DSA 210 FINAL REPORT

**EDİZ KARLI**

**33806**

**MY SOCIAL MEDIA USAGE**

**INSTRUCTOR: SELİM BALCISOY**

**WHAT IS IN THE REPORT?**

The following report analyzes the relation between my personal social media usage and daily activities. To achieve this 1 month of data was gathered for Instagram, Reddit and Youtube Usage with additions of other daily activity data such as number of hours spent in school or in traffic. By using statistical methods and visualizations this work sought to predict my daily social media usage across different platforms and give an insight about the effects of my daily activities on my technology usage.

**PARAMETERS IN THE REPORT:**

**Date**: Data entry day

**Having a Homework or a Project Deadline in Upcoming three days**: YES or NO, will be marked by me.

**Amount of Time in School**: In terms of hours, I will be using Google Maps to check how many hours I stayed in school.

**Travel Time:** In terms of minutes or hours, will be using Google Maps to check it.

**Sleep Time:** In terms of hours, will be using a smartwatch to collect the data.

**Sleep Quality**: 1-10, I will be using the same smartwatch and same app to collect the data.

**Social Interaction**: In terms of numbers, it will start with 0 and with every non-family member who I talk more than 3 minutes I will add one to it.

**Reddit Usage**: Will be collected from the settings part of my smart phone

**Instagram Usage**: Will be collected from the settings part of my smart phone

**Youtube Usage:**  Will be collected from the settings part of my smart phone

**INTRODUCTION**

In this project I will be using my School Time, Travel Time, Sleep Time, Sleep Quality, Social Interaction amount and Homework/Project Deadline amount to predict my Reddit, Instagram and Youtube usage. Through statistical analysis and visualization, I’ll be identifying the key factors that contribute to social media usage. This report is the result of multiple data analysis and visualization technics.

**WHAT WAS DONE?**

From 14th of March to 13th of April I gathered the given parameters’ data on an excel page. I used Google Maps to learn how many hours I stayed in school and how many hours I traveled, I used a smart watch to get how many hours I slept and what was my sleep quality, I used my smart phone’s usage track settings to get the data of Instagram-Reddit-Youtube usage, I kept in mind with how many people I spoke in a day and marked Yes or No for Homework/Project Deadline to gather all the data.

It is important to note that during the EDA process I decided to analyze the input data with Reddit and Instagram Usage only and did not analyze or visualize Youtube Usage. This was because Youtube usage was not a part of my initial planning and it was added on after EDA process. From this point of view, we can say that this way we will be able to see if my model can predict a dataset which was not considered while building the model.

Then I inspected my data, it did not have any NA entries so I didn’t need to fill them. Also, I converted the hours to minutes in order to make it easier for calculation. After that I decided that my data was ready to be used for Exploratory Data Analysis.

Before starting EDA I also created the null hypothesizes:

**1-School and travel time do not affect social media usage.**

**2-Sleep time and quality do not affect social media usage.**

**3-Homework/project deadlines do not affect social media usage.**

**4-Social interaction does not affect social media usage.**

After identifying Null Hypothesizes, I analyzed and visualized the data with Correlation Analysis, T Test and Chi-Square Test. Again, I did these for the relation between input data and Reddit-Instagram Usage.

Once EDA was completed, I reviewed the results and saw if a given data input was related to Reddit or Instagram Usage. With these results I rejected or failed to reject the null hypothesizes I had in my hand. The next step was Machine Learning process.

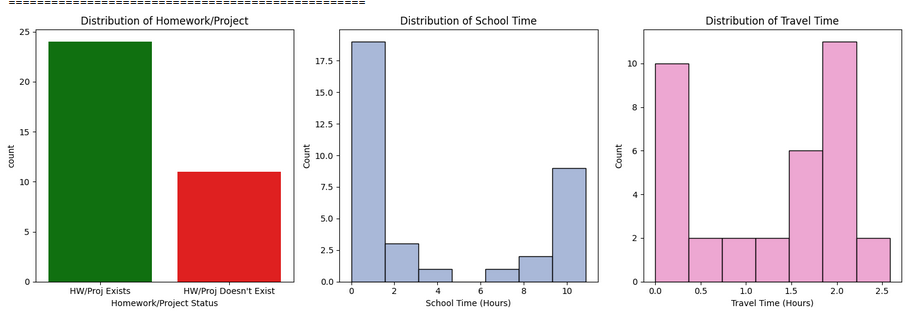
In Machine Learning process I tried to predict my Instagram, Reddit and Youtube usage. Before starting it I had rough ideas about the relationship of my data with Reddit and Instagram Usage but I didn’t know the relationship of my data and Youtube Usage. I used Linear Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), KNN Regression and Gradient Boosting separately for each of them by using the feature data I had in my hand.

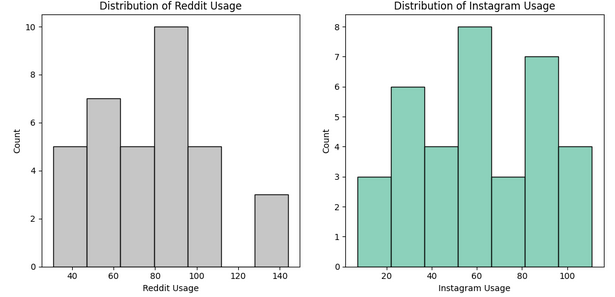
After checking the results of Machine Learning process and predicting, I decided whether my data was able to predict or not and if not, what could be the reason.

**1-GRAPHS AND VISUALIZATIONS**

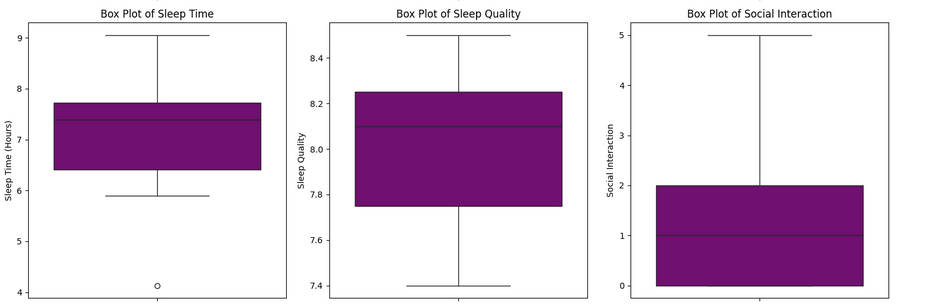
First, I wanted to see my data without any interpretation and used Histogram, Boxplot and Scatter Plot for visualizations. Here are some examples of them:

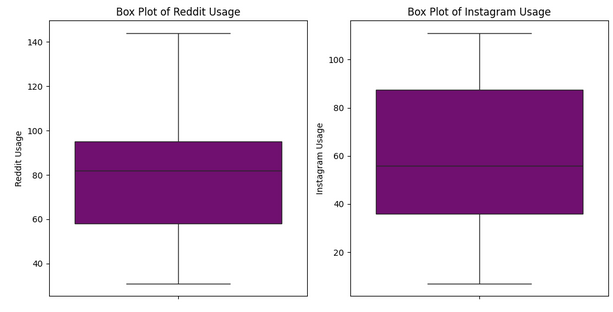
HISTOGRAM VISUALIZATIONS:



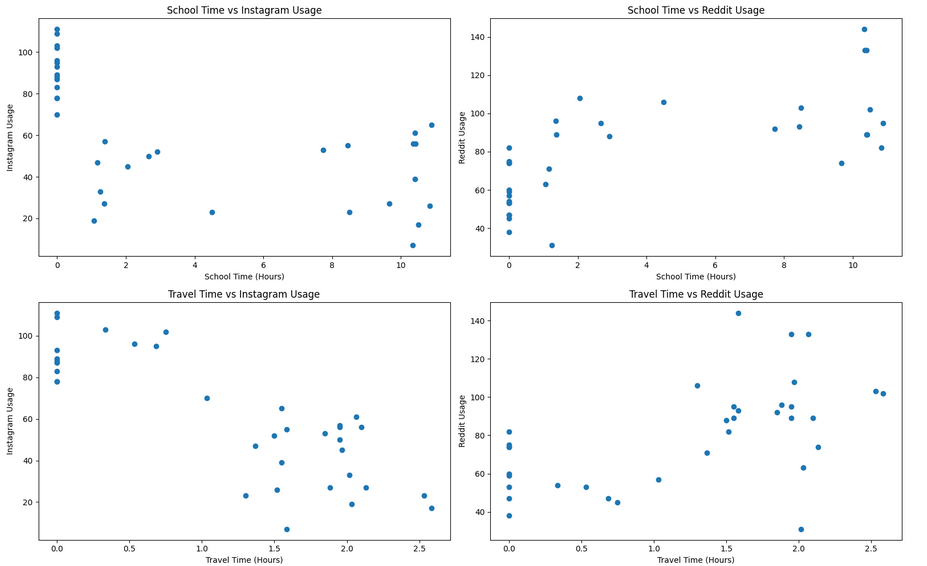


BOXPLOT VISUALIZATIONS:





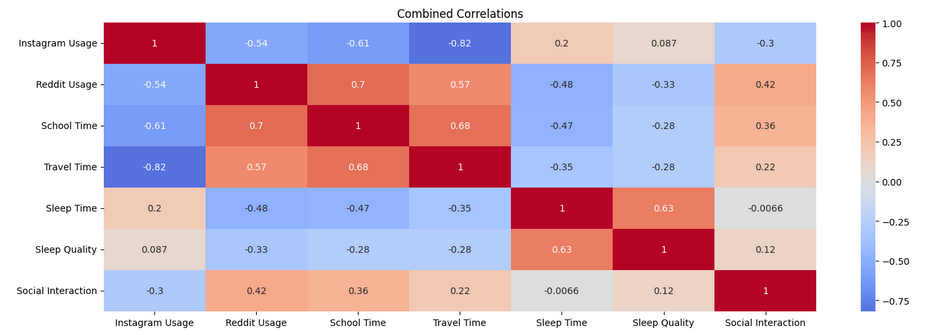
SCATTER PLOTS:



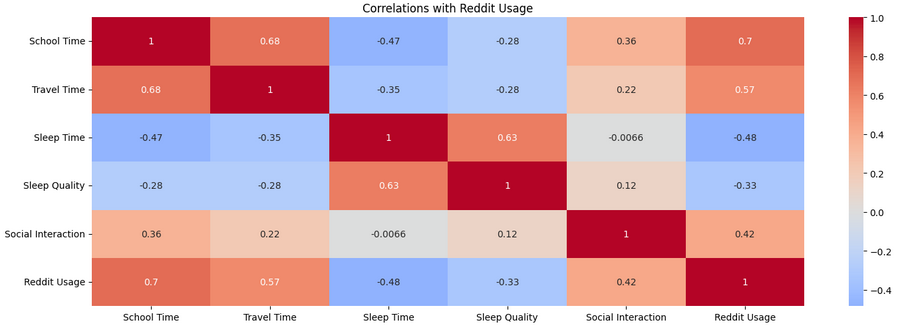
I used these visualizations to understand the constraints and limits of my data and applied analysis methods according to them.

**2-CORRELATION, T-TEST AND CHI-SQUARE TEST**

ALL CORRELATIONS:



REDDIT USAGE:



My Reddit Usage data’s highest correlation in terms of magnitude was School Time and Travel Time (0.7 and 0.57 respectively). They both had a very small p-value (smaller than 0.0000 for both of them) suggesting that we can reject the null hypothesis of school and travel time not having an effect on social media usage. This makes sense in a way, if we look at all correlations we can see that travel time and school time has a correlation of 0.68 indicating they are related with each thus they are affecting Reddit Usage in the same way which is a positive correlation. Also if we look at the t-test and chi-square test results we can see that the relation is significant:

School Time vs Reddit = (Chi2 statistic: 20.867 / P-value: 0.000) || (T-statistic: 6.805 / P-value: 0.000)

Travel Time vs Reddit = (Chi2 statistic: 15.125 / P-value: 0.000) || (T-statistic: 4.096 / P-value: 0.000)

Since correlation results are positive, we can say that Reddit Usage and School Time-Travel Time are positively related in a significant way.

Although not as high as School and Travel Time combined, Sleep Time and Sleep Quality also has a relation with Reddit Usage. For Sleep Time correlation value is -0.48 and for Sleep Quality it is -0.33. Their p-values are 0.003 and 0.055 respectively. For sleep quality p-value is bigger than 0.05 by a small margin so for correlation analysis we fail to reject null hypothesis but if we look at T-test and Chi-Square Test:

Sleep Time vs Reddit= (Chi2 statistic: 2.378 / P-value: 0.123) || (T-statistic: -3.333 / P-value: 0.002)

Sleep Quality vs Reddit= (Chi2 statistic: 5.322 / P-value: 0.021) || (T-statistic: -2.099 / P-value: 0.044)

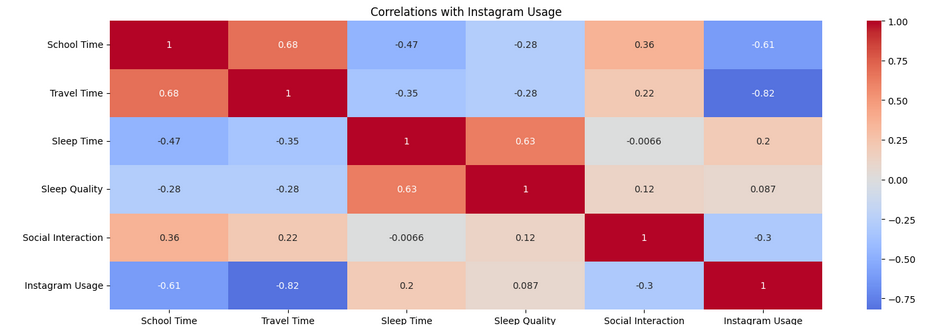
We can say that there is significant relationship between Sleep Quality and Reddit Usage when T-Test and Chi-Square test is applied. But also when Chi-Square is applied to Sleep Time vs Reddit there isn’t. So for both data groups there is at least one method that gives not significant result but most of the methods give significant relation. It is hard to give a direct reason for this and also it is hard to interpret the results. But overall if we accept certain methods and their results we can say that there is negative correlation between Sleep Time-Sleep Quality and Reddit Usage. According to this we can also say we can reject the null hypothesis.

For Social Interaction the correlation value is 0.423 and its p-value is 0.011, for T-test and Chi-Square Test:

Social Interaction vs Reddit = (Chi2 statistic: 3.263 / P-value: 0.071) || (T-statistic: 2.860 / P-value: 0.007)

From here we can see that for correlation graph and T-Test there is significant relation. According to this we can say that Reddit Usage and Social Interaction has a positive relation.

INSTAGRAM USAGE:



Highest correlation values for Instagram data are School Time-Travel Time features. Their correlation values are -0.612 / -0.822 and their p-values are 0.000/ 0.000 respectively. For T-Test and Chi-Square Test:

School Time vs Instagram= (Chi2 statistic: 10.362 / P-value: 0.001) || (T-statistic: -4.529 / P-value: 0.000)

Travel Time vs Instagram= (Chi2 statistic: 10.362 / P-value: 0.001) || (T-statistic: -5.073 / P-value: 0.000)

As can be seen from the results we can say that there is a significant negative relationship between School Time-Travel Time and Instagram. This is actually the opposite of the result we found for the relationship between School Time-Travel Time and Reddit Usage indicating these features affect Reddit and Instagram Usage in a different way.

For Sleep Time-Sleep Quality the correlation values are 0.201 / 0.087 and their p-values are 0.248 / 0.618. For the correlation graph both of the p-values are higher than 0.05 so we fail to reject null hypothesis here. If we check T-Test and Chi-Square Test:

Sleep Time vs Instagram= (Chi2 statistic: 2.303 / P-value: 0.129) || (T-statistic: 1.941 / P-value: 0.061)

Sleep Quality vs Instagram= (Chi2 statistic: 0.232 / P-value: 0.630) || (T-statistic: 0.702 / P-value: 0.487)

If we look at these methods as well, we can see all of them indicates that there were no significant relations found. So from here we fail to reject the null hypothesis for Instagram Usage that said “Sleep time and quality do not affect social media usage.”

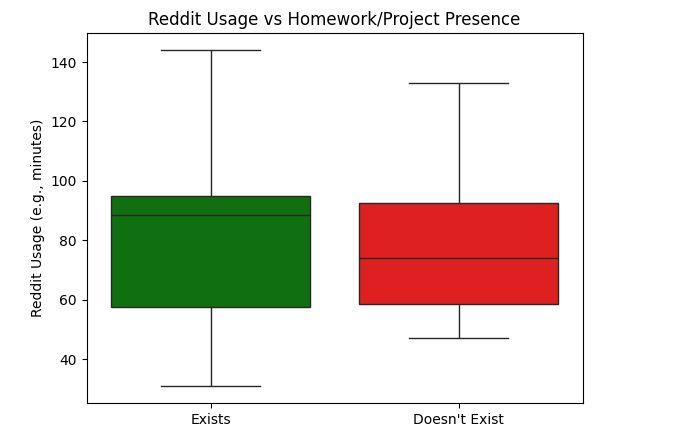
For Social Interaction the correlation value is -0.3 and its p-value is 0.076. For T-Test and Chi-Square Test:

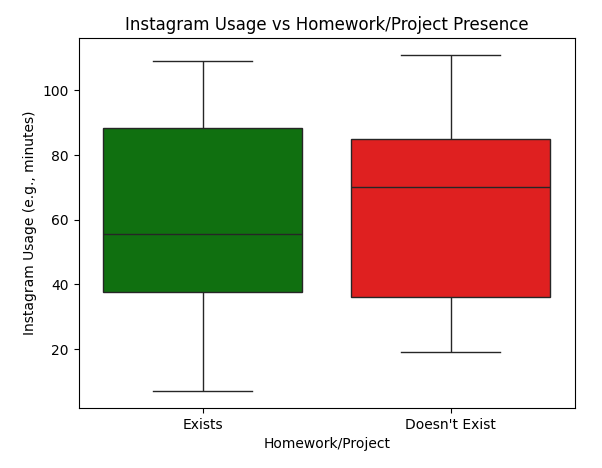
Social Interaction vs Instagram= (Chi2 statistic: 1.802 / P-value: 0.179) || (T-statistic: -2.022 / P-value: 0.051)

All of the methods indicate that there was no significant result between Social Interaction and Instagram Usage. So we fail to reject the null hypothesis of “Social interaction does not affect social media usage.” For Instagram Usage.

Homework/Project Amount:

Since homework/project data consisted of only 0’s and 1’s indicating Yes or No it was not possible to apply correlation graph, T-Test and Chi-Square Test for it. Instead I created boxplots to understand the relation between it and Reddit-Instagram Usage:





These graphs can’t really say much about the significance of the relation between Homework/Project Amount and Social Media Usage but they can at least give and idea about it. As can be seen from the graphs, when I have more Homeworks or Projects I used Reddit more frequently and when I didn’t, I used Instagram. Again, we can’t talk about the significance of this relationship but we can interpret that the similar path of having an opposite effect on Reddit and Instagram usage if there is a relation can also be seen here.

INSIGHTS AND REMARKS ABOUT EDA:

1-Reddit Usage had significant relationship with all the data sets I had, meaning Reddit Usage was highly influenced by my daily structed routine.

2-Reddit appears to be used more during the days I spent more time in school and ultimately traveling. Also from my interpretation I can say that the reason why my sleep time has a negative relation with Reddit Usage is because I sleep less when I have to spend more time in school and since I sleep less my sleep quality is also low on those days.

3-Social Interaction has a positive relation with it, I think this is because I have more interaction with other people when I spend more time in school.

4-Instagram Usage had significant relationship only with School Time-Travel Time.

5-Although it can be said that Instagram is being used less when I spend more time at school, since other data sets didn’t show significant relation with Instagram Usage it is open to other interpretations.

6-Maybe the dataset I have in hand doesn’t really explain my Instagram Usage or maybe my Instagram Usage is only depended on time I spend at school and travel. To answer these questions further observations should be made on the data or maybe new datasets should be added.

**3-MACHINE LEARNING**

In EDA part I only analyzed my dataset with Reddit and Instagram Usage. In Machine Learning Part I also added my Youtube Usage data to analysis. Here my purpose was to see whether my data would be able to understand a dataset I didn’t think of while collecting my data. And indeed, while I was collecting my data I didn’t plan to do anything with Youtube Usage, it was added later on.

I used Linear Regression, Decision Tree, Random Forest, Support Vector Machine, KNN and Gradient Boosting for Youtube, Reddit and Instagram Usage prediction:

1-YOUTUBE USAGE:

Linear Regression: (RMSE: 50.70 / MAE: 38.86 / R² Score: 0.21)

Decision Tree Regression: (RMSE: 47.85 / MAE: 39.00 / R² Score: 0.30)

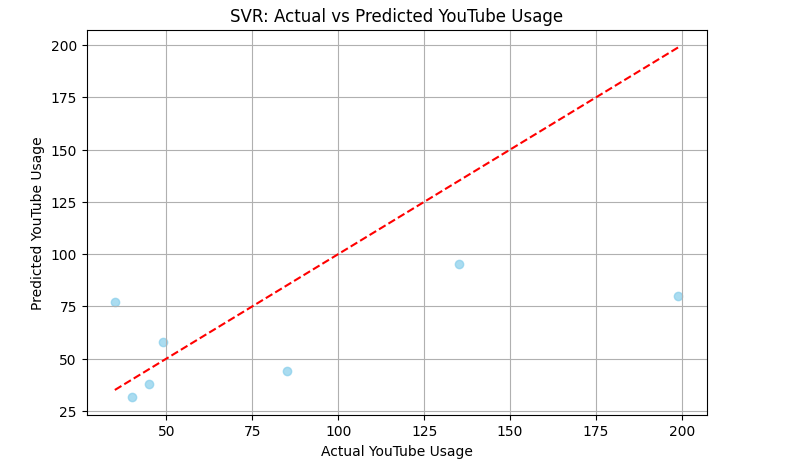
Random Forest Regression: (RMSE: 46.95 / MAE: 34.72 / R² Score: 0.33)

Support Vector Regression: (RMSE: 52.65 / MAE: 38.11 / R² Score: 0.15)

KNN Regression: (RMSE: 47.04 / MAE: 38.94 / R² Score: 0.32)

Gradient Boosting: (RMSE: 49.10 / MAE: 39.55 / R² Score: 0.26)

Here the best prediction result we have is for Support Vector Regression:



Here my Youtube Usage is best predicted by Support Vector Regression since it has the lowest error results for RMSE and MAE and has the highest R square value. It achieves the best balance between low error rates and variance explanation. KNN and Decision Tree also show similar results but they have higher error rates and lower R square value.

However, none of the models explain the data significantly, so from here we can say that maybe Youtube data is affected by other data sets I don’t have in my data or the relationship between Youtube usage and the indicators in my data is more complex than I think.

2-INSTAGRAM USAGE:

Instagram Usage - Linear Regression: (RMSE: 20.58 / MAE: 17.94 / R² Score: 0.27)

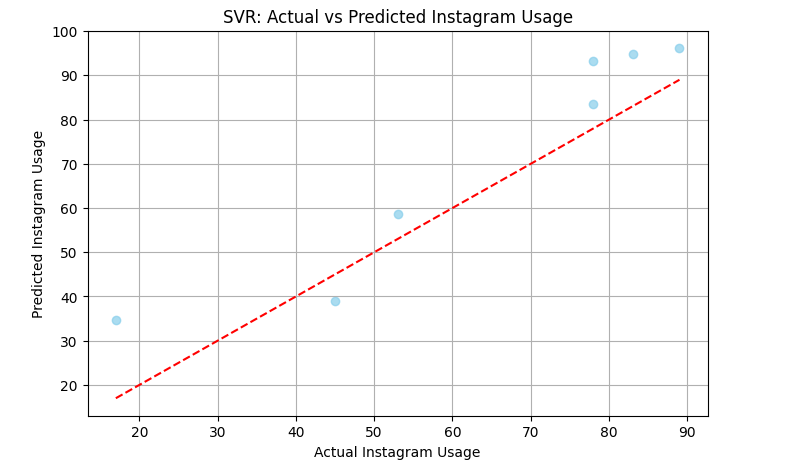
Instagram Usage - Decision Tree: (RMSE: 17.94 / MAE: 14.00 / R² Score: 0.45)

Instagram Usage - Random Forest: (RMSE: 15.97 / MAE: 14.41 / R² Score: 0.56)

Instagram Usage - Support Vector Regression: (RMSE: 10.92 / MAE: 9.88 / R² Score: 0.79)

Instagram Usage - KNN Regression: (RMSE: 15.72 / MAE: 14.43 / R² Score: 0.57)

Instagram Usage - Gradient Boosting: (RMSE: 16.84 / MAE: 13.99 / R² Score: 0.51)



Here if we look at the data, we can see that Support Vector Regression is the best model for predicting Instagram Usage with the highest R squared and lowest errors results. It captures nonlinear patterns well and outperforms other models. Random Forest and KNN also performs strong, but not as much as SVR.

Compared to the results for YouTube Usage the models for Instagram Usage performed better likely because the data was first gathered to make comparison between Reddit and Instagram Usage and Youtube usage was not considered in the early stages of project planning.

3-REDDIT USAGE:

Reddit Usage - Linear Regression: (RMSE: 19.91 / MAE: 16.19 / R² Score: 0.17)

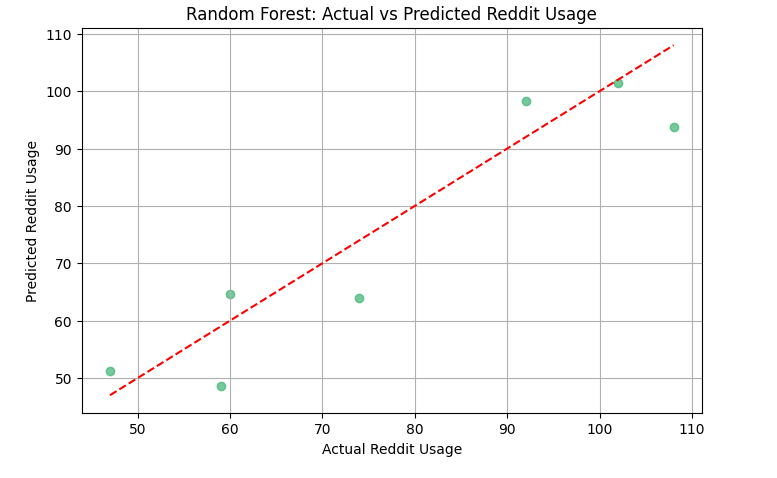
Reddit Usage - Decision Tree: (RMSE: 10.47 / MAE: 9.29 / R² Score: 0.77)

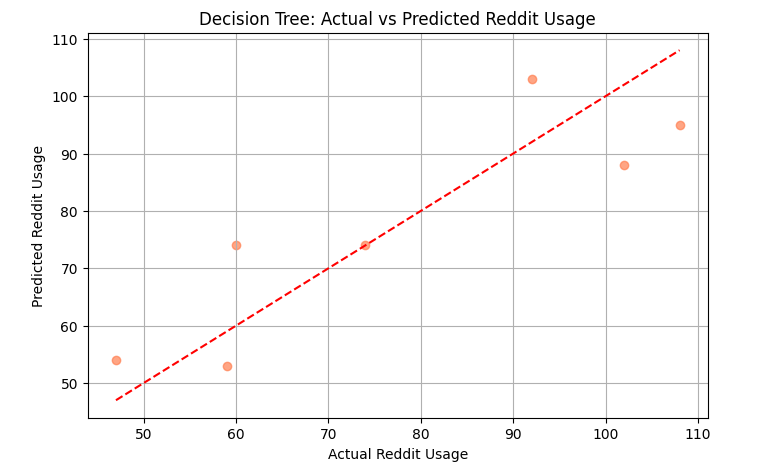
Reddit Usage - Random Forest: (RMSE: 8.33 / MAE: 7.16 / R² Score: 0.85)

Reddit Usage - Support Vector Regression: (RMSE: 17.02 / MAE: 11.67 / R² Score: 0.39)

Reddit Usage - KNN Regression: (RMSE: 15.41 / MAE: 12.14 / R² Score: 0.50)

Reddit Usage - Gradient Boosting: (RMSE: 12.05 / MAE: 10.80 / R² Score: 0.69)





Here if we look at the data, we can see that Random Forest Regression is the best model for Reddit Usage with the highest R squared and lowest error results. Decision Tree Regression also provides strong results with simpler structure. Since Linear Regression and SVR underperform, we can say that Reddit usage is influenced by nonlinear factors most of the time.

Out of all the result we had, the R square for Random Forest with 0.85 was the highest one. Suggesting that the data set I have is affective the most when we try to predict my Reddit Usage.

INSIGHTS AND REMARKS ABOUT MACHINE LEARNING:

1-The results suggest that YouTube usage is also driven by factors not included in my dataset. This makes sense since YouTube wasn’t considered in early project planning and when I decided on the features of my dataset I just thought about Reddit and Instagram Usage.

2-If we think in terms of the dataset I have in hand, we can say that YouTube usage may not be as routine based as Reddit and Instagram.

3-Instagram usage appears more structured, possibly related to features in my dataset.

4-The model confirms that Instagram usage is influenced by observable daily patterns, making it more predictable.

5-The results confirm that Reddit usage follows strong nonlinear patterns and is highly influenced by school and travel time which is consistent with my earlier correlation and hypothesis testing analysis.

6-My dataset is well suited for modeling Reddit usage.

7-Random Forest’s superior performance all throughout 3 social media dataset indicates complex interactions between variables.

**CONCLUSION AND WHAT COULD BE BETTER?**

We can say that this project successfully explored the relationship between daily routine factors and my social media usage, with a focus on Reddit and Instagram. Through a 30-day dataset, multiple results were uncovered using statistical tests, correlation analyses, and regression models. The analysis revealed:

* School and Travel Time significantly influence my social media habits. Instagram usage tends to decrease while Reddit usage increases on busier school days.
* Sleep Time and Sleep Quality have a noticeable effect on Reddit usage, although not consistently across all tests.
* Social Interaction shows a significant correlation with Reddit usage but not with Instagram.
* Homework and Project Deadlines did not exhibit statistically significant effects, although Reddit usage tended to rise during high workload periods based on boxplot visualization.

Using regression models, I extended the study to predict YouTube screen time. Models like Random Forest and Gradient Boosting performed the best, suggesting non-linear relationships in the data. The best prediction was made for Reddit and then Instagram. Youtube had the worst prediction but since I didn’t plan to use Youtube data at the beginning of the planning phase this makes sense.

To get better results for correlation and prediction we can:

1-Get a larger and more diverse dataset, maybe extending the 30 days to 60 or 90 days.

2-For Homework and Project data we can find a way to represent it quantitively rather than Yes or No only.

3-We can add new features like mood, stress or maybe we can divide the existing datasets into time zones such as morning-noon-night.

4-Additional machine learning models could be used especially classification models.

5-If we want to extend the research question, we can also focus on the social media context rather than only the number of hours I use it.

EDİZ KARLI 33806