**Data-science Project Documentation**

When working with huge real world data sets, we can find them messy and that’s why we first go through a process called Data wrangling, to organize and clean up the data and make it more suitable for creating regression, and classification models.

Data Wrangling steps:

1. We examine our datasets and profile it
2. Assess the Datasets quality
3. Clean the data (Handle missing values, convert data types, remove duplicates, … etc)

In this project we will be using 2 datasets.

Dataset 1: user behavior dataset

This dataset is from Kaggle. It explores the user behavior for 700 participants and explores the variables: Device model, Operating system, App usage time, Screen time, Battery drainage, and Data usage, Gender, and Age.

Why we choose this dataset?

This dataset is suitable for creating regression models because it contains multiple numerical variables that can be predicted, like App usage time(minutes), screen on time (hours), Battery Drain (percentage), Data usage (MBs), Numbers of apps installed. This dataset also has multiple predictors available such as:

* User characteristics: Age, Gender
* Device Attributes: Device model, operating system
* Usage statistics: App usage times, Screen on time, Data usage

This dataset also displays correlations between variables like:

* App usage time, and battery drain are positively correlated
* Screen on time, and app usage time are positively correlated
* Number of apps installed, and data usage seem to be positively correlated but not as strongly as the first two.

This dataset seems very fit for multiple different regression models.

Analyzing the dataset and cleaning the dataset:

After examining the dataset here are some initial notes:

* We find this dataset seems to be mostly consistent with its values
* Screen On time is > app usage time as it should be
* The dataset is mostly dominated by android users
* App usage time seem to be correlated with Battery drain and Data usage.
* Screen on time is also positively correlated with both App usage time and Battery drain.
* Number of apps Installed shows a mild positive correlation with Data usage.
* User behavior class correlates moderately with App usage time and Screen on time.
* User behavior class 1 and 5 are very few indicating that extreme behaviors such as minimal use, or excessive use are not that common.

After using the pandas library built in functions to further analyze this dataset here are some more notes:

* App usage time has a mean of 271.12, and a standard deviation of 177.19
* Screen on time has a mean of 5.27, and a standard deviation of 3.06
* Battery drain has a mean of 1525.16, and a standard deviation of 819.13
* Number of apps installed has a mean of 50.68, and a standard deviation of 26.94
* Data usage has a mean of 929.74, and a standard deviation of 640.45
* The mean, Std, min, max, and the quartiles for the remaining variables have also been found.
* Ran a check for null values (none were found)
* Ran a check for duplicates (none were found)
* After creating histograms with 100 bins for various variables, it seems that we have diverse range of ages, from 18 to 59.
* From the histogram App usage time seems to be mostly skewed to the left most users are moderate users. Also applies to screen on time.
* Looking at the correlation matrix and the heat map, most variables except for the user ID, and age seem to be strongly correlated to each other.
* The IQR method was used to find and flag outliers (none were found).

Fortunately, this dataset seems to be cleaned already so we don’t need to wrangle it.

Dataset 2: mobile addiction

This dataset is also from Kaggle. Investigates the addiction level of participants with a sample size of 13,600. With variables like daily screen time, app sessions, social media usage, gaming time, notifications, night usage, age, work study hours, stress level, apps installed, Addicted.

Why we choose this dataset?

This dataset seems suitable for creating classification models. Simply for example the Addicted column has two classes (addicted, and not addicted). The dataset also has a lot of numerical variables like Screen time, app sessions, notifications, social media usage, gaming time, apps installed, stress level, age, night usage, work study hours, and daily screen time. These variables are going to be useful for teaching the model classification boundaries. Furthermore, the number of samples is most likely going to guarantee that both groups are going to be well represented.

Analyzing the dataset and cleaning the dataset:

After taking an initial look on the dataset here are some notes:

* Daily screen time increases as app sessions increases.
* Stress levels increase as social media usage increases.
* Initial inspection shows varied screen times ranging from 2 to 9 hours and app sessions between 28 and 39.
* Addicted users tend to have higher screen time and more app sessions compared to non-addicted ones.
* The social media usage and gaming time columns show that addicted users often spend more time on these activities.
* Addicted users seem to have more frequent night usage and receive a higher number of notifications.
* Addicted users usually have higher stress levels and higher number of apps installed.
* Subjects who have minimal social media usage are usually not addicted.
* Subjects who are younger tend to have higher social media time.
* As the age decreases the likelihood of addiction increases.

After analyzing the file using pandas library here are some more notes:

* All columns have int data type except for the addicted column it is an object.
* Daily screen time has a mean of 3.76, and a standard deviation of 1.89
* App sessions has a mean of 30.03, and a standard deviation of 7.40
* Socia media usage has a mean of 1.54, and a standard deviation of 1.204
* Gaming time has a mean of 1.03, and a standard deviation of 0.993
* Work study hours has a mean of 5.97, and a standard deviation of 2.07
* The mean, Std, min, max, and the quartiles for the remaining variables have also been found.
* Ran a check for null values (none were found).
* Ran a check for duplicates (none were found).
* Created graphs to get a better understanding of the dataset.
* Created a heatmap and correlation matrix to see the correlations between variables.
* From the correlation matrix we see that daily screen time and app sessions are positively correlated.
* Social media usage and daily screen time are also positively correlated.
* Social media usage and app sessions are also positively correlated.
* Stress level and night usage shows a weak positive correlation.
* Tested for outliers using z-test method (271 outliers were found)
* Removed all outliers (dataset size is now 13,331)
* Performed the Anderson-Darling test for normality to see if the dataset follows normal distribution (It doesn’t).

Overall, this dataset shows a lot of potential for being used in multiple classification models.