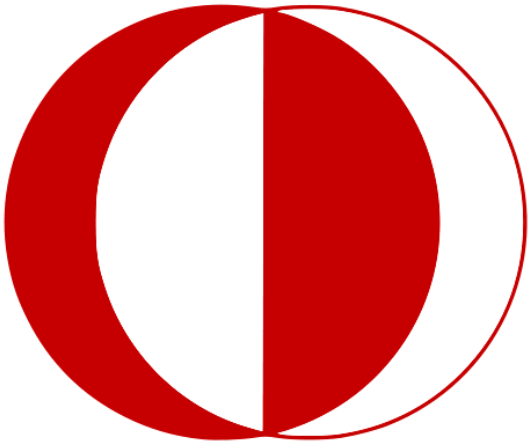

MIDDLE EAST TECHNICAL UNIVERSITY

Department of Statistics

STAT 412 STATISTICAL DATA ANALYSIS TERM PROJECT

FINAL REPORT

"A study into the impacts of appearance and abilities of a fictional character with super powers on the character alignment "

Submitted by: Ebru ŞAKAR

Advisor: Assoc. Prof. CEYLAN TALU YOZGATLIGİL

June 21, 2020

TABLE OF CONTENT

0.ABSTRACT

There are two most essential characters to consider when creating a story. The good one and The evil one. All authors know this. However, are there any physical features or abilities that they oftenly use when creating these characters? Based on this question, this research investigates predicting whether the characters are good or bad by looking at these features, and which method is the most appropriate one. To answer these research questions, two different data set are merged and used.The more details about the merger process and these two data sets are shown in the data description & pre-processing part. Confirmatory data analysis & results part includes the 6 different method results such as XGBoost, SVM, Linear Regression etc. What can be interpreted by looking at analysis results is discussed in discussion and conclusion part of the study.

1.INTRODUCTION

Since the beginning of the world, there is a struggle between good and evil. Therefore, the battle of superheroes and supervillains in movies has always attracted viewers. Then, is there a specific feature that differentiates good people from evil people? The study sought to answer this research question. In other words, the purpose of this research is predicting whether a fictional character with super powers is superhero or supervillain by his\her own appearance and abilities.

2. DATA DESCRIPTION & PRE-PROCESSING

There are two datasets used for this study. The first dataset contains physical characteristics of 734 super villains and superheroes, and there are 10 physical characteristics for each subject. The second one contains 611 samples, each with 9 features, and it records the superpower and ability scores such as speed score, power score, strength score etc. These two dataset which is collected in June/2017 from Superhero Database have been merged according to name of the fictional characters with super powers.

2.1.Explaratory Data Analysis

name	Gender	Eye color	Race	Hair color	Height	Publisher	Skin color
A-Bomb : 1	Female:196	blue :220	Human :200	Black :155	Min. : 15.2	Marvel Comics :376	green : 21
Abe Sapien : 1	Male :491	brown :122	Mutant : 62	Blond : 96	1st Qu.:173.0	DC Comics :208	blue : 9
Abin Sur : 1	NA's : 28	green : 71	God / Eternal : 14	Brown : 81	Median :183.0	NBC - Heroes : 19	red : 9
Abomination : 1		red : 45	Cyborg : 11	No Hair: 75	Mean :186.9	Dark Horse Comics: 18	white : 7
Abraxas : 1		black : 23	Human / Radiation: 11	Red : 49	3rd Qu.:191.0	George Lucas : 14	grey : 5
Absorbing Man: 1		(Other): 68	(Other) :119	(Other): 92	Max. :975.0	(Other) : 65	(Other): 21
(Other) :709		NA's :166	NA's :298	NA's :167	NA's :209	NA's : 15	NA's :643

Alignment	Weight	Intelligence	Strength	Speed	Durability	Power	Combat
bad :206	Min. : 2	Min. : 1.0	Min. : 1.00	Min. : 1.00	Min. : 1.00	Min. : 0.00	Min. : 1.00
good :478	1st Qu.: 61	1st Qu.: 1.0	1st Qu.: 1.00	1st Qu.: 1.00	1st Qu.: 1.00	1st Qu.: 0.00	1st Qu.: 1.00
neutral: 24	Median : 81	Median : 50.0	Median : 11.00	Median : 23.00	Median : 40.00	Median : 38.00	Median : 50.00
NA's : 7	Mean :113	Mean : 45.5	Mean : 29.79	Mean : 27.75	Mean : 42.87	Mean : 41.12	Mean : 44.03
	3rd Qu.:108	3rd Qu.: 75.0	3rd Qu.: 53.00	3rd Qu.: 42.00	3rd Qu.: 80.00	3rd Qu.: 69.00	3rd Qu.: 70.00
	Max. :900	Max. :113.0	Max. :100.00	Max. :100.00	Max. :120.00	Max. :100.00	Max. :101.00
	NA's :231	NA's :134	NA's :134	NA's :134	NA's :134	NA's :134	NA's :134

Table 1. Descriptive Statistics of the Heroes Dataset

According to table 1, on average, a hero or a villain has aproximately 45 points for the power, durability, intelligence and combat scores. On the other hand, a hero or a villain has aproximately 28 points for the strength and speed scores, on average. Surprisingly, 25 percent of these scores is equal to 0 or 1. Additionally, some of the scores is greater than 100 for exceptional heroes or villains.The number of good or male characters is greater than number of bad or female characters. Also, any of these characters is very likely to be blue-eyed, black-haired or a human.

2.1.1.Research Questions

- What is the frequency distribution of character alignment?

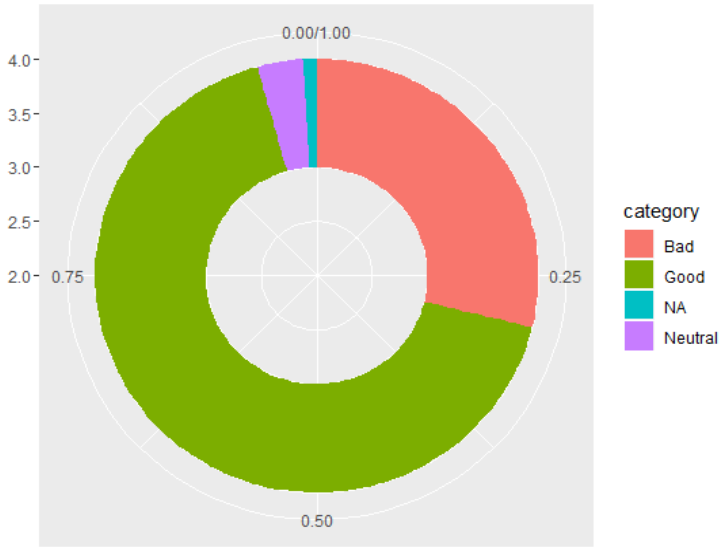

Figure 1: Pie Chart of Character Alignment

The chart above supports that good character is the largest proportion. Bad character is the second largest proportion. Lastly, the proportion of neutral character is larger than the proportion of missing values.

- Who are some of the most intelligent characters in dataset?

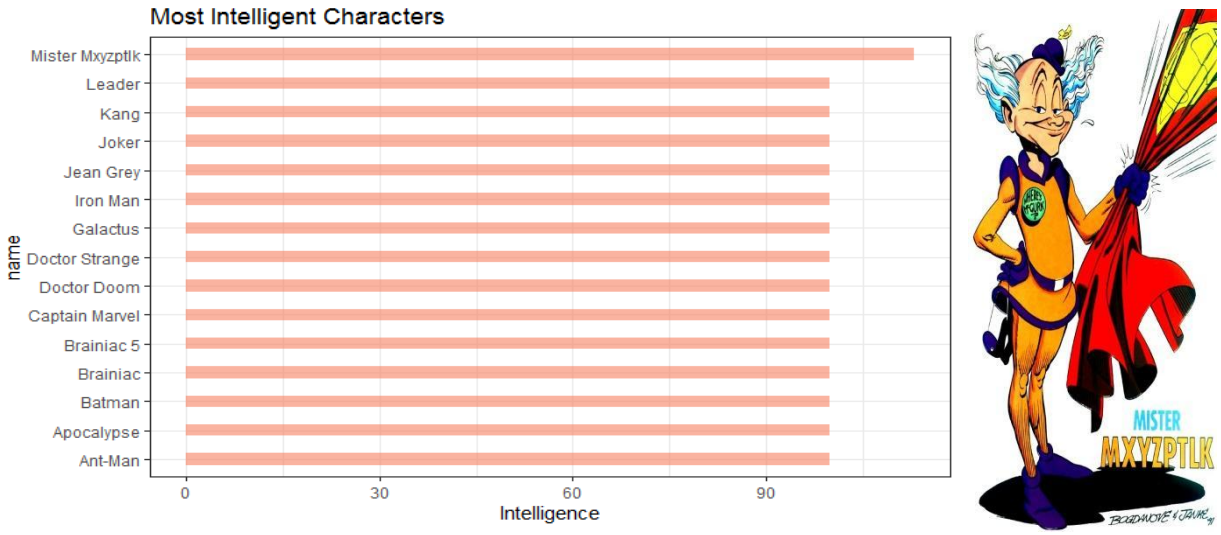

Figure 2: Bar Plot of Most Intelligent Characters

With 113 points, mister mxyzptlk reaches the highest intelligence.

- Is there a relationship between gender and character alignment?

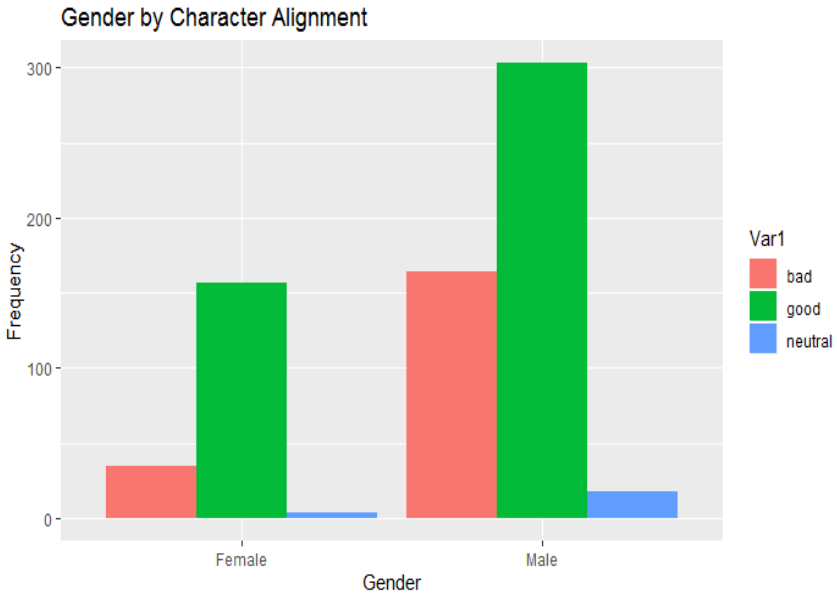

Figure 3: Bar Plot of Gender by Character Alignment

According to Figure 3, compared to the number of bad male or bad female characters, the number of good male or good female characters are smaller.

The Assumptions of the Chi-square Test

Firstly, these two variables are categorical data. Secondly, the variables consist of two or more categorical, independent groups. Finally, no cell is less than one.

Chi-square Test

As the p-value 2.699e-05 ($\chi 2$ (df = 1) = 17.619)

is less than the .05 significance level, we can reject the null hypothesis that gender is independent of the character alignment.

- Is there a relationship between skin color and character alignment?

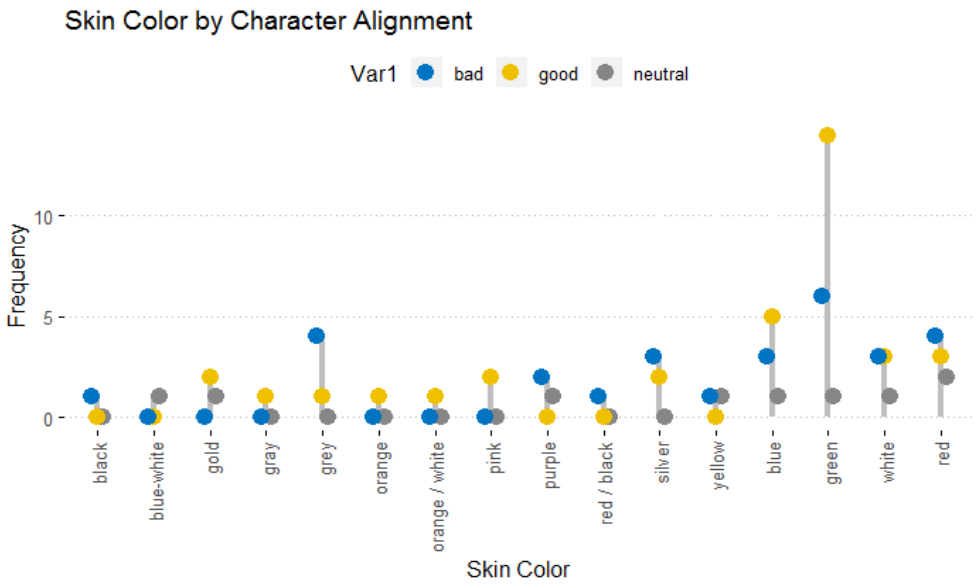

Figure 4: Lollipop Plot of Skin Color by Character Alignment

As can be seen in Figure 4, it indicates that gold, grey, orange, green, pink and blue skin colors are more dominant for good characters compared with bad ones .

- Is there any relationship between continuous variables?

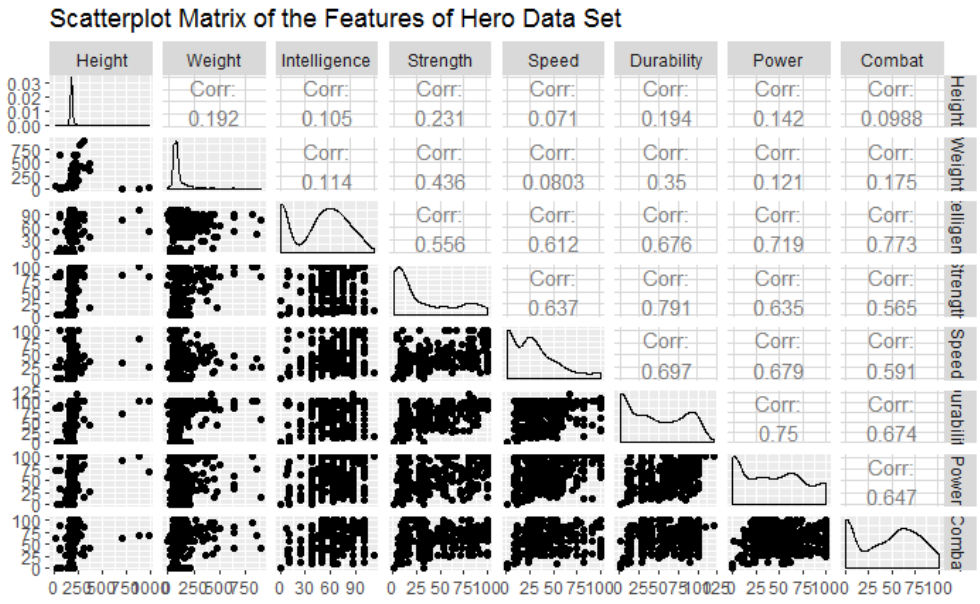

Figure 5. Scatter Plot of Continuous Variables

There is a moderate positive relationship between ability scores such as combat, speed, durability etc.

Name of Variable	VIF value
Weight	1.209465
Height	1.070947
Intelligence	2.773319
Strength	2.679282
Speed	2.318333
Durability	3.504301
Combat	2.639238
Power	3.120003

Fortunately, as it can be seen in the table above, all VIF values are less than 5 which means there is no multicollinearity problem.

2.2. Missing Data Imputation

In deciding whether data is MAR or MCAR, an approach is applied to check for associations between missing and observed data. And this approach is producing pairs plots for missing values and observed values in all variables.

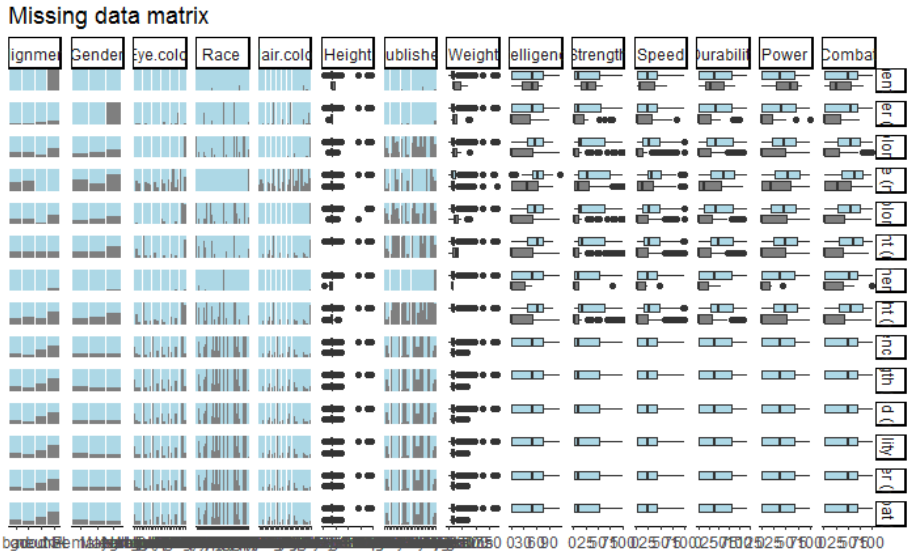

Figure 6. Missing Data Matrix

The data are not missing completely at random (MCAR) because there is a relationship between the missingness of the data. Obviously, missingness in all of the variables above differs by another variable, and the distributions of observed and missing data are not similar to each other. Missing at random (MAR) means the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data. In the light of this information, missing data mechanism for this data is missing at random. In statistically, Little's MCAR Test proves that ($\chi 2$ (df=608) = 953.0772, p<.00001) the missing data is not MCAR.

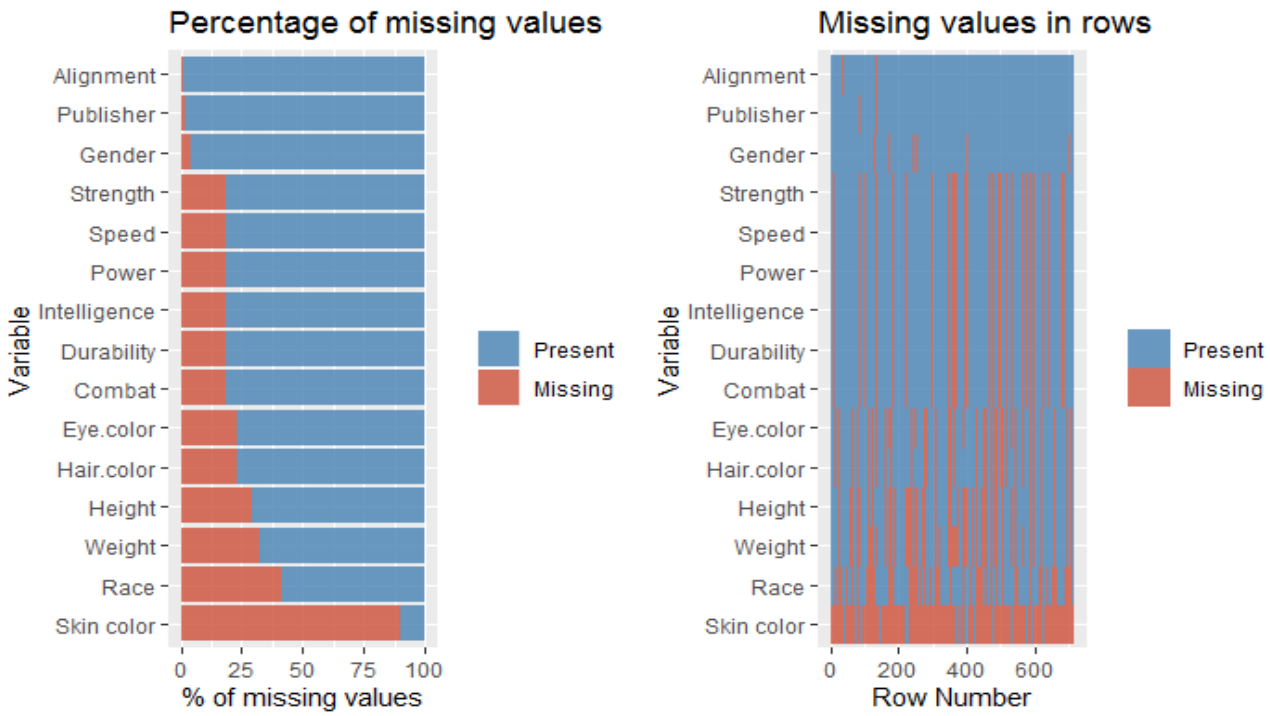

Figure 7. Missing Data of Heroes Dataset

The first plot shows that we have a problem with Skin Color variable since almost 90% of the values are missing. In addition to this, there is the same amount of missing values for ability scores, which means that these values are probably correlated with each other. Links between missing values for different features can be easily seen in the second plot. Related variables seem linked to each other.

- **Deletion**

Skin color variable is deleted due to the high missingness. Likewise, by using list wise deletion method, the whole row of observations is deleted where any of the alignment variable is missing.

- **Mean/ Mode Imputation**

By using mode imputation method which is the most frequently used one , missing values of physical appearance variables are replaced by mode values of these categorical variables. Then,

with mean imputation method, missing values of height and weight variables are replace by the mean of its group with respect to gender.

- **Multiple Imputation By Chained Equations(MİCE)**

Predictive mean matching (PMM) is an plausible way of Multiple Imputation for imputing non-normally distributed quantitative data. Also, this method is used to handle missing data when data are missing at random as it is now. In order to focus numerical values, categorical variables are removed.

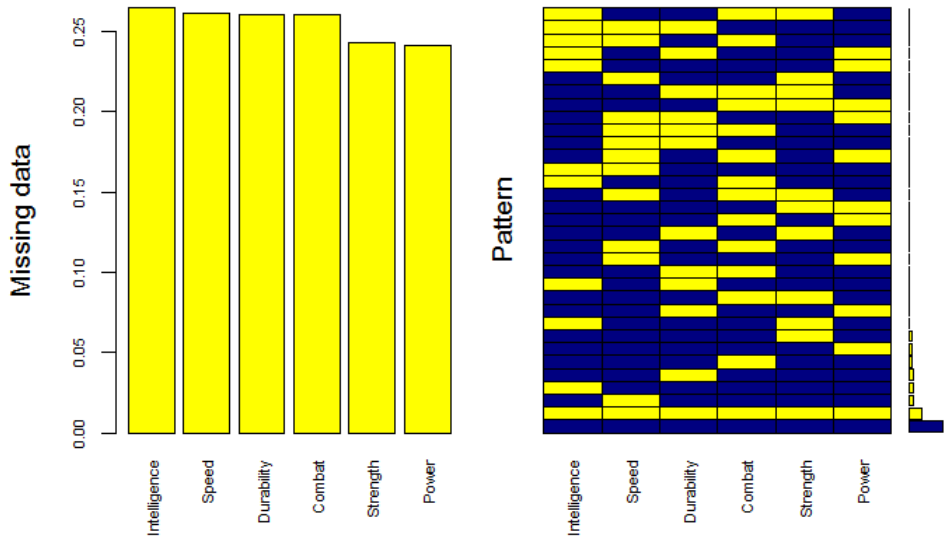

Figure 8. Missing Data of Heroes Dataset

The plot above gives the frequencies for different combination of missing variables. Moreover, blue and yellow represent the observed data and the missing data respectively. According to the bar plot, it seems that more than 25% of values in Intelligence variable is missing. Further, if all of the variables are missing, the probability of this pattern occurring is the highest than other patterns with missing variable.

Checking Imputations Visually

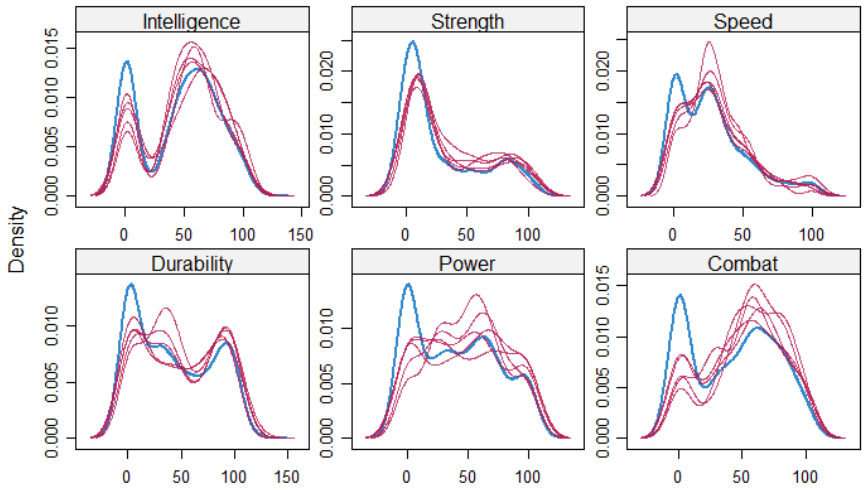

Figure 9. Checking Imputation Visually

This plot compares the density of observed data with the ones of imputed data. As we expected, their densitys are similar to each other.

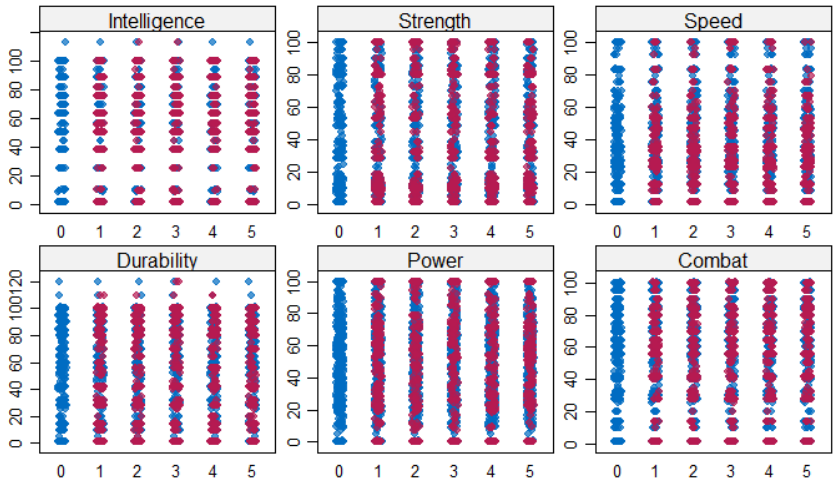

Figure 10. Density of Imputations

Blue and red represents the observed and the imputed data, respectively. As you can see in the scatterplot, these colours are consistent with what they represent from now on. In other words, we

expect the red points that show the imputed data have almost the same shape as blue points that show the observed data, and, as a result, the findings are satisfactory.

3.CONFIRMATORY DATA ANALYSİS & RESULTS

The data are split into train and test set in order to apply cross validation. %80 of data is used to conduct the model as a training set and remained part is used to test the model, and both sets are constructed by using the same proportion of 1 – 0 for binary dependent variable. Then, several methods are applied to provide a best fitted model by using the same train and test sets. Before these applications, it is applied one-hot encoding for categorical variables, removed the low variance features, and it has paid sufficient attention for scaling of continuous variables.

3.1. Modelling Tools

Artificial Neural Network

The first model(NN1) is constructed by using a 1-hidden layer ANN with 1 neuron. The second one(NN2) is constructed by using 2-hidden layer ANN with 1 neuron and 2 neuron, respectively. The third one(NN3) is constructed by using a 1-hidden layer ANN with 2 neuron.The last one(NN4) is constructed by using a 1-hidden layer ANN with 3 neuron.

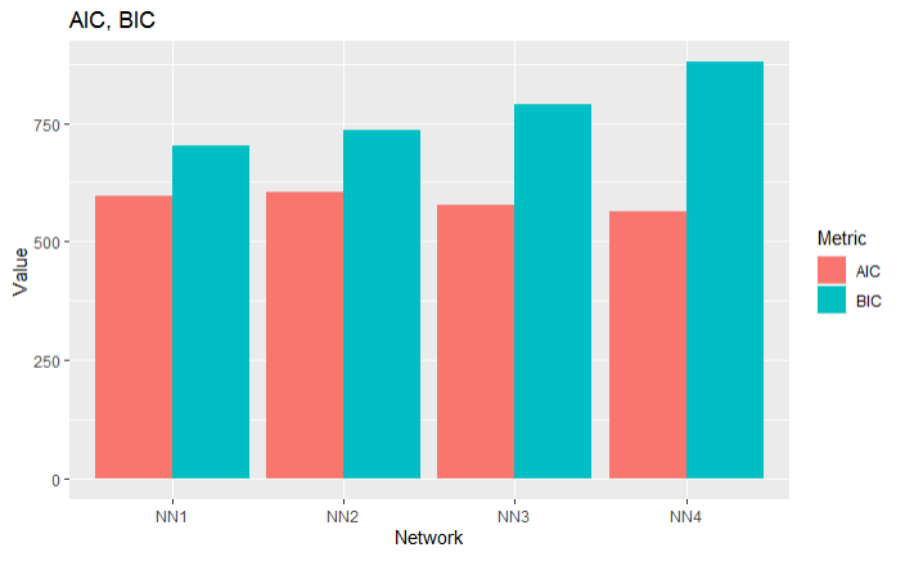

Figure 11. Bar Plot of ANN Models

The plot indicates that as we add hidden layers and nodes within those layers, the BIC increases, but the AIC decreases. Hence, it seems that the 'best' classification ANN is the simplest.

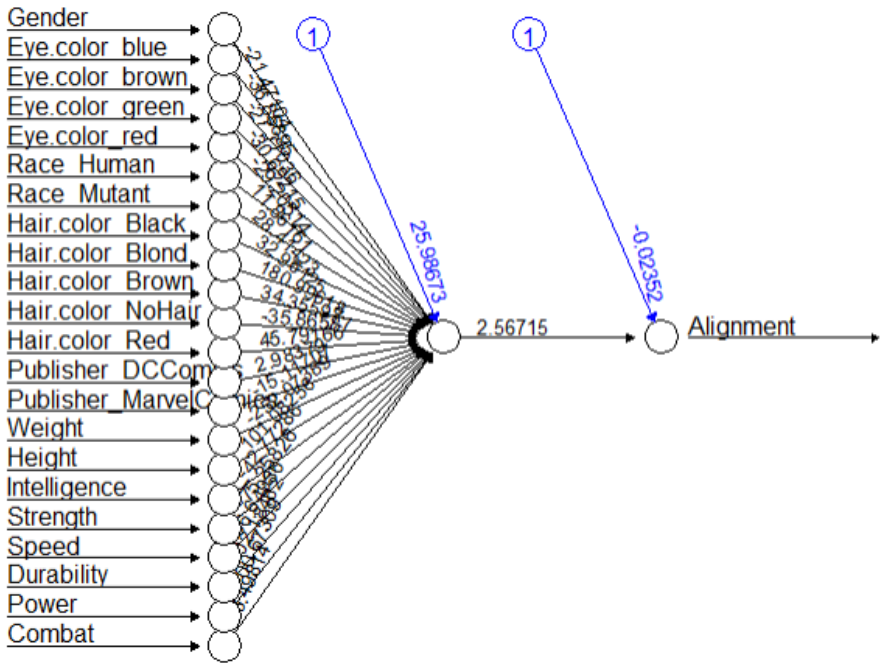

Figure 12. Best Model (NN1)

XGBoost

As can be seen in the ROC curve below, Youden index method is used to find the optimal threshold point which is equal to 0.709.

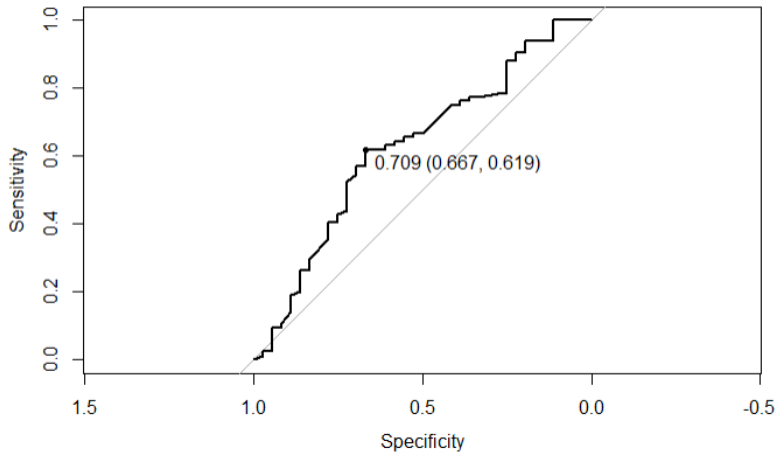

Figure 13. ROC with Youden Index Method

The predictions of the model, which is established with the optimal threshold point, are 1.8% more accurate than the predictions of the other one.

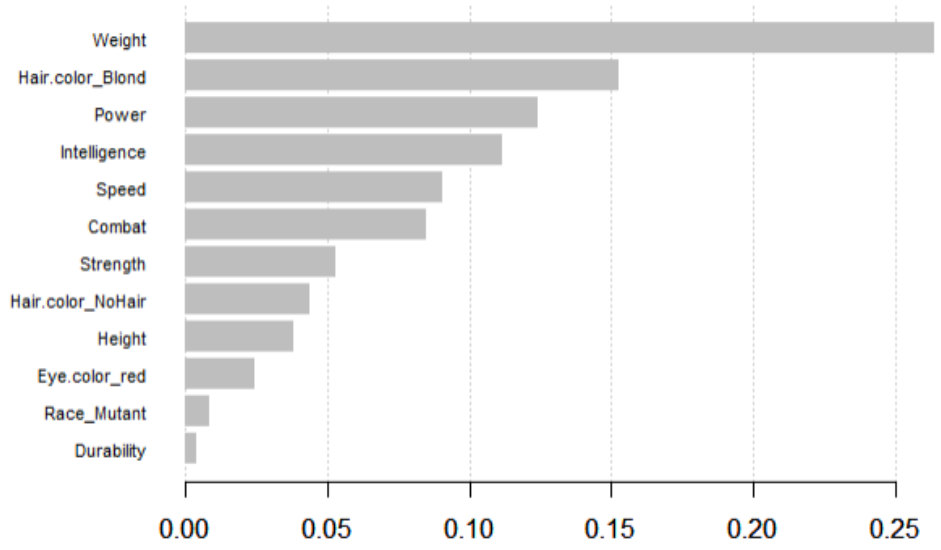

Figure 14. Variable Importance Plot

It seems that the top 3 most important features are weight, blonde hair color and power.

Random Forest

In order to assign the optimal number of trees (ntree) and the optimal number of variable per level (mtry), the plot below is constructed. In this way, the optimal number of trees is determined as 2000, and the optimal number of variable per level is determined as 2.

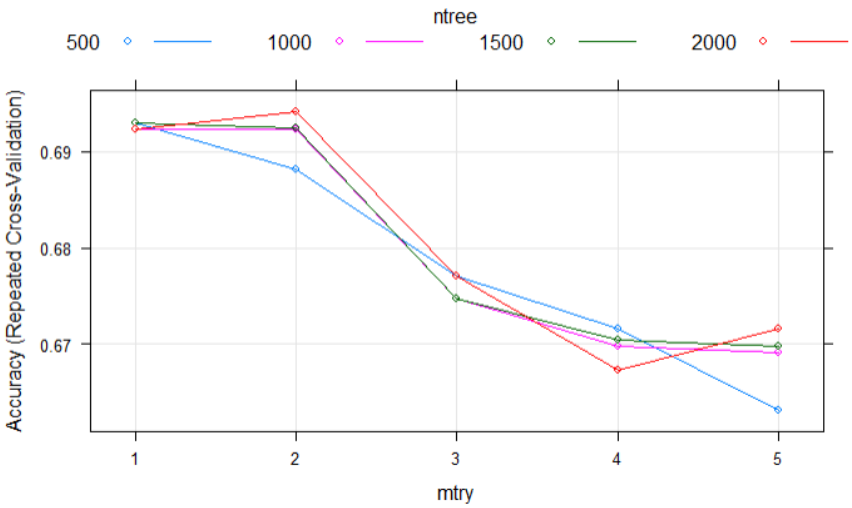

Figure 15. Line Plot of ANN Models with Different Parameters

Support Vector Machine

	Train				Test				
Model	Classification Error Rate	Accuracy	Sensitivity	Specificity	Classification Error Rate	Accuracy	Sensitivity	Specificity	Pearson Chi_sq P_value
Default	0.2249	0.7751	0.30538	0.9815	0.2750	0.7250	0.2777	0.91666	0.0119
Tuned	0.3053	0.6946	0.000	1.000	0.300	0.700	0.000	1.000	0.6612

Table 2. Table of SVM Models

After tuning the cost and gamma parameters, the p-value of Pearson's chi-squared test is greater than 0.05. Therefore, it can be concluded that the frequency distribution of observed events in the sample is consistent with the theoretical distribution. Consequently, the tuned model is selected based on the largest p-value.

Naive Bayes Classification

Logistic Regression

As can be seen in the ROC curve below, Youden index method is used to find the optimal threshold point which is equal to 0.666.

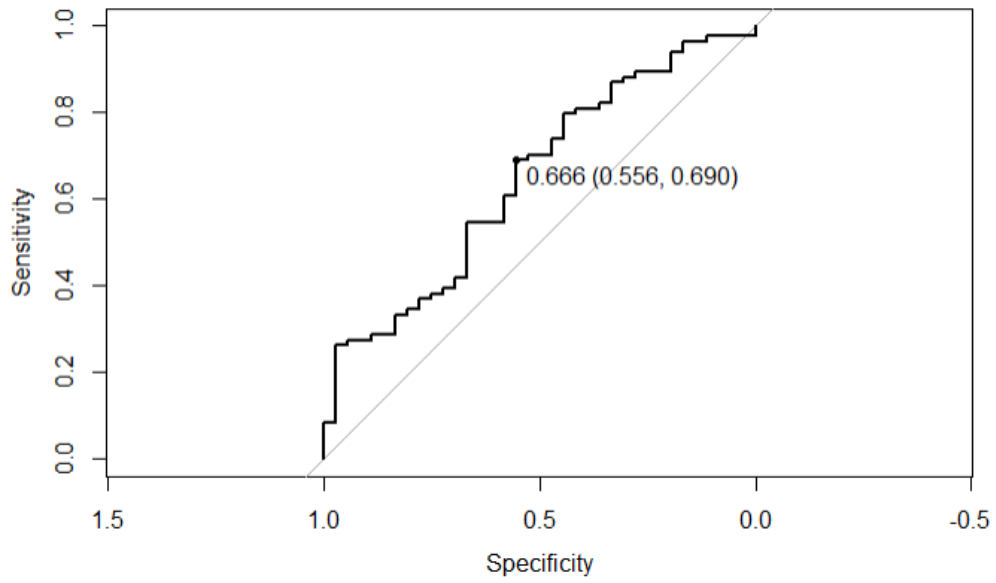

Figure 16. ROC with Youden Index Method

3.2. Model Selection

	Train				Test				
Model	Classification Error Rate	Accuracy	Sensitivity	Specificity	Classification Error Rate	Accuracy	Sensitivity	Specificity	Pearson Chi_squ P_value
LR	0.309	0.691	0.7132	0.6407	0.3583	0.6417	0.6786	0.5556	0.0119
NBC	0.3126	0.6874	0.4790	0.7789	0.325	0.675	0.500	0.750	0.0136
XGBoost	0.3345	0.6654	0.8743	0.5337	0.3666	0.6333	0.6667	0.6190	0.074
RF	0.3089	0.6929	0.9500	0.1078	0.3083	0.6917	0.9405	0.1111	0.4858
SVM	0.3053	0.6946	0.000	1.000	0.3000	0.700	0.000	1.000	0.6612
Nnet	0.3200	0.6800	0.7485	0.6500	0.4500	0.5500	0.5277	0.5600	0.4973

Tablo 3 Summary Table of Different Methods

The Support Vector Machine model seems the best in all 6 models. Since its p-value is greater than other p values and 0.05. Moreover, Classification Error Rates of both test and train samples are the lowest for 6 models and its Accuracy and Specificity values of both train and test sample are the greatest. Unfortunately, the SVM model correctly returns a positive result for 0% of good characters because of the 0% sensitivity, but it correctly returns a negative result for 100% of bad characters due to 100% specificity. Obviously, SVM is not the best method for the data set because it does not work properly in spite of all of the good outputs. In contrast to SVM, XGBoost method correctly returns a positive result for 67% of good character, and it correctly returns a negative result for 62% of bad characters.The p-value of Pearson's chi-squared test indicates that we fail to reject null hypothesis which is stating that there is no significant

difference between the observed and the expected value. As a result, it is decided to use the XGBoost to predict whether a fictional character with superpowers is a superhero or supervillain by his\her own appearance and abilities.

4.DISCUSSION & CONCLUSION

First of all, there is a relationship between gender and character alignment. Secondly, evil characters have as high intelligence as good ones. In fact, the most intelligent character is a villain. Additionally, the total count of heroes is greater than the total count of villains. This shows how difficult it is to deal with them. Accordingly, the villains tend to be more skilled than the heroes. Thirdly, there is a moderate positive relationship between ability scores. According to this information, these characters have multiple skills. However, I really wonder which is better, multiple average skills or having a single great skill? Anyway, they have both of them. As a conclusion, it is decided to use the XGBoost algorithm, since it gives the most appropriate result with respect to other 5 methods'. It seems that in order to predict character alignment, the top 4 most important features are weight, blonde hair color, power and intelligence, respectively.

5.REFERENCES

- ClaudioDavi. (2018, May 14). Super Heroes Dataset. Retrieved June 20, 2020, from https://www.kaggle.com/claudiodavi/superhero-set?select=heroes_information.csv

- Camp, M. S. (2018, August 11). Superheroes info and stats. Retrieved June 20, 2020, from https://www.kaggle.com/magshimimsummercamp/superheroes-info-and-stats?select=superheroes_stats.csv