Vs. Loneliness |

In all the cases, polynomial regression showed slightly less RMSE error.

Although I found some results from the research, they are not that significant. This is because I could not use most of this dataset's data for my research for the lack of relevance to my work. Besides, many students' grades are missing in the dataset; otherwise, the result could be more accurate.

# 5. Conclusion and Future work

The primary purpose of this research was to find out what activities and mental condition drives students’ academic performance and what does the opposite. It is found that usage of phones negatively affects academic performance. On the other hand, Students who feel lonely tend to perform better in exams. The study also shows that students who feel lonely lack confidence.

In future, the study can be extended to take more parameters like food habit, study time, gossiping time, etc. into account so that the prediction becomes more accurate, and also an application can be developed accordingly that will alert students when they exceed a limit on any activity that may severely affect his grade or health.

# References

# Bibliography

Amazon, 2021. *sustainability.aboutamazon.com.* [Online]   
Available at: https://sustainability.aboutamazon.com/?energyType=true&workerCount=true&engagementProgram=true&productCategory=true.

Bek, H., 2017. Understanding the Effect of Loneliness on Academic Participation and Success among International University Students. *Journal of Education and Practice,* 8(14), pp. 46-50.

Belkhir, L., 2018. *theconversion.com.* [Online]   
Available at: https://theconversation.com/how-smartphones-are-heating-up-the-planet-92793

Corrales, D. C., Ledezma, A. & Corrales, J. C., 2015. *A Conceptual Framework for Data Quality in Knowledge Discovery Tasks (FDQ-KDT): A Proposal.* s.l., s.n.

Costagliola, G., Fuccella, V., Giordano, M. & Polese, G., 2009. Monitoring Online Tests through Data Visualization. *IEEE Transactions on Knowledge and Data Engineering,* 21(6), pp. 773-784.

Felisoni, D. D. & Godoi, A. S., 2018. Cell phone usage and academic performance: An experiment. *Computers & Education,* 117(0360-1315), pp. 175-187.

Google, 2020. *Environmental Report.* [Online]   
Available at: https://www.gstatic.com/gumdrop/sustainability/google-2020-environmental-report.pdf

Kibona, L. & and Mgaya, G., 2015. Smartphones’ effects on academic performance of higher learning students. *Journal of Multidisciplinary Engineering Science and Technology,* Volume 4, pp. 777-784.

Morphitou, R. N., 2014. The use of smartphones among students in relation to their education and social life. *2014 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL2014),* pp. 315-319.

Pervez, S. I., 2018. The relationship between Loneliness and Academic Achievement: A study on Undergraduate level students.. *International Journal of Creative Research Thoughts (IJCRT),* 6(1), pp. 1302-1305.

Reboxed.co, 2021. *reboxed.co.* [Online]   
Available at: https://reboxed.co/blogs/outsidethebox/the-carbon-footprint-of-your-phone-and-how-you-can-reduce-it

Rui, W. et al., 2014. *StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones..* s.l., ACM.

Sanchez, J. M. P., Dolera, R. M. E., Rosalita, R. G. & Elnar, A. R. B., 2018. Carbon Footprints from Mobile Phone Use among Science Undergraduate Students. *International Journal of Current Research,* Volume 10, pp. 72484-72487.

Yi, Y., You, S. & Bae, B., 2016. The influence of smartphones on academic performance. *Library Hi Tech.*

# Appendix

*import pandas as pd*

*import numpy as np*

*from datetime import datetime*

*import matplotlib.pyplot as plt*

*%matplotlib inline*

*table = pd.read\_csv("D:\Lectures\England\DAV\Report\dataset\call\_log\call\_log\_u59.csv")*

*#table[['timestamp']].head()*

*for row in table.iterrows():*

*table['timestamp'] = table['timestamp'].replace(row[1][2],datetime.fromtimestamp(row[1][2]))*

*table=table.drop(columns= ['id', 'device', 'CALLS\_name', 'CALLS\_number', 'CALLS\_numbertype', 'CALLS\_type'])*

*date\_list = table*

*processed\_data= date\_list.drop(['CALLS\_\_id', 'CALLS\_numberlabel'], axis=1)*

*df\_nonan = processed\_data.dropna()*

*df\_nonan['CALLS\_date'] = df\_nonan['CALLS\_date'].apply(pd.to\_datetime, unit='ms')*

*for\_datenduration = df\_nonan*

*pd.to\_datetime(for\_datenduration['CALLS\_date'])*

*for\_datenduration*

Table

Description automatically generated

49 files are processed in this way and then combined by this code:

*import os*

*import glob*

*import pandas as pd*

*#working directory*

*os.chdir(r"C:\Users\muhad\DAV\formerge\Phone\_Lock")*

*#find all csv files in the folder*

*#use glob pattern matching -> extension = 'csv'*

*#save result in list -> all\_filenames*

*extension = 'csv'*

*all\_filenames = [i for i in glob.glob('\*.{}'.format(extension))]*

*#print(all\_filenames)*

*#combine all files in the list*

*combined\_csv = pd.concat([pd.read\_csv(f) for f in all\_filenames ])*

*#export to csv*

*combined\_csv.to\_csv( "Call\_Combined.csv", index=False, encoding='utf-8-sig')*

The output contains around 6400 rows like this:

Table

Description automatically generated

Same kind of processing is done for Phone\_Lock, SMS, Phone\_Charge, App usage data.

CO2 calculation from phone calls:

*call\_made = pd.read\_csv(r'C:\Users\muhad\DAV\Level\_2\call\_made.csv')*

*call\_made['CO2\_Emission'] = (call\_made['CALLS\_duration']/60)\*0.06*

*Table

Description automatically generated*

Phone Charge:

*table =pd.read\_csv("D:\Lectures\England\DAV\Report\dataset\sensing\phonecharge\phonecharge\_u01.csv")*

*Table

Description automatically generated*

*for row in table.iterrows():*

*table['start'] = table['start'].replace(row[1][0],datetime.fromtimestamp(row[1][0]))*

*for row in table.iterrows():*

*table['end'] = table['end'].replace(row[1][1],datetime.fromtimestamp(row[1][1]))*

*table['Charge Duration']= table['end']- table['start']*

*Graphical user interface, text

Description automatically generated*

SMS:

*table = pd.read\_csv(r"C:\Users\muhad\DAV\formerge\SMS\SMS\_Combined.csv")*

*table['CO2 Emission 10 Weeks'] = ((table['SMS Sent']\*0.014)/1000)*

*myplot = table['CO2 Emission 10 Weeks'].plot(figsize=(15, 10), kind='line', rot=90, color = "blue", title = "A Graph of Users against their CO2 Emission from sent SMS")*

*#myplot.figure(figsize=(3, 3))*

*myplot.set\_xlabel("Students")*

*myplot.set\_ylabel("Co2 Emission in 10 Weeks")*

*plt.savefig('user-wise emission from SMS.jpg')*

*table['CO2 Emission 10 Weeks'].cumsum()*

Phone Lock:

*table = pd.read\_csv("D:\Lectures\England\DAV\Report\dataset\sensing\phonelock\phonelock\_u00.csv")*

*#table.head()*

*for row in table.iterrows():*

*table['start'] = table['start'].replace(row[1][0],datetime.fromtimestamp(row[1][0]))*

*table['end'] = table['end'].replace(row[1][1],datetime.fromtimestamp(row[1][1]))*

*table['Time in Locked Mode'] = table['end'] - table['start']*

*table['Time in Locked Mode'] = table['Time in Locked Mode'].dt.total\_seconds()*

*#pd.to\_datetime(table['Time in Locked Mode'])*

*#type(table['Time in Locked Mode'])*

*#table.head()*

*x= table['Time in Locked Mode']*

*#print(table.dtypes)*

*time\_in\_sec= []*

*for i in range(table['start'].count()):*

*time\_in\_sec.append(x[i])*

*#print(x[i].total\_seconds())*

*Text, table

Description automatically generated with medium confidence*

*table['Phone Usage'] = 24 - table['Time in Locked Mode']*

*Table

Description automatically generated*

Correlation:

*table = pd.read\_csv(r"C:\Users\muhad\DAV\Level\_2\Final DataSet.csv")*

*table.head()*

*df = table.dropna()*

*df.head()*

*df.corr(method='pearson', min\_periods=1)*

*Table

Description automatically generated*

*import seaborn as sn*

*import matplotlib.pyplot as plt*

*x = df.corr()*

*sn.heatmap(x,annot=True,fmt="f",*

*xticklabels=x.columns,*

*yticklabels=x.columns)*

*plt.show()*

*from scipy import stats*

*pearsonCoeff\_rvalue, p\_value = stats.pearsonr(table["Phone\_Usage"], df["Grades"])*

*print("Pearson Correlation Coefficient r value : ", pearsonCoeff\_rvalue.round(decimals=3), "and P-value: ", p\_value.round(decimals =3))*

*alpha = 0.05*

*alpha\_half = 0.025*

*if p\_value < alpha\_half:*

*print("Conclusion drawn: The null hypothesis can be rejected")*

*else:*

*print("Conclusion drawn: The null hypothesis is accepted")*

*from scipy import stats*

*r, p = stats.pearsonr(df["Time\_Spent\_On\_Entertainment\_App"], df["Phone\_Usage"])*

*print("Pearson Correlation Coefficient r value : ", r.round(decimals=3), "and P-value: ", p.round(decimals =3))*

Linear Regression:

*import matplotlib.pyplot as plt*

*import numpy as np*

*import pandas as pd*

*from sklearn import datasets, linear\_model*

*from sklearn.linear\_model import LinearRegression*

*from sklearn.model\_selection import train\_test\_split*

*from sklearn.metrics import mean\_squared\_error, r2\_score*

*table = pd.read\_csv(r"C:\Users\muhad\DAV\Level\_2\Final DataSet.csv")*

*grades\_x = table[['Time\_Spent\_On\_Entertainment\_App']]*

*grades\_y = table[['Phone\_Usage']]*

*plt.scatter(grades\_x, grades\_y)*

*x\_train,x\_test,y\_train,y\_test=train\_test\_split(grades\_x, grades\_y,test\_size=0.2)*

*regr = LinearRegression()*

*regr.fit(x\_train, y\_train)*

*#y\_pred = regr.predict(x\_test)*

*#y\_pred2 = regr.predict([[900]])*

*#print("\n")*

*#print("The predicted Grade for a student who used phone for 900 hours is: ", y\_pred2)*

*print('Coefficients: \n', regr.coef\_)*

*print("\n")*

*print('Mean squared error: ', mean\_squared\_error(y\_test, y\_pred))*

*print('Coefficient of determination: ', r2\_score(y\_test, y\_pred))*

*plt.scatter(x\_test, y\_test, color='black')*

*plt.plot(x\_test, y\_pred, color='blue', linewidth=3)*

*plt.xticks(())*

*plt.yticks(())*

*plt.xlabel("Loneliness Scale", labelpad=10, fontsize=15)*

*plt.ylabel("Confidence Level", labelpad=10, fontsize=15)*

*plt.show()*

Polynomial Regression:

*import seaborn as sns*

*import pandas as pd*

*import matplotlib.pyplot as plt*

*import numpy as np*

*from sklearn.model\_selection import train\_test\_split*

*from sklearn.linear\_model import LinearRegression*

*from sklearn import metrics*

*from sklearn.metrics import r2\_score*

*from sklearn.preprocessing import PolynomialFeatures*

*import statsmodels.api as sm*

*import statsmodels.formula.api as smf*

*X = table[['Loneliness\_Scale']]*

*X1 = table['Loneliness\_Scale']*

*y = table[['Confidence\_Scale']]*

*poly\_reg = PolynomialFeatures(degree=4)*

*X\_poly = poly\_reg.fit\_transform(X)*

*pol\_reg = LinearRegression()*

*model = pol\_reg.fit(X\_poly, y)*

*y\_pred = model.predict(X\_poly)*

*reg\_label = "Coefficients:%s - b:%0.2f" % \*

*(np.array2string(model.coef\_,*

*formatter={'float\_kind': lambda fk: "%.3f" % fk}),*

*model.intercept\_)*

*print(reg\_label)*

*print("\n")*

*print("The model coefficients are: for degree 4 - x^0, x^1, x^2, x^3, X^4 ")*

*print(model.coef\_)*

*print("The model intercepts are ", model.intercept\_)*

*fig, ax = plt.subplots(1,2, figsize = (10,5))*

*sns.scatterplot(data=table, x = "Loneliness\_Scale", y = "Confidence\_Scale", ax = ax[0])*

*sns.regplot(data = table, x = "Loneliness\_Scale", y = "Confidence\_Scale", ax = ax[1])*

*sns.lmplot(x="Loneliness\_Scale", y="Confidence\_Scale", data=table)*

*sns.lmplot(x="Loneliness\_Scale", y="Confidence\_Scale", data=table, order = 2)*

*sns.lmplot(x="Loneliness\_Scale", y="Confidence\_Scale", data=table, order = 5)*

*#o5.savefig("Abc.jpg")*

*#Visualise Polynomial Regression Graphs*

*fig1, ax1 = plt.subplots()*

*ax1.scatter(X1,y, color = 'green')*

*ax1.plot(X1, y\_pred, color = 'red')*

*ax1.scatter(X1, y\_pred, color = 'black', marker = 'x')*

*ax1.set\_title("Polynomial Regression Model Plot")*

*ax1.set\_xlabel("Loneliness\_Scale")*

*ax1.set\_ylabel("Confidence\_Scale")*

*lm1 = smf.ols(formula='Grades~ Loneliness\_Scale + Loneliness\_Scale\*\*2 + Loneliness\_Scale\*\*3 + Loneliness\_Scale\*\*4', data=table).fit()*

*print(lm1.params)*

*print("\n")*

*print("The model coefficients are: for degree 4 - x^0, x^1, x^2, x^3, X^4 ")*

*print(model.coef\_)*

*print("The model intercepts are ", model.intercept\_)*

*print("Compare the results of statsmodel and sk-learn for polynomial regression model")*

*print("\n")*

*print('Mean Absolute Error, MAE is ', metrics.mean\_absolute\_error(y, y\_pred))*

*print('Mean Squared Error, MSE is ', metrics.mean\_squared\_error(y, y\_pred))*

*print('Root Mean Square Error is ', np.sqrt(metrics.mean\_squared\_error(y, y\_pred)))*

*print("\n")*

*#Rsquared value of model*

*print("Coefficient of determination or R squared value is ")*

*print(r2\_score(y, y\_pred))*

*print("\n")*

Multiple Variable Linear Regression:

*import matplotlib.pyplot as plt*

*import numpy as np*

*import pandas as pd*

*import statsmodels.api as sm*

*from sklearn import datasets, linear\_model*

*from sklearn.linear\_model import LinearRegression*

*from sklearn.model\_selection import train\_test\_split*

*from sklearn.metrics import mean\_squared\_error, r2\_score*

*#table = pd.read\_csv(r'C:\Users\muhad\DAV\Level\_2\Final DataSet.csv')*

*table=table.drop('User',axis=1)*

*table.head()*

*Application

Description automatically generated with low confidence*

*reg = linear\_model.LinearRegression()*

*data\_x = table[['Phone\_Usage','Time\_Spent\_On\_Entertainment\_App', 'Time\_Spent\_On\_Educational\_App', 'Health\_Fitness\_App', 'Loneliness\_Scale', 'Confidence\_Scale']]*

*data\_y = table[['Grades']]*

*x\_train,x\_test,y\_train,y\_test=train\_test\_split(data\_x, data\_y,test\_size=0.2)*

*reg.fit(x\_train,y\_train)*

*y\_pred = reg.predict(x\_test)*

*reg.predict([[200,2,2,5,2,4]])*

*reg.coef\_*

*print('Mean squared error: ', mean\_squared\_error(y\_test, y\_pred))*

*print('Coefficient of determination: ', r2\_score(y\_test, y\_pred))*

*data\_x = sm.add\_constant(data\_x)*

*model = sm.OLS(data\_y, data\_x).fit()*

*predictions = model.predict(data\_x)*

*print\_model = model.summary()*

*print(print\_model)*

*import seaborn as sns*

*import pandas as pd*

*sns.set\_style()*

*g = sns.lmplot(*

*data=table,*

*x="Phone\_Usage", y="Grades", hue="Loneliness\_Scale",*

*height=5*

*)*

*g.set\_axis\_labels("Phone Usage", "Grades")*

*g = sns.PairGrid(table)*

*g.map(sns.scatterplot)*

*g.savefig("Multiple Linear Regression.jpg")*

*X = table[['Phone\_Usage','Loneliness\_Scale', ]]*

*X1 = table['Phone\_Usage']*

*X2 = table['Loneliness\_Scale']*

*y = table['Grades']*

*#Split Dataset to Train and Test Dataset*

*x\_train, x\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=0.10, random\_state=42)*

*X11 = table['Phone\_Usage']*

*X12 = table['Loneliness\_Scale']*

*fig = plt.figure(figsize = (30,10))*

*ax = fig.add\_subplot(1,2,1, projection = '3d')*

*ax.scatter3D(X1, X2, y, c=y, cmap='Greens')*

*ax.plot3D(X1, X2, y, 'gray')*

*# set up the axes for the second plot*

*ax = fig.add\_subplot(1, 2, 2, projection='3d')*

*ax.plot\_trisurf(X1, X2, y, cmap=plt.cm.viridis, linewidth=0.2)*

*plt.show()*

*ax.set\_xlabel('X1')*

*ax.set\_ylabel('X2')*

*ax.set\_zlabel('y')*

*#Build the model*

*ax = fig.add\_subplot(2, 2, 1, projection='3d')*

*for degree in [1,2,3,4,5,6,7]:*

*model = make\_pipeline(PolynomialFeatures(degree), LinearRegression())*

*model.fit(x\_train,y\_train)*

*y\_pred = model.predict(x\_test)*

*print("Return the coefficient of determination R^2 of the prediction, i.e. the model score is: ",model.score(x\_test, y\_test))*

*print("\n")*

*plt.plot(y\_test, y\_pred, label="degree %d" % degree)*

*plt.legend(loc='upper right')*

*plt.xlabel("Grades Test Data")*

*plt.ylabel("Predicted Grades")*