INTER-RELATIONSHIP BETWEEN SENSOR DATA SERIES USING GAA DATA

Programme:	MSc. In Computing
Course code:	MCM
Module:	Data Analytics and Data Mining
Module code:	CA683
Date of submission:	17/04/2018

Haree Varshan Jeyaram - 17211965

Mithra Devi Veeramohan - 17210404

Sai Sree Malepati Ravikumar - 17210622

Viswanathan Umamaheswaran - 17211729

Acknowledgement

A report submitted to Dublin City University, School of Computing for module CA683 Cloud Architecture, 2017/2018. We understand that the University regards breaches of academic integrity and plagiarism as grave and serious. We have read and understood the DCU Academic Integrity and Plagiarism Policy. We accept the penalties that may be imposed should we engage in practice or practices that breach this policy. We have identified and included the source of all facts, ideas, opinions, viewpoints of others in the assignment references. Direct quotations, paraphrasing, discussion of ideas from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the sources cited are identified in the report references. We declare that this material, which we now submit for assessment, is entirely our own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of our work. By signing this form or by submitting this material online we confirm that this assignment, or any part of it, has not been previously submitted by us or any other person for assessment on this or any other course of study. By signing this form or by submitting material for assessment online we confirm that we have read and understood DCU Academic Integrity and Plagiarism Policy (available at http://www.dcu.ie/registry/examinations/index.shtml).

Name:	Haree Varshan Jeyaram -	17211965
	Mithra Devi Veeramohan -	17210404
	Sai Sree Malepati Ravikumar -	17210622
	Viswanathan Umamaheswaran -	17211729

Date: 17/04/2018

1. Abstract

A project has been proposed to analyze the performance of football players who play under GAA(Gaelic Athletic Association). It is the fact that training is the key to success in any sport. And a game like football needs serious, dedicated and monitored training sessions. A Data Mining technique called CRISP-DM has been used to Understand, Prepare, Model and Evaluate the data collected. RPE(Rate of Perceived Exertion) is the key to predict the performance of the players. "P-value" helped in determining the most important variables that are responsible to predict RPE. Finally, amongst all the models, the decision tree Regression has given the better results.

2. Problem Statement

Predicting what variables determine the Rate of Perceived Exertion (RPE) based on the key attributes like sessions, duration minutes, load, RTT, distance, speed, acceleration, deceleration, body heart rate, sprints, pitches, muscle load etc.

3. Introduction

Practice is the key to success in any game. At the same time, too much of practicing leads to serious injuries as well. The strengths and weaknesses differ from player to player. High-intensity training might bring serious injuries whereas low-intensity intensity training might influence the performance of the player in the game. Finally, in order to optimize the performance of the soccer team, the training process should be modified before assessing its outcome. RPE is the Rate of Perceived Exertion which ranges on a scale from 0 to 10 that measures the intensity of any physical activity. It helps in monitoring the player's stress and physical strength.

4. Data Mining techniques

CRISP-DM is a standard data mining technique used to project planning and management. It stands for **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining. It defines the commonly used approached to handle data mining projects. It is a robust and well-proven data mining technique [3].

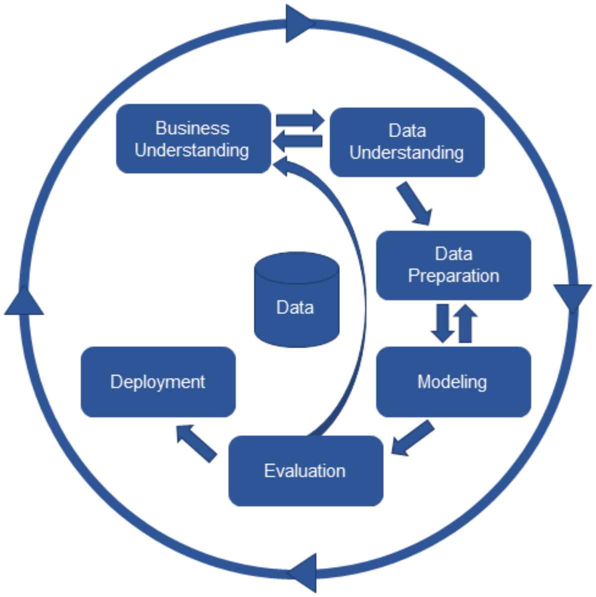


Figure 1 CRISP-DM Methodology [4]

4.1 Business Understanding

RPE is a valuable metric for athletes. It helps in self-regulating their efforts. It is used to measure the physical activity intensity level. It scales from 1-10 with 1 being no effort and 10 being the maximum effort. Distance, speed, acceleration, body heart rate, sprints are some of the variables which affect RPE.

4.2 Data Understanding

Data Collection:

- Data has been taken from GAA (Gaelic Athletic Association)
- GAA is an Irish International amateur sporting and cultural organization which promotes Gaelic games. The association has got major influence in Irish sporting and cultural life. Football and hurling are the two main games. Here we have taken football data to predict the performance of the players.

Data Description:

- *RTT & RPE Project database (Training Database.csv)* – Contains the username, each players corresponding session date and their RPE, Load, Duration in mins, RTT, RTT-1, RTT+1.
- *Derivedkpi.csv* – Contains information about the speed, accelerations, decelerations, distance, body load and other KPI measures (116 measures) for the combination of each player and session.
- *Players.csv* – Contains details of the player such as name, dob, height, weight, heart rate, position etc.
- *Sessions.csv* – Contains complete information about the sessions (Activity date, start date, start time, end date, end time)
- *Database_name.csv* – Contains details about the player (first name, last name, display name etc.,)

4.3 Data Preparation

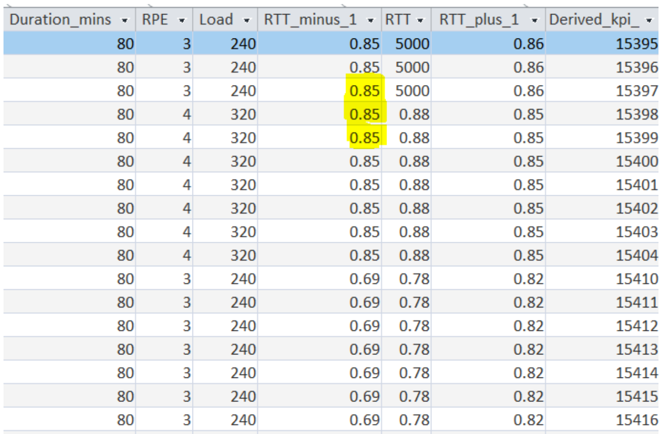
Data Integration and Construction

After gaining a better understanding of the data, all the tables of the dataset have been merged meaningfully to construct a final dataset. INNER JOIN was used to join all the datasets.

- **Step1**: A dataset named Training_db_with_PlayerId was constructed by joining Training_database.csv with database_names.csv with "username" as key. This join is performed to get the Player_id in Training_database.
- **Step2:** A dataset named Training_db_with_PlayerId_Sessions was constructed by joining the dataset created in Step 1 with Sessions_csv with "data" as key. This join is performed to get the session_id in Training_database.
- **Step3:** A dataset named Final_Training_db was constructed by joining the dataset created in Step 2 with derived_kpi.csv with "player_id" and "session_id" as keys. This is the dataset consisting of all the required columns for training the dataset with the models.

Data Formatting

- After constructing the final dataset, the percentage values of the columns RTT, RTT-1, RTT+1 were converted to float values to perform calculations in an easier way.



Duration_mins	RPE	Load	RTT_minus_1	RTT	RTT_plus_1	Derived_kpi_
80	3	240	0.85	5000	0.86	15395
80	3	240	0.85	5000	0.86	15396
80	3	240	0.85	5000	0.86	15397
80	4	320	0.85	0.88	0.85	15398
80	4	320	0.85	0.88	0.85	15399
80	4	320	0.85	0.88	0.85	15400
80	4	320	0.85	0.88	0.85	15401
80	4	320	0.85	0.88	0.85	15402
80	4	320	0.85	0.88	0.85	15403
80	4	320	0.85	0.88	0.85	15404
80	3	240	0.69	0.78	0.82	15410
80	3	240	0.69	0.78	0.82	15411
80	3	240	0.69	0.78	0.82	15412
80	3	240	0.69	0.78	0.82	15413
80	3	240	0.69	0.78	0.82	15414
80	3	240	0.69	0.78	0.82	15415
80	3	240	0.69	0.78	0.82	15416

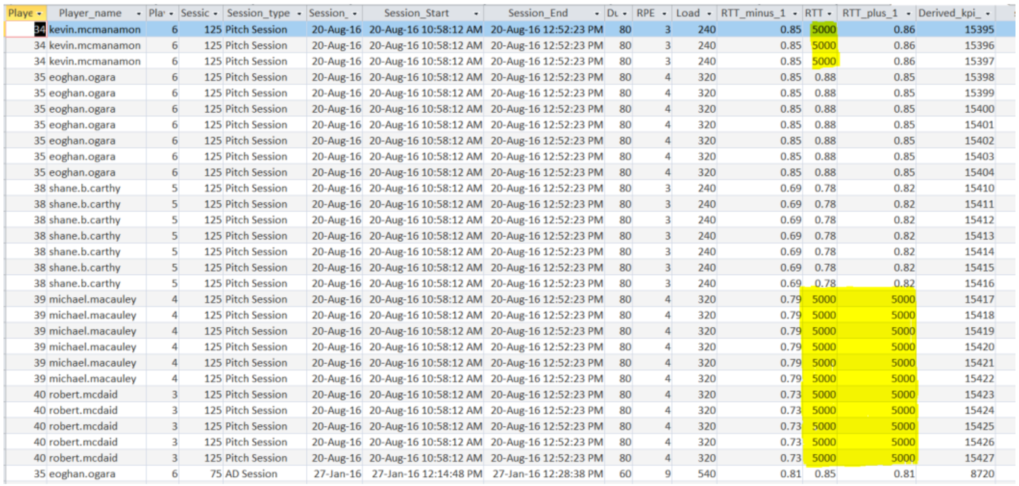
Figure 2: Percentage values converted to Float values

Data Cleaning

In the final dataset, missing values are present in the following columns:

- RPE – 142 values
- RTT – 3736 values
- RTT_Min_1 – 3228 values
- RTT_Plus_1 – 3569 values
- Load – 142 values

The missing values are present as "5000" values. These missing values are randomly distributed in the complete dataset. In the above mentioned five columns there are no null values. Hence, the missing value pattern present in this dataset is **Missing Completely At Random (MCAR)**.



Playe	Player_name	Pla	Sessic	Session_type	Session_	Session_Start	Session_End	Du	RPE	Load	RTT_minus_1	RTT	RTT_plus_1	Derived_kpi_
34	kevin.mcmanamon	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.85	5000	0.86	15395
34	kevin.mcmanamon	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.85	5000	0.86	15396
34	kevin.mcmanamon	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.85	5000	0.86	15397
35	eoghan.ogara	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.85	0.88	0.85	15398
35	eoghan.ogara	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.85	0.88	0.85	15399
35	eoghan.ogara	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.85	0.88	0.85	15400
35	eoghan.ogara	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.85	0.88	0.85	15401
35	eoghan.ogara	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.85	0.88	0.85	15402
35	eoghan.ogara	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.85	0.88	0.85	15403
35	eoghan.ogara	6	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.85	0.88	0.85	15404
38	shane.b.carthy	5	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.69	0.78	0.82	15410
38	shane.b.carthy	5	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.69	0.78	0.82	15411
38	shane.b.carthy	5	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.69	0.78	0.82	15412
38	shane.b.carthy	5	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.69	0.78	0.82	15413
38	shane.b.carthy	5	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.69	0.78	0.82	15414
38	shane.b.carthy	5	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.69	0.78	0.82	15415
38	shane.b.carthy	5	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	3	240	0.69	0.78	0.82	15416
39	michael.macauley	4	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.79	5000	5000	15417
39	michael.macauley	4	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.79	5000	5000	15418
39	michael.macauley	4	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.79	5000	5000	15419
39	michael.macauley	4	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.79	5000	5000	15420
39	michael.macauley	4	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.79	5000	5000	15421
39	michael.macauley	4	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.79	5000	5000	15422
40	robert.mcdaid	3	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.73	5000	5000	15423
40	robert.mcdaid	3	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.73	5000	5000	15424
40	robert.mcdaid	3	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.73	5000	5000	15425
40	robert.mcdaid	3	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.73	5000	5000	15426
40	robert.mcdaid	3	125	Pitch Session	20-Aug-16	20-Aug-16 10:58:12 AM	20-Aug-16 12:52:23 PM	80	4	320	0.73	5000	5000	15427
35	eoghan.ogara	6	75	AD Session	27-Jan-16	27-Jan-16 12:14:48 PM	27-Jan-16 12:28:38 PM	60	9	540	0.81	0.85	0.81	8720

Figure 3: Missing values present as "5000"

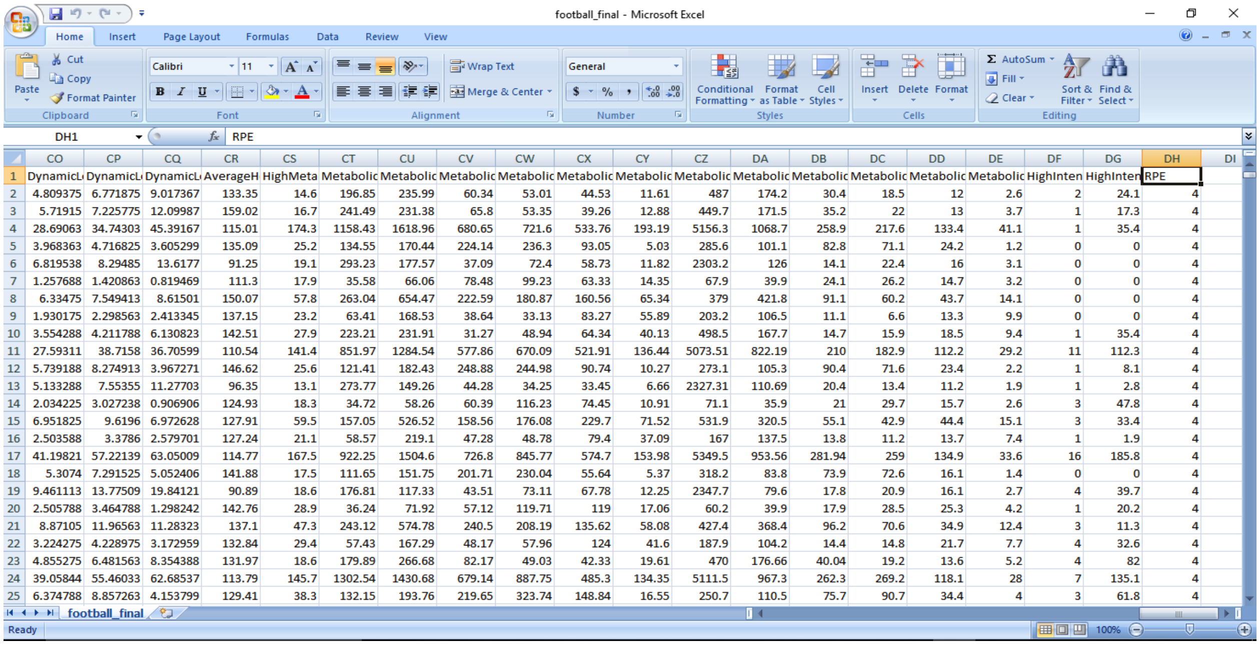
Based on the given dataset, two possibilities were derived to impute the missing values :

- Imputing the "5000" values with the mean value acquired by grouping the respective variable according to each player.
- Imputing the "5000" values with the mean value acquired by grouping the respective variable according to each session and comparing the corresponding variable values with other players having all valid values.

In few cases, all the columns had missing values for a particular player, therefore approach 2 could not compare the corresponding values with other players and impute the missing ones. Hence approach 1 was chosen for the imputation of missing values.

There were few variables whose complete columns had NULL values for each player and session. Hence these columns were removed from the final dataset. The dataset includes a categorical variable session_type. This column was subjected to label encoding and oneHot encoding to convert the categorical variable into a nominal variable. Label encoding converts each parameter of the variable into a number and OneHot encoding convert the numbers into an array of binary numbers (0,1).

The final dataset after cleaning consists of one dependent variable – RPE and 120 independent variables which are shown in below figure.



DH1 fx RPE

	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH
1	DynamicL	DynamicL	DynamicL	AverageH	HighMeta	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	Metabolic	HighInten	HighInten	RPE
2	4.809375	6.771875	9.017367	133.35	14.6	196.85	235.99	60.34	53.01	44.53	11.61	487	174.2	30.4	18.5	12	2.6	2	24.1	4
3	5.71915	7.225775	12.09987	159.02	16.7	241.49	231.38	65.8	53.35	39.26	12.88	449.7	171.5	35.2	22	13	3.7	1	17.3	4
4	28.69063	34.74303	45.39167	115.01	174.3	1158.43	1618.96	680.65	721.6	533.76	193.19	5156.3	1068.7	258.9	217.6	133.4	41.1	1	35.4	4
5	3.968363	4.716825	3.605299	135.09	25.2	134.55	170.44	224.14	236.3	93.05	5.03	285.6	101.1	82.8	71.1	24.2	1.2	0	0	4
6	6.819538	8.29485	13.6177	91.25	19.1	293.23	177.57	37.09	72.4	58.73	11.82	2303.2	126	14.1	22.4	16	3.1	0	0	4
7	1.257688	1.420863	0.819469	111.3	17.9	35.58	66.06	78.48	99.23	63.33	14.35	67.9	39.9	24.1	26.2	14.7	3.2	0	0	4
8	6.33475	7.549413	8.61501	150.07	57.8	263.04	654.47	222.59	180.87	160.56	65.34	379	421.8	91.1	60.2	43.7	14.1	0	0	4
9	1.930175	2.298563	2.413345	137.15	23.2	63.41	168.53	38.64	33.13	83.27	55.89	203.2	106.5	11.1	6.6	13.3	9.9	0	0	4
10	3.554288	4.211788	6.130823	142.51	27.9	223.21	231.91	31.27	48.94	64.34	40.13	498.5	167.7	14.7	15.9	18.5	9.4	1	35.4	4
11	27.59311	38.7158	36.70599	110.54	141.4	851.97	1284.54	577.86	670.09	521.91	136.44	5073.51	822.19	210	182.9	112.2	29.2	11	112.3	4
12	5.739188	8.274913	3.967271	146.62	25.6	121.41	182.43	248.88	244.98	90.74	10.27	273.1	105.3	90.4	71.6	23.4	2.2	1	8.1	4
13	5.133288	7.55355	11.27703	96.35	13.1	273.77	149.26	44.28	34.25	33.45	6.66	2327.31	110.69	20.4	13.4	11.2	1.9	1	2.8	4
14	2.034225	3.027238	0.906906	124.93	18.3	34.72	58.26	60.39	116.23	74.45	10.91	71.1	35.9	21	29.7	15.7	2.6	3	47.8	4
15	6.951825	9.6196	6.972628	127.91	59.5	157.05	526.52	158.56	176.08	229.7	71.52	531.9	320.5	55.1	42.9	44.4	15.1	3	33.4	4
16	2.503588	3.3786	2.579701	127.24	21.1	58.57	219.1	47.28	48.78	79.4	37.09	167	137.5	13.8	11.2	13.7	7.4	1	1.9	4
17	41.19821	57.22139	63.05009	114.77	167.5	922.25	1504.6	726.8	845.77	574.7	153.98	5349.5	953.56	281.94	259	134.9	33.6	16	185.8	4
18	5.3074	7.291525	5.052406	141.88	17.5	111.65	151.75	201.71	230.04	55.64	5.37	318.2	83.8	73.9	72.6	16.1	1.4	0	0	4
19	9.461113	13.77509	19.84121	90.89	18.6	176.81	117.33	43.51	73.11	67.78	12.25	2347.7	79.6	17.8	20.9	16.1	2.7	4	39.7	4
20	2.505788	3.464788	1.298242	142.76	28.9	36.24	71.92	57.12	119.71	119	17.06	60.2	39.9	17.9	28.5	25.3	4.2	1	20.2	4
21	8.87105	11.96563	11.28323	137.1	47.3	243.12	574.78	240.5	208.19	135.62	58.08	427.4	368.4	96.2	70.6	34.9	12.4	3	11.3	4
22	3.224275	4.228975	3.172959	132.84	29.4	57.43	167.29	48.17	57.96	124	41.6	187.9	104.2	14.4	14.8	21.7	7.7	4	32.6	4
23	4.855275	6.481563	8.354388	131.97	18.6	179.89	266.68	82.17	49.03	42.33	19.61	470	176.66	40.04	19.2	13.6	5.2	4	82	4
24	39.05844	55.46033	62.68537	113.79	145.7	1302.54	1430.68	679.14	887.75	485.3	134.35	5111.5	967.3	262.3	269.2	118.1	28	7	135.1	4
25	6.374788	8.857263	4.153799	129.41	38.3	132.15	193.76	219.65	323.74	148.84	16.55	250.7	110.5	75.7	90.7	34.4	4	3	61.8	4

Figure 4: Final dataset with one dependent variable (RPE) and 120 independent variables

5. Modeling

There are totally 121 variables in the dataset. As we see from figure1, RPE is the dependent variable and it is continuous, hence it has been identified as a regression problem. Therefore, linear regression was implemented between the dependent variable (RPE) and the independent variables (remaining 120 variables). And a test design was generated using backward elimination technique.

5.1 Linear Regression

Since the dependent variable is continuous, it has been identified as a Regression problem. So, linear regression was implemented between RPE and all independent variables.

- **Goodness of fit -** It depicts a model which fits the data well if the difference between the observed values and predicted values are small.
- **p-value -** The p-value is the level of significance in a statistical hypothesis test which represents the likelihood of the occurrence of the given event. It is an alternative to the rejection points which provides the smallest level of significance at which the null hypothesis would be rejected.

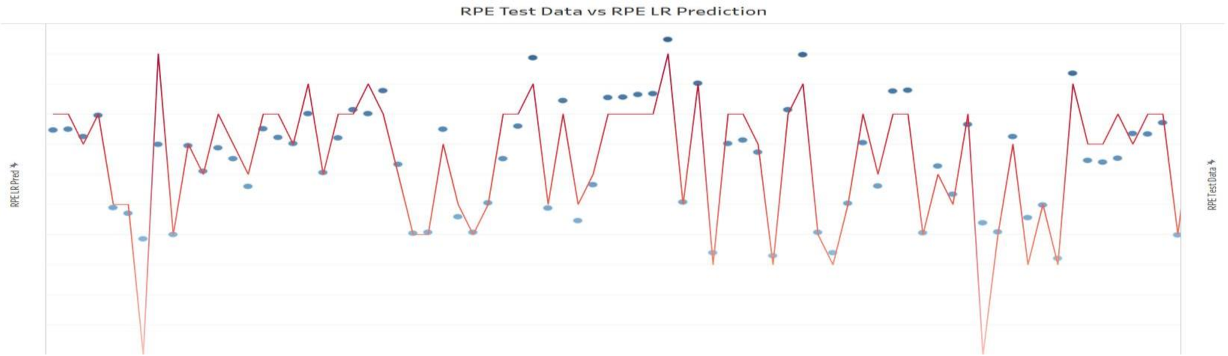


Figure 5: — *Actual Test Data of RPE values* ● *Predicted RPE values*

- **Feature Selection – Technique chosen : Backward Elimination using p-value**

¨Sometimes, less is better¨ Selecting which independent variables contribute to the dependent variable is very important rather than feeding the whole dataset into the model. The performance is improved by selecting the maximum tolerable error rate, the smallest number of features necessary to reach that classification performance in the chosen algorithm.

5.2 Decision Tree Regression

It has been believed that the Backward elimination of linear regression has given the exact variables (62 variables) which affect RPE. So the Decision Tree Regression has been implemented using the 62 variables and it has produced better accuracy than the other models.

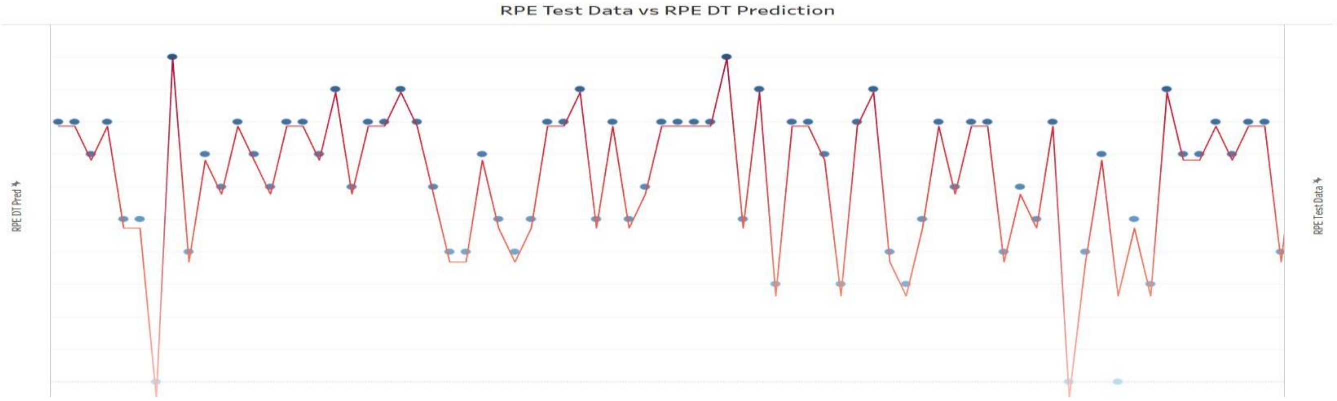


Figure 6: — *Actual Test Data of RPE values* ● *Predicted RPE values*

5.3 Artificial Neural Network

In order to enhance the comparison, a deep learning model called Artificial Neural Network (ANN) has been implemented using the 62 variables. Considering the system's configuration a small ANN was created and the results were yielded.

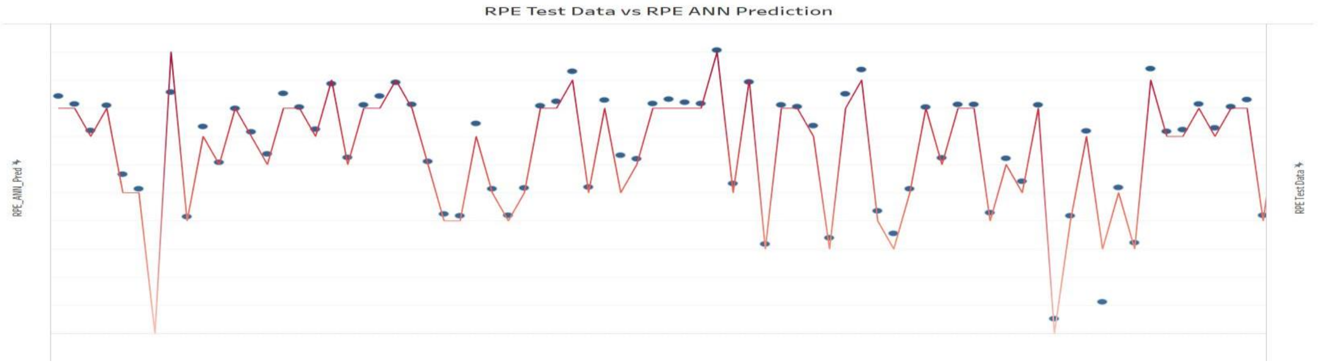


Figure 7: —— *Actual Test Data of RPE values* ● *Predicted RPE values*

6. Evaluation

To evaluate the implemented models, metrics such as Mean Squared Error, Explained Variance Score, Median Absolute Error were used. Below table depicts the metrics score for each model and shows that Decision Tree Regression has given the better results when comparing with other two models.

	Linear Regression	Artificial Neural Network	Decision Tree Regression
Mean Squared Error	0.67	0.34	**0.04**
Explained Variance Score	0.84	0.92	**0.99**
Median Absolute Error	0.32	0.14	**0.0**

7. Feedback

From the presentation, we were asked to improve on the following

1. Removal of RTT+1 variable.
2. Choice of the Model as the dependent variable (RPE) was ordinal.
3. Reason for choosing Backward Elimination.

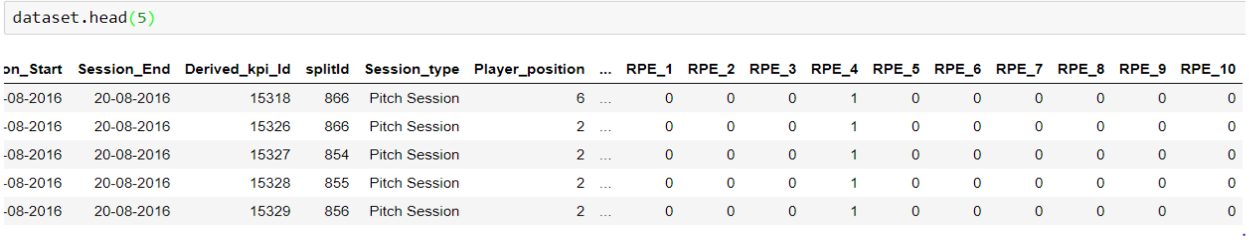
1. Removal of RTT+1 variable

The “RTT+1” variable was removed from the analysis as they represent a future score and hence they may not be useful for the prediction of RPE.

2. Regression vs. Classification

As the dependent variable RPE values were over the scale 0-10, we assumed it to be continuous initially. Linear Regression, Decision Tree Regression, and Artificial Neural Network were implemented on the data considering RPE as continuous and we were able to get good prediction results on the test data as seen above in the validation table.

Based on the feedback received, RPE was considered as an ordinal variable and hence we did the modeling as a classification problem. The RPE value, which had values over the scale 1-10, was one hot encoded and converted into 11 dummy variables, with each variable representing its corresponding RPE(RPE_0, RPE_1…RPE_10).



```
dataset.head(5)
```

on_Start	Session_End	Derived_kpi_Id	splitId	Session_type	Player_position	...	RPE_1	RPE_2	RPE_3	RPE_4	RPE_5	RPE_6	RPE_7	RPE_8	RPE_9	RPE_10
-08-2016	20-08-2016	15318	866	Pitch Session	6	...	0	0	0	1	0	0	0	0	0	0
-08-2016	20-08-2016	15326	866	Pitch Session	2	...	0	0	0	1	0	0	0	0	0	0
-08-2016	20-08-2016	15327	854	Pitch Session	2	...	0	0	0	1	0	0	0	0	0	0
-08-2016	20-08-2016	15328	855	Pitch Session	2	...	0	0	0	1	0	0	0	0	0	0
-08-2016	20-08-2016	15329	856	Pitch Session	2	...	0	0	0	1	0	0	0	0	0	0

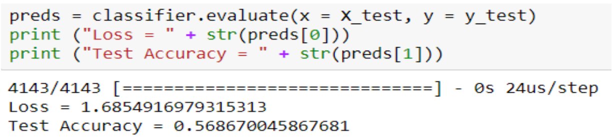
Hence, we implemented an Artificial Neural Network on

1. Complete Football dataset
   - Input Layer (30 neurons, input dimensions = 109)
   - 2 hidden layers
   - Output Layer (10 neurons (predicted RPE))

2. Optimized Football dataset (backward elimination using p-value)
   - Input Layer (30 neurons, input dimensions = 109)
   - 2 hidden layers
   - Output Layer (10 neurons (predicted RPE))

The results were evaluated based on loss and accuracy and we found out that the ANN on the optimized data had better results than the complete dataset.

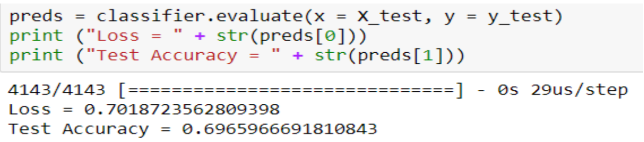
	ANN(full dataset)	ANN (p-value)
Loss	1.69	0.70
Accuracy	0.57	0.70



```
preds = classifier.evaluate(x = X_test, y = y_test)
print ("Loss = " + str(preds[0]))
print ("Test Accuracy = " + str(preds[1]))
```

```
4143/4143 [==============================] - 0s 24us/step
Loss = 1.6854916979315313
Test Accuracy = 0.568670045867681
```

a) Full dataset



```
preds = classifier.evaluate(x = X_test, y = y_test)
print ("Loss = " + str(preds[0]))
print ("Test Accuracy = " + str(preds[1]))
```

```
4143/4143 [==============================] - 0s 29us/step
Loss = 0.7018723562809398
Test Accuracy = 0.6965966691810843
```

b) p-value reduced dataset

3. Reason to choose Backward Elimination

The main reason to select Backward Elimination as the feature selection technique in our project is to reduce the number of noisy variables and only consider those important attributes while training the model. We assumed that this approach would be efficient because it removes the least significant feature at each iteration based on the p-value and is repeated until no improvement is observed on removal of features. This improves the performance of the model, reduces overfitting and enhances the model testing time by interpreting more accurately.

8. Conclusion

To conclude, data understanding, data preparation and data cleaning are the major factors which impact highly in yielding better results. Data mining technique like CRISP-DM gives the clear view of data. Backward elimination method in linear regression helped in predicting the variables which affected RPE. The machine learning and deep learning models that were used for predicting the RPE were evaluated with metrics like accuracy, and we are able to conclude that the data that was optimized using Backward Elimination (p-value) yielded better results.

The data, presentation, report, sql and code files can be found in the below github repository.

https://github.com/haree-vj/data-mining-assignment-gaa.git

9. Future work

Since the data consist of sessions information with start date and end date with timestamps, a time series model could be implemented. A more complex neural network could be implemented with increased neurons and hidden layers and by enhancing the system’s configuration.

10. References

1. Anon, (n.d.). *How To Use The RPE Scale For Strength Training (Plus What The Research Suggests) - BarBend*. [online] Available at: https://barbend.com/how-to-use-rpe-scale-strength-training/.

2. Anon, (n.d.). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?*. [online] Available at: http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.

3. Smart Vision Europe. *What is the CRISP-DM methodology?*.[online] Available at: https://www.sv-europe.com/crisp-dm-methodology/.

4. James Taylor. *Four Problems in Using CRISP-DM and How To Fix Them –Kdnuggets*. [online] Available at: https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html.