LUT
Lappeenranta
University of Technology

A240A0050 Big Data in Business and Industry

PRACTICAL ASSINGMENT

Daniela Maldonado

Dilhara Liyanaaratchi

Shageerthana Sathiyamoorthy

Contents

Data preprocessing

Treating missing values

The data set has 19 variables with 3456 observations. Initially the missing values were checked variable wise and observations wise. And it was found out that variable hydration has the highest number of missing values which was 922 and variables start lag, start long, end lag and end long had 414 values were missing and ascend and descend variables had 411 missing values.

Table 1 Nan per Variable

hydration_l	922
start_lat	414
start_long	414
end_lat	414
end_long	414
speed_max_kmh	413
ascend_m	411
descend_m	411
altitude_min_m	408
altitude_max_m	408
calories_kcal	1

But after getting the grouped data based on the sport activity, it was found that we need to keep the observations, even more than there are 7 missing values if we removed those it will remove 6 different sport activity entirely. Hence those missing values were interpolated with zero as there were no prior knowledge in the data values and the rest of the missing values were interpolate with the grouped mean of the specific variables.

Figure 1 Mean of the Sport per Variable

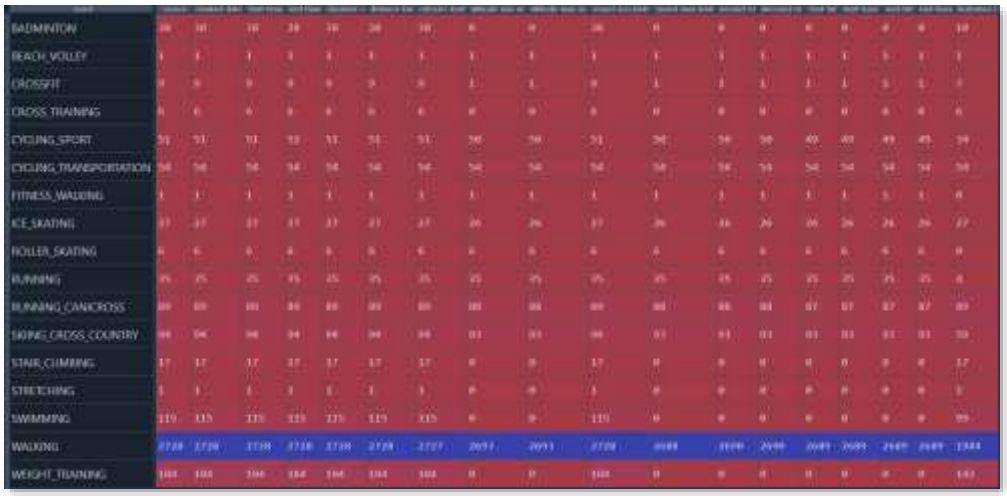


Figure 2 # of observation of the Sport per Variable

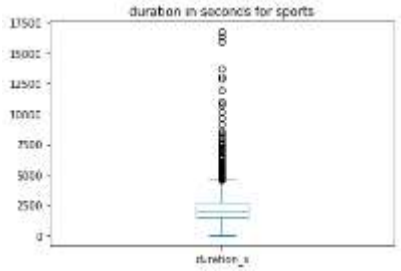
As per the above table the entire sport activity cross training, stretching, stair climbing might have to remove as there were lots of missing values in more than 7 variables in all the observations.

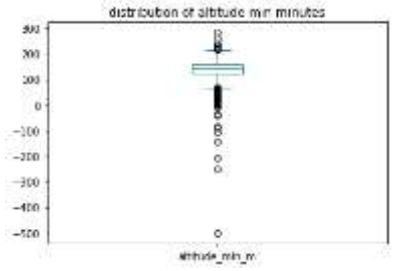
Therefore, at the end of the treatment for the missing values, none of the variables or observations were removed, and the whole data set were used for the explanatory analysis.

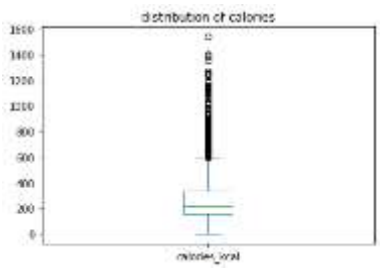
Explanatory Data Analysis

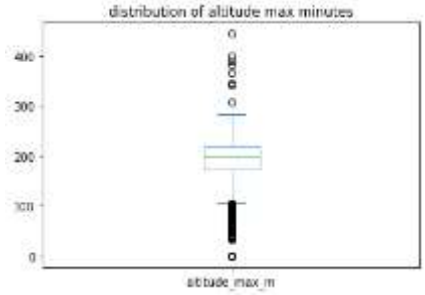
After the treatment for the explanatory data analysis was performed using graphic. For this purpose, box plots and histograms were used, and it is found that none of the variables followed a normal distribution and mostly the variables are skewed to right.

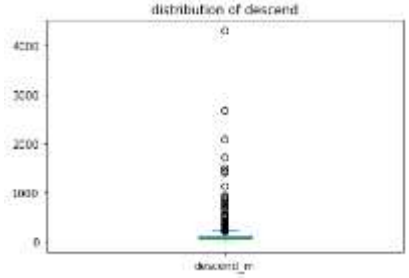
Distribution Boxplots

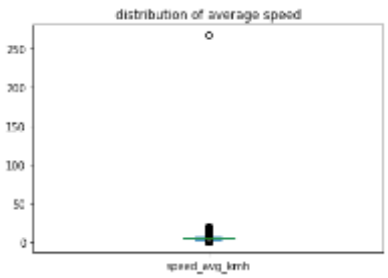












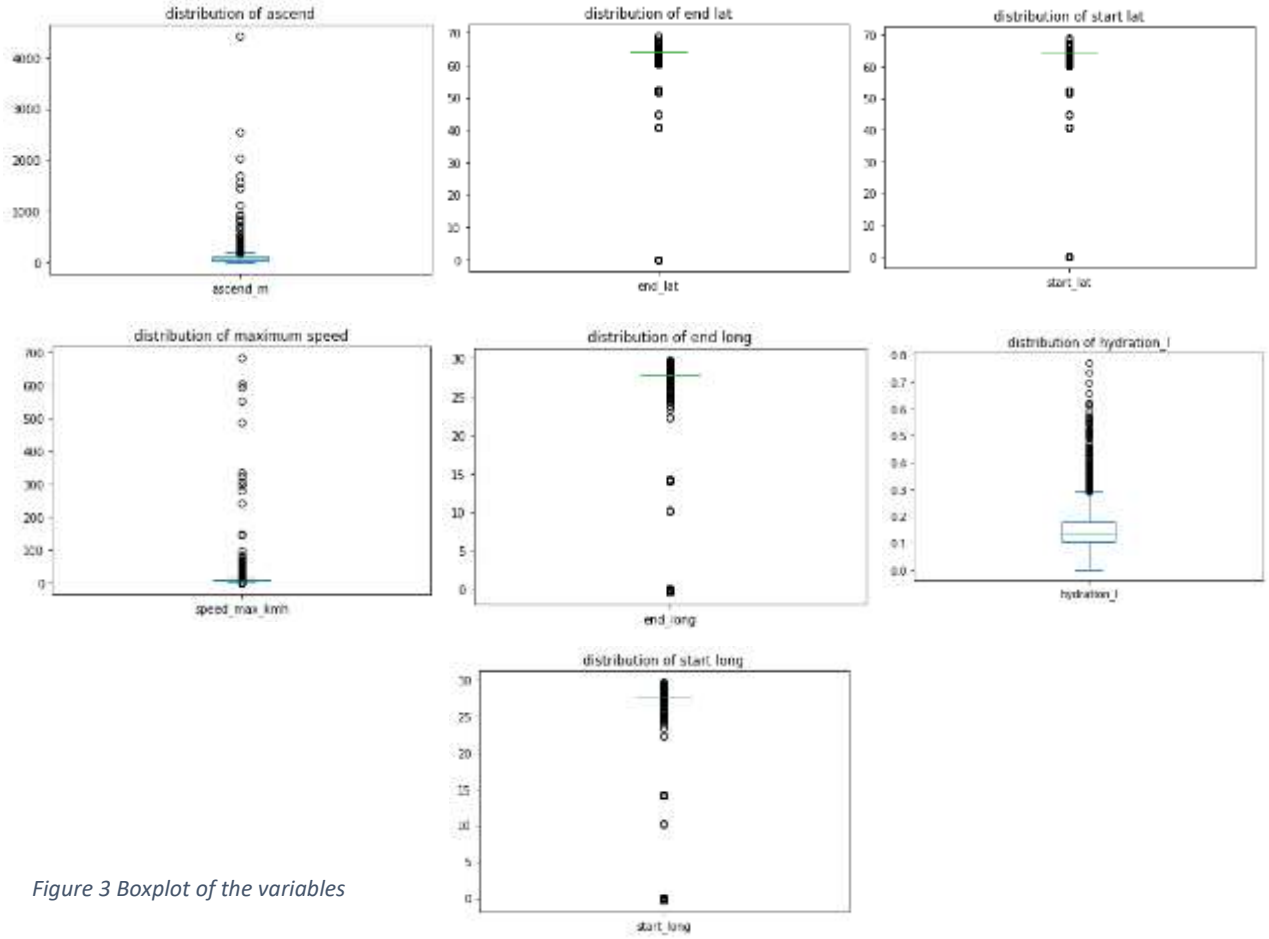


Figure 3 Boxplot of the variables

Occurence

As per the following graph, the occurrence of the sport activities was shown, and as per the most common activity is walking and the second most common activity is weight training and swimming, skiing cross country and running cani cross were placed 3rd, 4th and 5th place as per the occurrence.

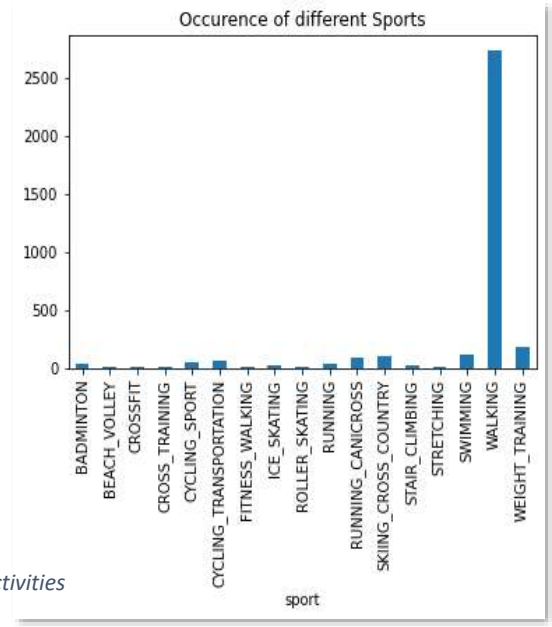


Figure 4 Occurrence of the sport activities

The activities were analyzed graphically with respect to the time to the day the activity started, and it seems like, mostly the sports were performed evening and the morning, and least number of activities were done after 9 at night and early morning around 2 o'clock.

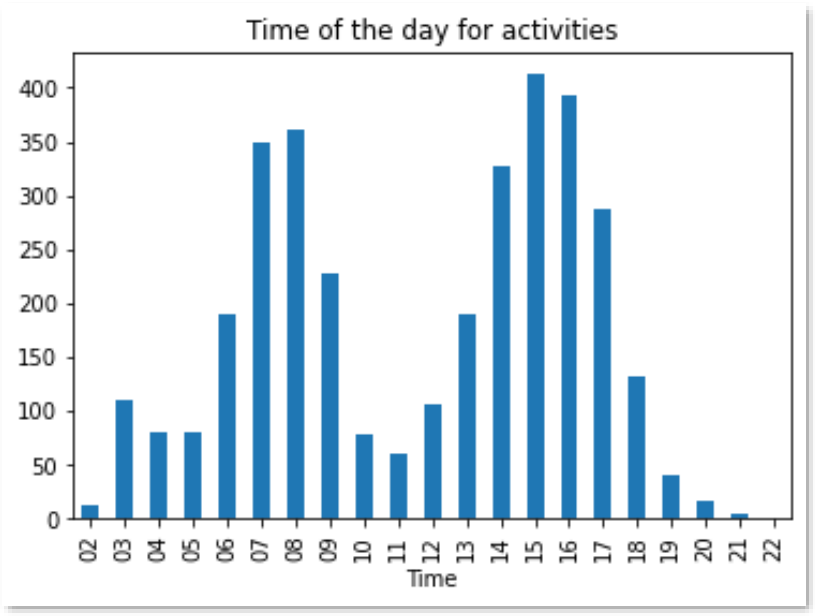


Figure 5 # of observation per Time of the day

The below pie chart shows the occurrence of the activities during the week. As per that, the activities were done equally throughout the week.

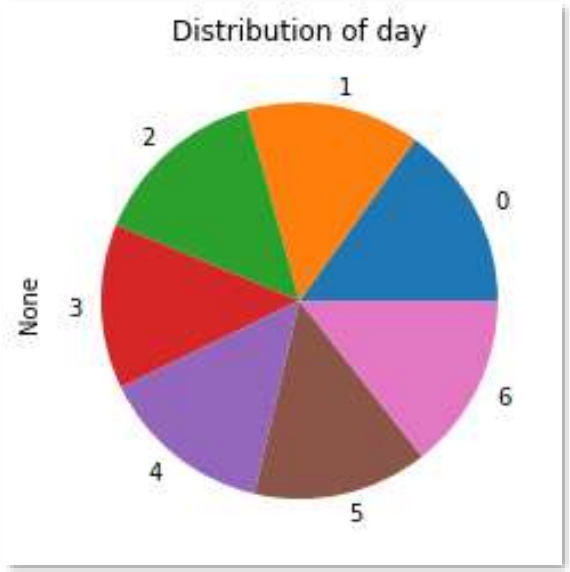


Figure 6 Observations regarding the day of the week

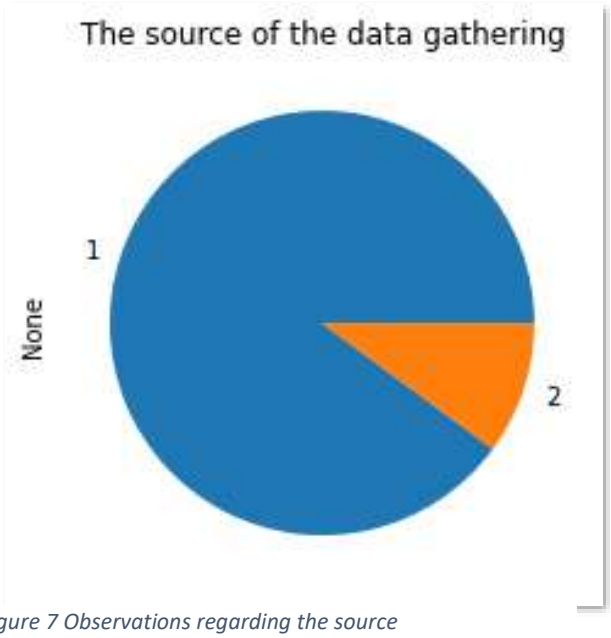


Figure 7 Observations regarding the source

Classification visualizations

The two methods of collecting the data were track mobile and input manually. And the most common method of data collection was done through tracking the mobile.

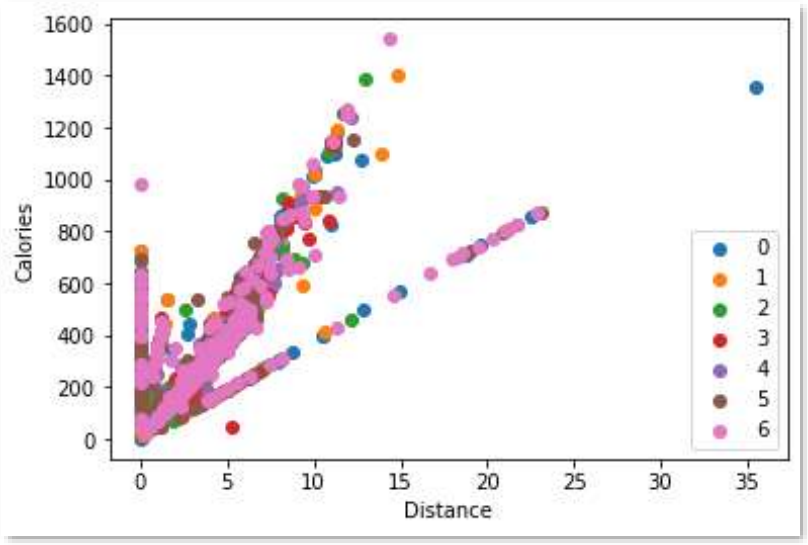


Figure 8 Distance (km) per day of the week vs calories

As per the above graph, we cannot see distinguish clusters to identify the distance vs the calories burned based on the day.

Further pair plots were performed based on the sport groups and the relationship between the variables were identified graphically.

And it is identified that hydration and calories variables are behaving the same with all the other variables.

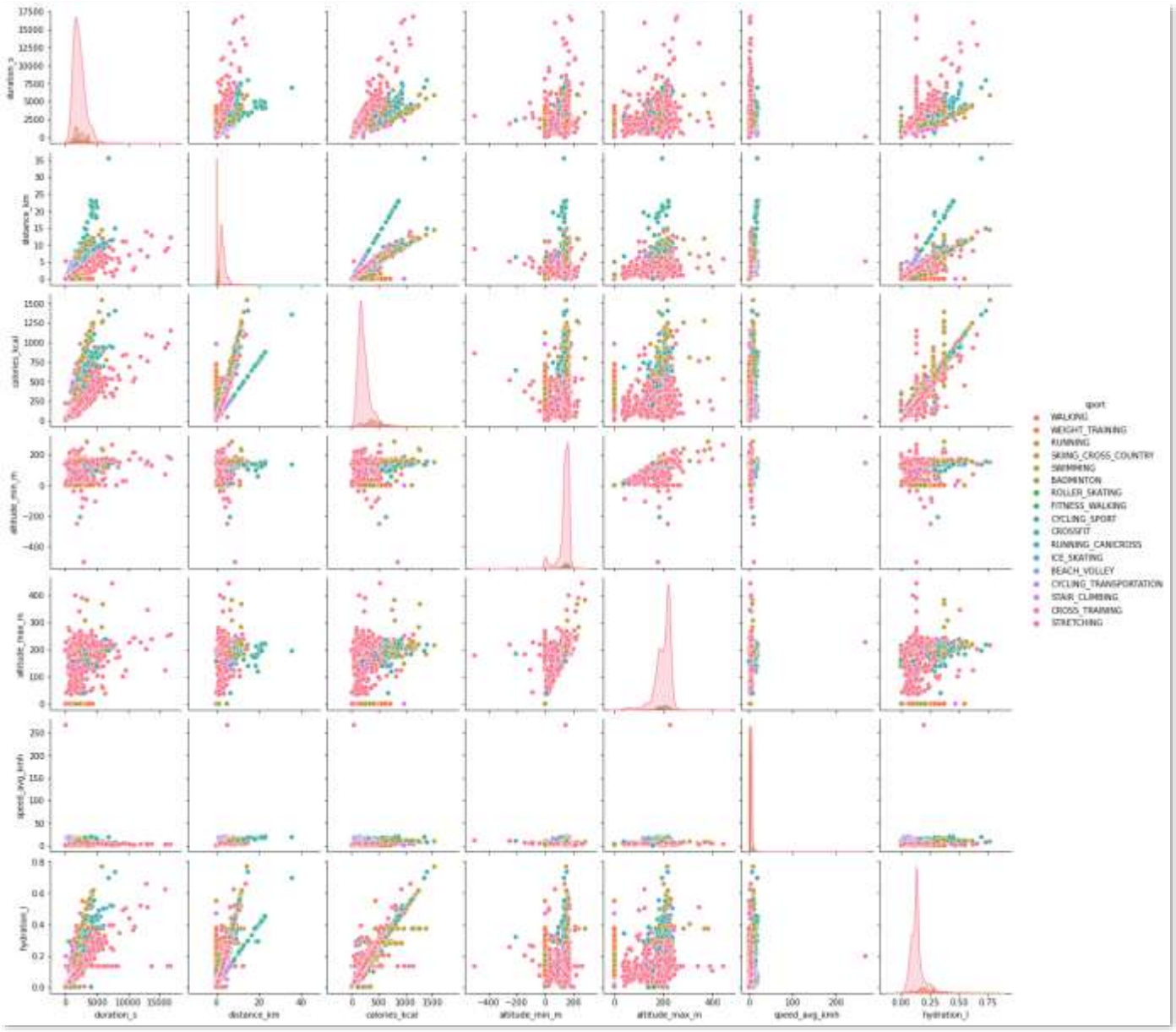
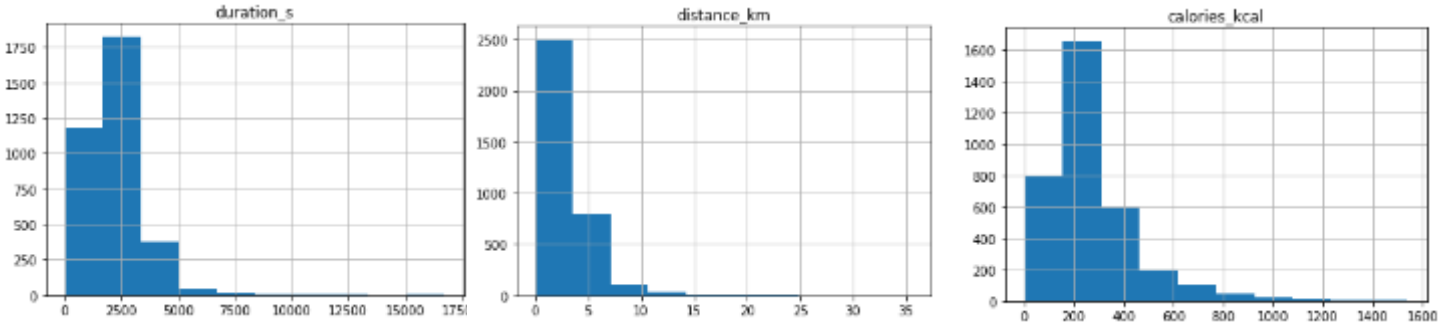


Figure 9 Clustering per variable vs variable

It is further observed that most of the variable's pairs such as, hydration and distance, hydration and duration, hydration and calories tend to have a positive relationship. But mainly the variables are seeming to be not correlated with each other.

Histogram



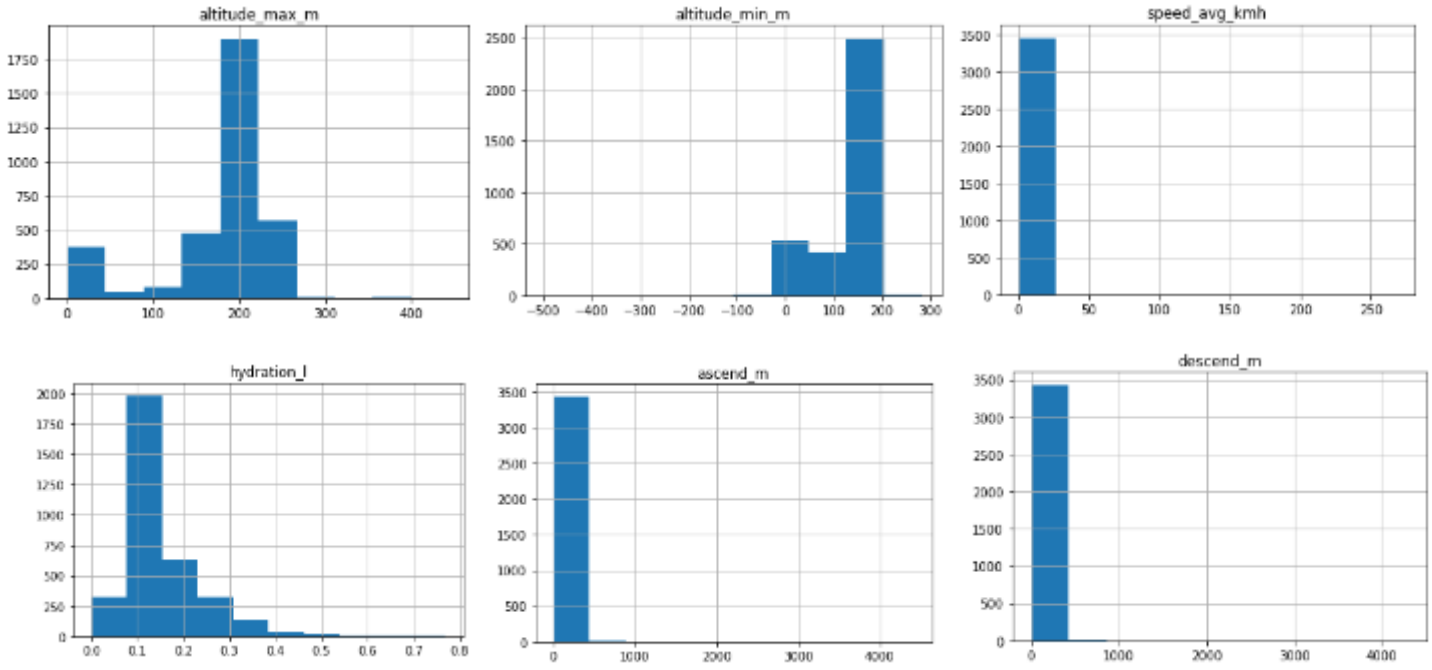


Figure 10 Distribution of all variables

We can see how the variables have been occurred. We have considered the variables and their frequency distribution. Accordingly, it can be observed that most of the variables are positively skewed. Such as - duration_s, distance_km, calories_kcal, speed_avg_kmh, speed_max_kmh, ascend_m, descend_m and hydration_l. Except for the variable – Altitude_min_m which is negatively skewed.

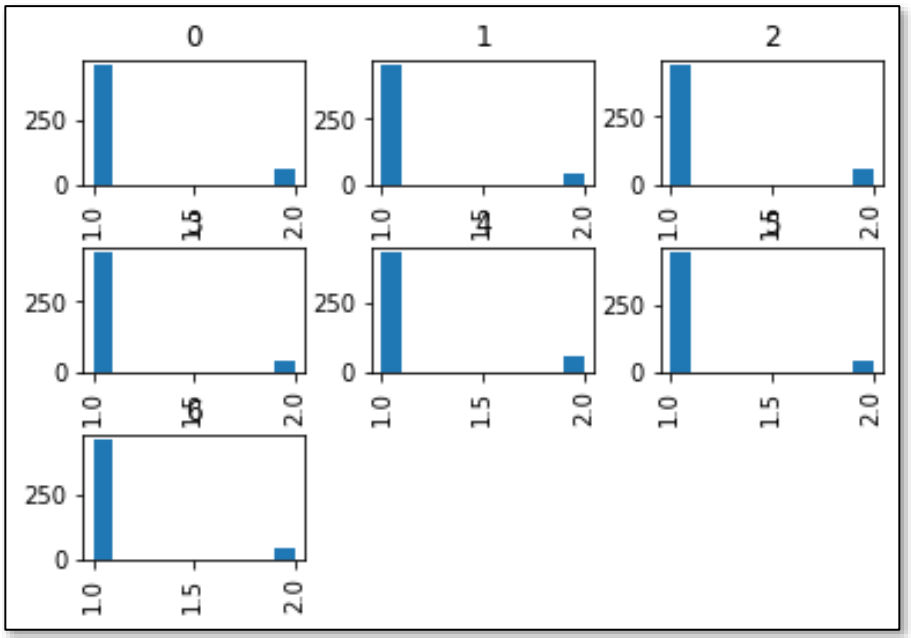


Figure 11 Days of the week, Distribution

The data has been gathered in 2 different ways - TRACK_MOBILE & INPUT_MANUAL_MOBILE. And the above plots show how the individual has input the data daily. The days of the week has been denoted from 0 – 6 (Sunday to Saturday). And it can be said that the data has been input by means of Mobile. (By tracking the mobile)

Feature Extraction and Plan for the modelling goal

In order to achieve the modelling goal of predicting the activity of the person, from the original data set, the start time and end time were removed and instead the start time in hours and the day of the week where the activity was performed was generated and taken into consideration.

The idea behind choosing those 2 features was, the assumption of having a habit of a person would make better prediction of the next sport activity.

As for modelling goal, performing a regression and decision tree and KNN was taken into consideration and, in order to perform the regression, the correlation of explanatory variables was checked.

Correlation

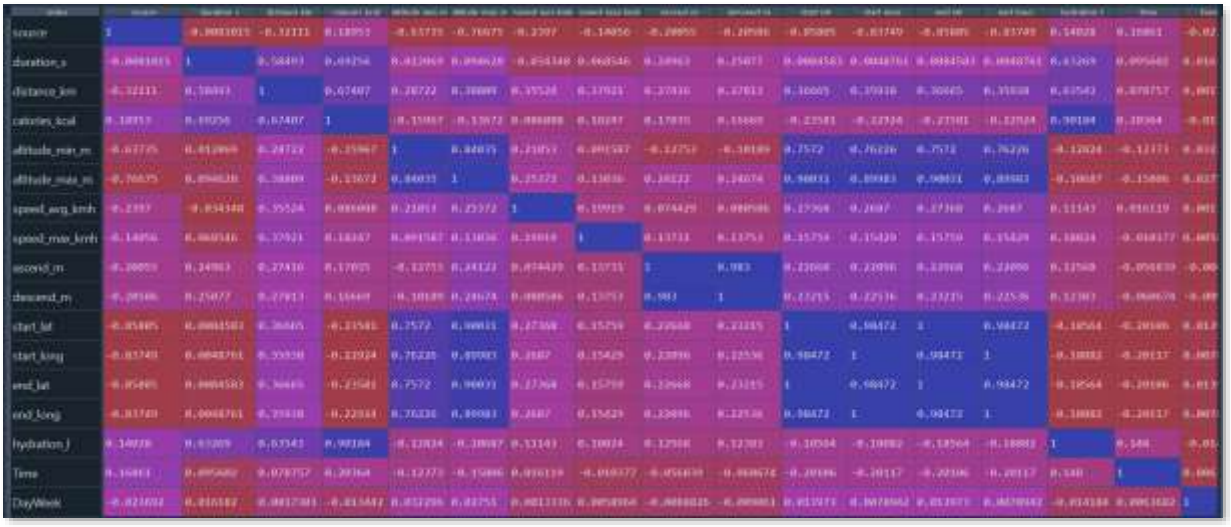


Figure 12 Correlation Matrix with x variables

As per the initial correlation plot, there are 6 variables that are seems to be highly correlated. Hence, out of the 17 variables, 5 variables were removed, and correlation was calculated again and presented in correlation matrix as follows.

	source	duration_s	distance_km	calories_kcal	altitude_min_m	speed_avg_kmh	speed_max_kmh	ascend_m	start_lat	Time	DayWeek
source	1	-0.0081015	-0.32111	0.18953	-0.63735	-0.2397	-0.14056	-0.20055	-0.85805	0.16861	-0.023692
duration_s	-0.0081015	1	0.58493	0.69256	0.012069	-0.034348	0.068546	0.24963	0.0084583	0.095602	0.016182
distance_km	-0.32111	0.58493	1	0.67407	0.28722	0.35524	0.37921	0.27416	0.36605	0.078757	0.0017303
calories_kcal	0.18953	0.69256	0.67407	1	-0.15967	0.086088	0.18247	0.17035	-0.23581	0.20364	-0.013442
altitude_min_m	-0.63735	0.012069	0.28722	-0.15967	1	0.21853	0.091587	-0.12753	0.7572	-0.12373	0.032296
speed_avg_kmh	-0.2397	-0.034348	0.35524	0.086088	0.21853	1	0.19919	0.074429	0.27368	0.016119	0.0013336
speed_max_kmh	-0.14056	0.068546	0.37921	0.18247	0.091587	0.19919	1	0.13731	0.15759	-0.010377	0.0058964
ascend_m	-0.20055	0.24963	0.27416	0.17035	-0.12753	0.074429	0.13731	1	0.22668	-0.056839	-0.0088826
start_lat	-0.85805	0.0084583	0.36605	-0.23581	0.7572	0.27368	0.15759	0.22668	1	-0.20106	0.013973
Time	0.16861	0.095602	0.078757	0.20364	-0.12373	0.016119	-0.010377	-0.056839	-0.20106	1	0.0063682
DayWeek	-0.023692	0.016182	0.0017303	-0.013442	0.032296	0.0013336	0.0058964	-0.0088826	0.013973	0.0063682	1

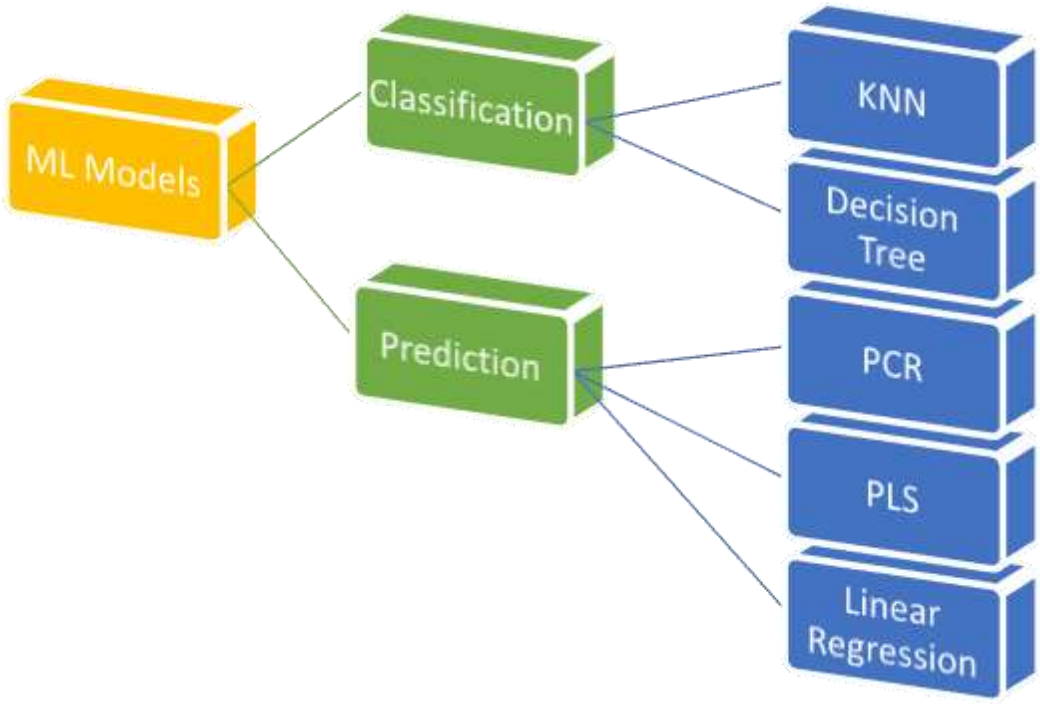
Figure 13 Correlation matrix after taking out high correlated variables

As per the above correlation matrix, the highest correlation is 0.7 and, all the variables that the correlation coefficient is more than 0.8 were removed.

Further, the string variable source was converted to a nominal variable for the convenience of the analysis.

ML Models for classification and prediction

For the classification and prediction of the sport activity, we used 2 different models for each goal.



As per the results of each model for the classification and prediction, the best model for the prediction and classification was chosen.

Classification Models

KNN

As a classification model, KNN was performed and up to 20 neighbors the error rate was calculated, and it is found that the least error rate is at when k equals to 1 and the highest error rate is at when the k is equal to 18. As for the explanatory variables, the final data set that was explained under feature extraction and plan for the modelling goal was used and the normalized data of that were used for the KNN.

Apart from k equals to 1, when k is 5, 6 and 9 the error rate is still very low. If we consider the accuracy difference between when k is 1 and 18, still the accuracy is really good as both accuracies are greater than 90%.

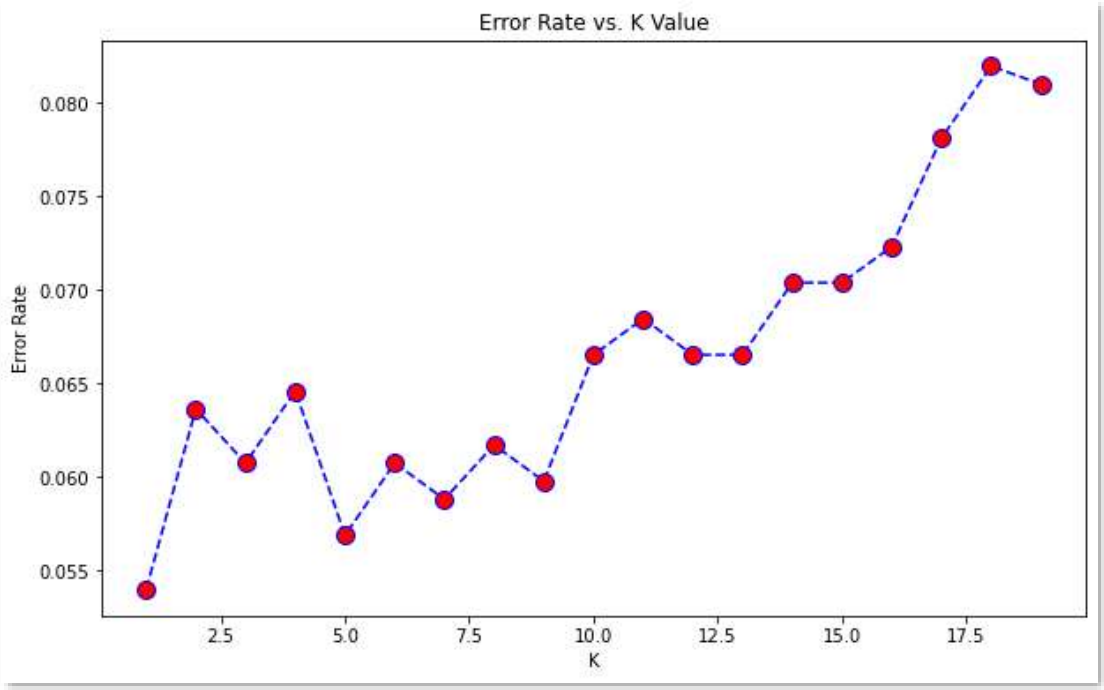


Figure 14 Knn Error Rate per K choosen

Decision Tree

The root node starts with 3456 samples and the Gini index is 0.371. the feature that best split the data is the speed_max_kmh with a threshold value 4.05, resulting in 2 nodes - distance_km and speed_avg_kmh. But the Gini index is low for speed_avg_kmh.

As you can see in the below figure, it can be predicted that we should consider the variables that are on the left-hand side of the tree. Because of the high gini value compared to the right-hand side variables.

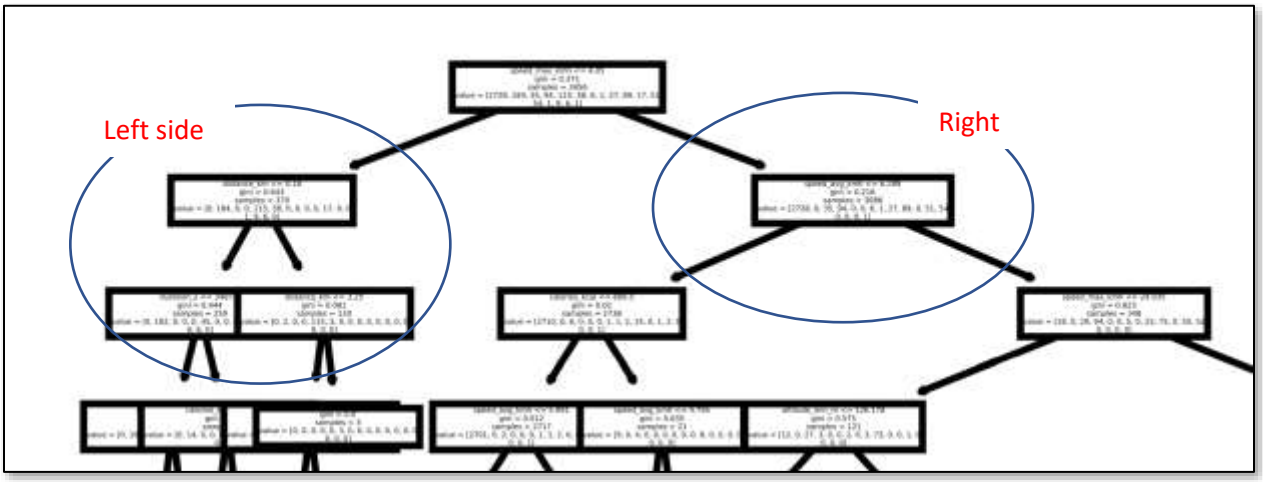


Figure 15 Figure of the top of the tree

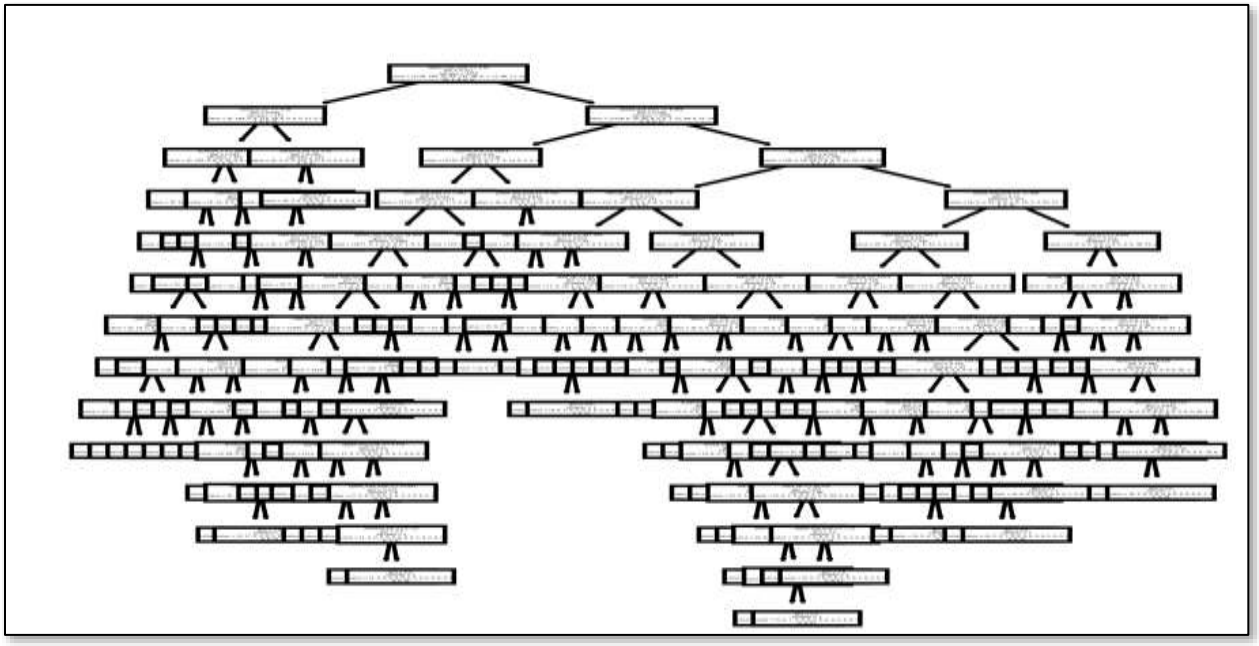


Figure 16 Figure of the entire Decision tree

Prediction Models

Linear Regression

A linear regression is performed with the following variables -'source','duration_s','calories_kcal', 'altitude_min_m','speed_avg_kmh','speed_max_kmh','ascend_m','start_lat','Time','DayWeek','distance_km' (these are considered as the features)
And the predictor variable is sport.
The data has been split into training and testing with a ratio of 80:20.
It has been calculated that the,

- R^2 – 0.418
- Intercept – 4.33
- Coefficients for each variable:

Table 2 Coefficients of LR

Source –	- 0.0011
Duration_s –	0.0037
Calories_kcal –	- 0.0006
Altitude_min –	0.0395
Speed_avg_kmh –	0.0082
Ascend_m –	- 0.0010
Start_lat –	- 0.0595
Time –	0.0057
DayWeek --	0.0003
Distance_km –	0.6212

According to the coefficients the distance is the variable with more importance over all the variables chosen.

PCR

The variables chosen for this PCA where the same as the ML. Regarding the number of compone nts, 1 was choose because it explains most of the variables according to de graphic.

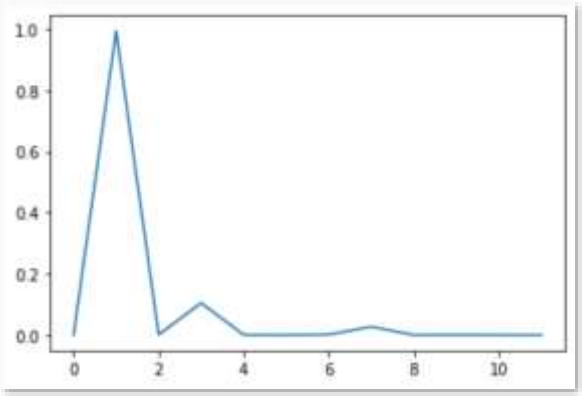


Figure 17 Variance explained by # of PCA

For the graphic above we can see that the PCR does not predict very well the sport variable, it does not look like a tendency overall but a certain points over the graphic.

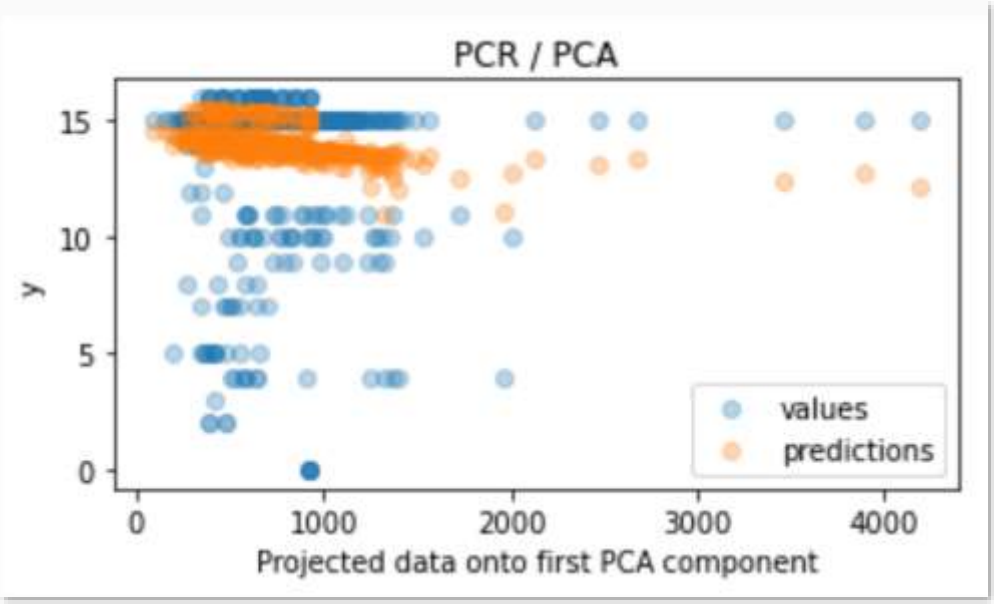


Figure 18 Projected values using PCR

PLS

As the PCA we choose 1 component for the PLS and the same variables, taking out the variables with high correlation mention before. We got a better result in the prediction of the variable, not only it can be seen in the graphic, but the R-square give us above 70%.

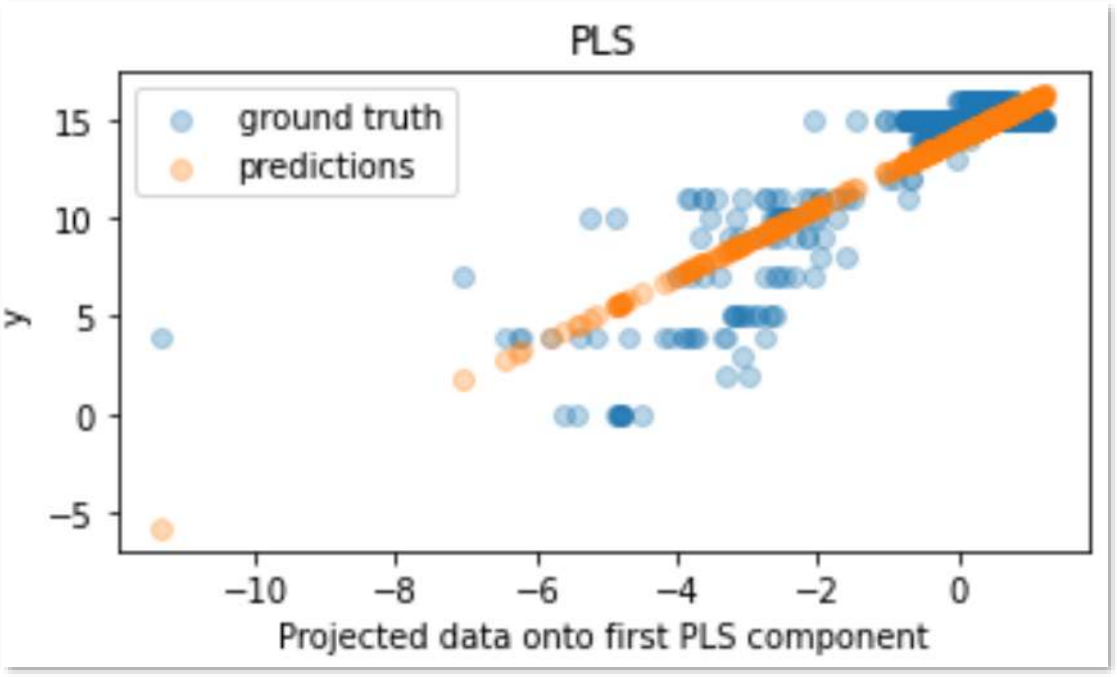


Figure 19 Projected values using PLS

Conclusion

The final accuracy of the 5 models is as follows.

	Classification	
	Decision Tree	**KNN**
Accuracy	95%	95%

	Prediction		
	PCR	**Linear Regression**	**PLS**
Accuracy	42.5%	41%	72%

As per the results it can conclude that out of the 3 prediction models the PLS model is the best option, not only because of the high accuracy compared to the other predictive models but also because it takes less resource since is taking into account once PCA instead of all variables as the linear regression is.

When it comes to the classification models both decision tree and KNN are similarly good as both models has given the same accuracy of 95%.