

Trend analysis of anime and an intelligent analytical framework for anime recommendation

Himamshu Chandrashekara
Department of Information School
Syracuse University
Syracuse, New York, USA 13210
hchandra@syr.edu

Abstract— The popularity of anime has increased rapidly in the recent years. Like other forms of media, viewers often face a critical issue: “What do I watch next?”. In this study, we thoroughly examined the current method of solving this issue and determined that the learning curve to effectively utilize the current solution is too high. We developed a program to ensure easier answers to the issue. The program uses a Python-based machine learning algorithm from ScikitLearn and data from My AnimeList Kaggle to create a model that delivers what consumers want, good recommendations. We also carried out different experiments with several iterations to study the difference in accuracy when applying different factors. Through these tests, we have successfully created a reliable Support vector machine model with 67% accuracy in recommending users what to watch. We created a machine learning algorithm to classify anime to ensure that algorithm is optimized and identify possible factors that can affect it.

Keywords—EDA - Exploratory Data Analysis, ML-Machine Learning, SVM-Support Vector Machines

I. INTRODUCTION

Anime commonly refers to Japanese animated cartoons and despite common belief it is aimed at both adults and children . Over the last decade the popularity of anime has significantly increased especially in North America. We are starting to see subscription streaming services such as Crunchyroll pop up. For reference the anime industry as a whole is worth around 20 billion USD and still growing. However anime watchers suffer the same problem as other forms of entertainment watchers: the dilemma of what to watch next . The simple solution is to go on the internet and ask; however like any other medium, taste is subjective. We developed a solution that recommends anime to you. Using AI we can provide quality and objective recommendations . This solves the issues of scrolling past pages of forums of people arguing about who has the better taste. This solution is geared towards you and can be corrected to fit your taste perfectly. We are using data from MyAnimeList (MAL) Kaggle. What you rate in the past is automatically in the programming, significantly reducing training time. By extension this program can also be attuned to manga and light novels, to give users the full experience of this culture. With the growth of these systems more people are going to be watching anime and thus moving the industry forward while creating a positive feedback loop.

We have also performed data mining, data cleaning and preparation on 15278 rows and 27 columns, Exploratory data analysis and provided trend analysis of anime providing inferences from it such as Top 10 animes based on rank, popularity and analysis on genre. We have also performed visualizations such as score vs rank, popularity vs favourites, score vs favourites and histogram based on popularity of genre.

The method that we are using is building a machine learning model in Python that classifies anime similar to the user input. To do this we are using a non linear support vector machine (SVM) which classifies data points (vectors) using linear regression. This method allows us to have multiple variables of consideration which makes the predictions more accurate. SVMs are also very easy to build, come with built-in optimization, designed to work with unstructured data and scale well to high dimensional data. This makes SVMs perfect for recommending anime. We are using data from myanimelist.net which is the most well known and largest anime wiki/forum to ensure accuracy. Experiments are performed to calculate and compare accuracy for determining the most appropriate machine learning model through the implementation of support vector learning (SVM) and regression models.

II. DATASET, VARIABLES AND CHALLENGES

A. What dataset are we using?

The dataset that we are using consists of an unfiltered list of animes on the popular site <https://myanimelist.net> updated in Feb. 04, 2019. The list contains only animes. It has 15,278 anime records and 20 columns.

```
#Create dataframe to read from Anime csv
```

```
df=pd.read_csv('/content/sample_data/Anime.csv')
```

```
#Find info summary, head of each Data Frame
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 15278 entries, 0 to 15277
```

```
Data columns (total 27 columns):
```

#	Column	Non-Null Count	Dtype
0	animeID	15278 non-null	int64
1	name	15278 non-null	object
2	title_english	6122 non-null	object
3	title_japanese	15230 non-null	object
4	title_synonyms	15273 non-null	object
5	type	15273 non-null	object
6	source	15273 non-null	object
7	producers	15273 non-null	object
8	genre	15273 non-null	object
9	studio	15273 non-null	object
10	episodes	14732 non-null	float64

```

11 status      15273 non-null object
12 airing      15273 non-null object
13 aired       15273 non-null object
14 duration     15273 non-null object
15 rating       15273 non-null object
16 score       14778 non-null float64
17 scored_by    15273 non-null float64
18 rank        13669 non-null float64
19 popularity   15273 non-null float64
20 members     15273 non-null float64
21 favorites    15273 non-null float64
22 synopsis    14565 non-null object
23 background   1118 non-null object
24 premiered    4179 non-null object
25 broadcast    4402 non-null object
26 related     15273 non-null object
dtypes: float64(7), int64(1), object(19)

```

memory usage: 3.1+ MB

B. Which variables did we use?

Initially, we used all the variables to complete our Exploratory Data Analysis (EDA). However, while developing the SVM models, to predict accuracy for different scenarios we have added extra columns of genre types binary labelling to perform the analysis.

Code: df.columns.values

Output: array(['animeID', 'name', 'title_english', 'title_japanese', 'title_synonyms', 'type', 'source', 'producers', 'genre', 'studio', 'episodes', 'status', 'airing', 'aired', 'duration', 'rating', 'score', 'scored_by', 'rank', 'popularity', 'members', 'favorites', 'synopsis', 'background', 'premiered', 'broadcast', 'related'], dtype=object)

C. Challenges

- The most challenging thing of carrying out the analysis is to understand the data. If the data is not well understood, it becomes very difficult to perform the analysis because there is no clear direction on what to do on the data.
- The anime data is updated in Feb. 04, 2019, which is quite old data for performing the analysis.

III. THE PROCESS : THIS SECTION DESCRIBES THE SEQUENCE OF STEPS WE TOOK IN PERFORMING EDA, TREND ANALYSIS AND PERFORMING ACCURACY TESTS USING SVM SVC MODEL.

A. Step 1: Data Preparation and Cleaning

i. Explore columns and rows

We first downloaded the anime data which is in csv format from the Kaggle platform

(<https://www.kaggle.com/datasets/aludosan/myanimelist-anime-dataset-as-20190204>) ,

uploaded it online on google collab content,

sample_data file and created a dataframe 'df' to load the contents from csv file into the dataframe 'df'.

#Create dataframe to read from Anime csv

```
df = pd.read_csv('/content/sample_data/Anime.csv')
```

While exploring the columns and rows we found there are a lot of empty spaces before column names. Therefore we removed the empty spaces before column names using strip function

Before

```
df.columns.values
```

```
array(['animeID', 'name', 'title_english', 'title_japanese', 'title_synonyms', 'type', 'source', 'producers', 'genre', 'studio', 'episodes', 'status', 'airing', 'aired', 'duration', 'rating', 'score', 'scored_by', 'rank', 'popularity', 'members', 'favorites', 'synopsis', 'background', 'premiered', 'broadcast', 'related'], dtype=object)
```

After

```
df.columns = df.columns.str.strip()
```

```
df.columns.values
```

Output

```
array(['animeID', 'name', 'title_english', 'title_japanese', 'title_synonyms', 'type', 'source', 'producers', 'genre', 'studio', 'episodes', 'status', 'airing', 'aired', 'duration', 'rating', 'score', 'scored_by', 'rank', 'popularity', 'members', 'favorites', 'synopsis', 'background', 'premiered', 'broadcast', 'related'], dtype=object)
```

ii. Check for NA and the missing values

```
print("\n Description of each column with missing values\n", df.isnull().sum())
```

#Total number of missing values NaN at each column in a DataFrame

```
print("\n Total number of missing values NaN in the DataFrame : \n\n", df.isnull().sum().sum())
```

#Total number of columns with missing values with axis = 0 for column wise operation

```
print("\n Total number of columns with missing values : \n\n", df.isnull().any(axis=0).sum())
```

#Total number of records with missing values axis = 1 for row wise operation

```
print("\n Total number of records with missing values : \n\n", df.isnull().any(axis=1).sum() )
```

Output

```
Description of each column with missing values
animeID      0
```

name	0
title_english	9156
title_japanese	48
title_synonyms	5
type	5
source	5
producers	5
genre	5
studio	5
episodes	546
status	5
airing	5
aired	5
duration	5
rating	5
score	500
scored_by	5
rank	1609
popularity	5
members	5
favorites	5
synopsis	713
background	14160
premiered	11099
broadcast	10876
related	5
dtype:	int64

Total number of missing values NaN in the DataFrame : 48787

Total number of columns with missing values : 25

Total number of records with missing values : 14822

Code:

Remove all the null values from the dataframe

`#remove all null row`

`df.dropna(inplace=True)`

`#Check for null na value rows after updation`

`df.isnull().sum()`

Output

```
animeID 0 name 0 title_english 0 title_japanese 0
title_synonyms 0 type 0 source 0 producers 0 genre 0 studio
0 episodes 0 status 0 airing 0 aired 0 duration 0 rating 0
score 0 scored_by 0 rank 0 popularity 0 members 0 favorites
0 synopsis 0 background 0 premiered 0 broadcast 0 related 0
dtype: int64
```

IV. EXPLORATORY ANALYSIS OF DATA (EAD) - VISUALIZATIONS

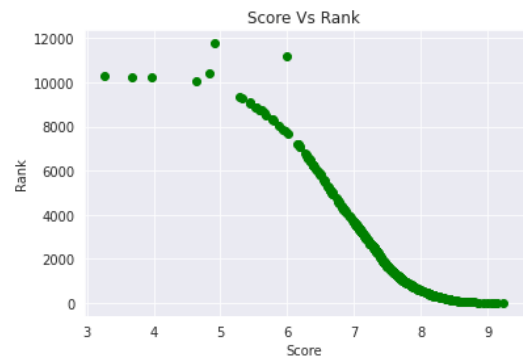


Fig 1: Scatterplot of score vs rank

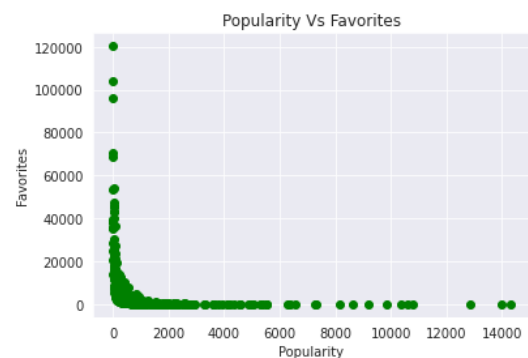


Fig 2: Scatterplot of popularity vs favorites

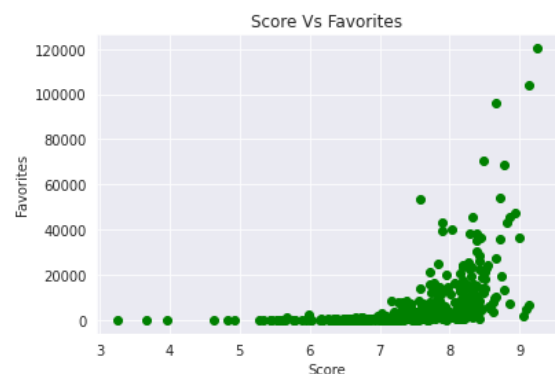


Fig 3: Scatterplot of score vs favorites

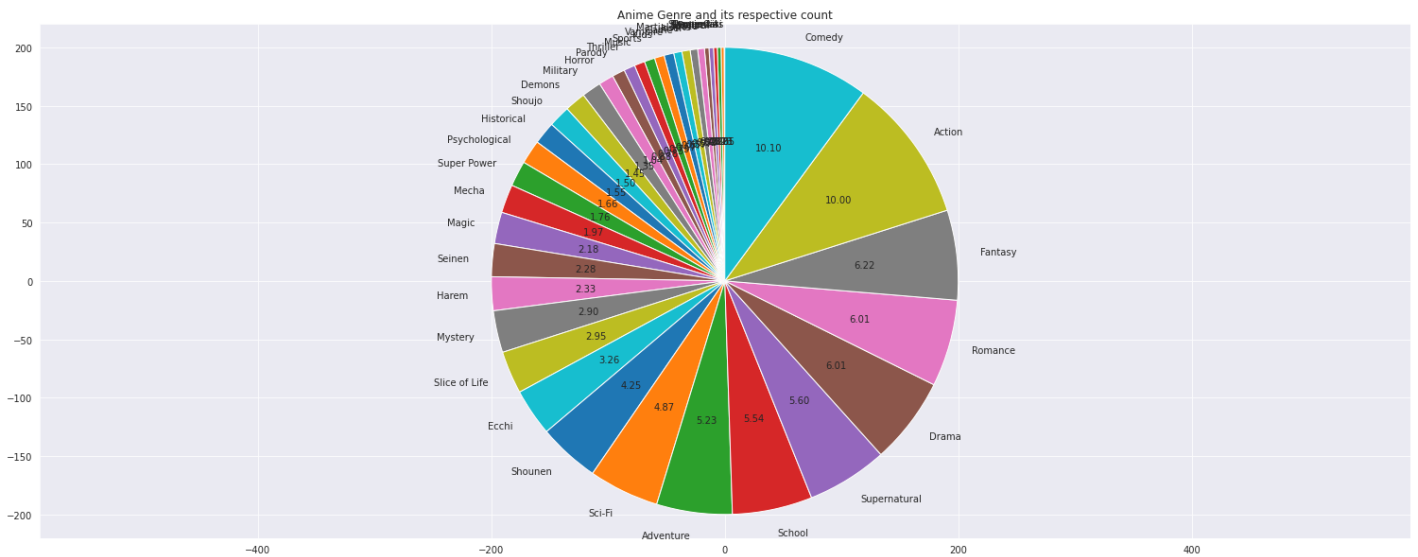


Fig 4: Pie Chart of Anime Genre and its respective count

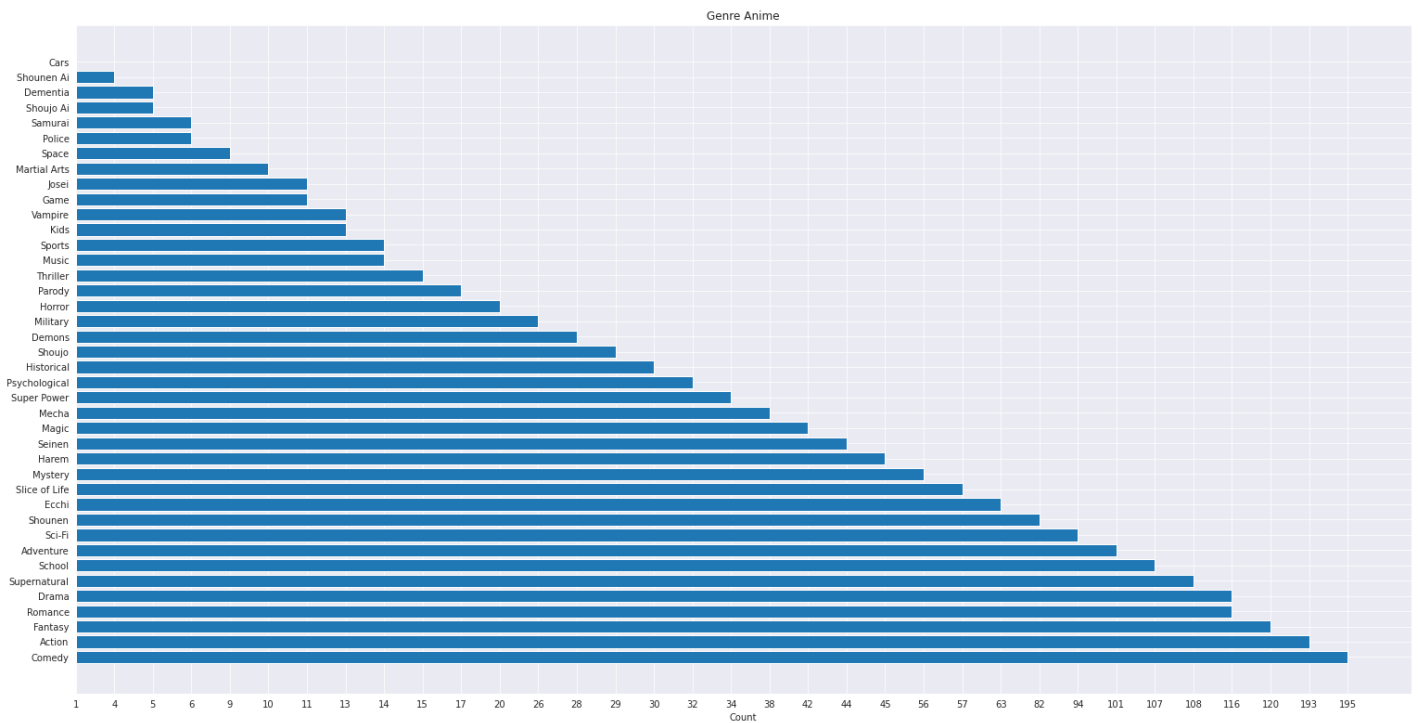


Fig 5: Histogram of Genre Anime

```
#Sort anime based on popularity from highest to lowest
df.sort_values('popularity',ascending=True).head(5)
```

animeID	name	title_english	title_japanese	title_synonyms	type	source	producers	genre	studio	...	scored_by	rank	popularity	members	favorites	synopsis	background
1311	1535	Death Note	Death Note	デスノート	[DN]	TV Manga	['VAP', 'Konami', 'Ashi Production', 'Nippon T...	['Mystery', 'Police', 'Psychological', 'Superm...	[Madhouse]	...	1107955.0	54.0	1.0	1610561.0	96146.0	A shinigami, as a god of death, can kill any p...	Death Note has been adapted into live action f...
7233	16498	Shingeki no Kyojin	Attack on Titan	進撃の巨人	[AoT]	TV Manga	['Production I G', 'Dentsu', 'Mainichi Broadca...	['Action', 'Military', 'Mystery', 'Super Power...	[Wit Studio]	...	1038161.0	116.0	2.0	1500958.0	70555.0	Centuries ago, mankind was slaughtered to near...	Shingeki no Kyojin adapts content from the fir...
6396	11757	Sword Art Online	Sword Art Online	ソードアート・オンライン	[S.A.O, 'SAO']	TV Light novel	['Aniplex', 'Genco', 'DAX Production', 'ASCII...	['Action', 'Adventure', 'Fantasy', 'Game', 'Ro...	[A-1 Pictures]	...	1007035.0	1440.0	3.0	1442099.0	53268.0	In the year 2022, virtual reality has progress...	Sword Art Online adapts the first 4 novels of ...
3769	5114	Fullmetal Alchemist: Brotherhood	Fullmetal Alchemist: Brotherhood	鋼の錬金術師 FULLMETAL ALCHEMIST	['Hagane no Renkinjutsushi: Fullmetal Alchemis...	TV Manga	['Aniplex', 'Square Enix', 'Mainichi Broadcast...	['Action', 'Adventure', 'Comedy', 'Drama', 'Fa...	[Bones]	...	826899.0	1.0	4.0	1355349.0	120331.0	"In order for something to be obtained, someth...	Fullmetal Alchemist: Brotherhood is an alterna...
10238	30276	One Punch Man	One Punch Man	ワンパンマン	['One Punch-Man', 'One-Punch Man', 'OPM']	TV Web manga	['TV Tokyo', 'Bandai Visual', 'Lantis', 'Asats...	['Action', 'Sci-Fi', 'Comedy', 'Parody', 'Supe...	[Madhouse]	...	806647.0	46.0	5.0	1195384.0	35969.0	The seemingly ordinary and unimpressive Saitam...	Episodes 1 and 2 were previewed at a screening...

5 rows x 27 columns

Fig 6: Top 5 popular anime

```
#Sort anime based on rank from highest to lowest
df.sort_values('rank',ascending=True).head(5)
```

animeID	name	title_english	title_japanese	title_synonyms	type	source	producers	genre	studio	...	scored_by	rank	popularity	members	favorites	synopsis	background	premiered
3769	5114	Fullmetal Alchemist: Brotherhood	Fullmetal Alchemist: Brotherhood	鋼の錬金術師 FULLMETAL ALCHEMIST	['Hagane no Renkinjutsushi: Fullmetal Alchemis...	TV Manga	['Aniplex', 'Square Enix', 'Mainichi Broadcast...	['Action', 'Adventure', 'Comedy', 'Drama', 'Fa...	[Bones]	...	826899.0	1.0	4.0	1355349.0	120331.0	"In order for something to be obtained, someth...	Fullmetal Alchemist: Brotherhood is an alterna...	Spring 2009
5470	9253	Steins;Gate	Steins;Gate	STEINS;GATE	[]	TV Visual novel	['Frontier Works', 'Media Factory', 'Movic', '...	['Thriller', 'Sci-Fi']	[White Fox]	...	633590.0	3.0	7.0	1139182.0	104173.0	The self-proclaimed mad scientist Rintarou Oka...	Steins;Gate is based on 5pb. and Nitroplus' .	Spring 2011
9699	28977	Gintama*	Gintama Season 4	銀魂*	['Gintama' (2015)']	TV Manga	['TV Tokyo', 'Bandai Visual', 'Dentsu']	['Action', 'Comedy', 'Historical', 'Parody', '...	[Bandai Namco Pictures]	...	82835.0	4.0	362.0	232437.0	6375.0	Gintoki, Shinpachi, and Kagura return as the f...	This is a fourth season of Gintama. In the epi...	Spring 2015
5792	9969	Gintama*	Gintama Season 2	銀魂*	['Gintama' (2011)']	TV Manga	['TV Tokyo', 'Aniplex', 'Dentsu', 'Trinity Sou...	['Action', 'Sci-Fi', 'Comedy', 'Historical', '...	[Sunrise]	...	101593.0	7.0	344.0	239983.0	4754.0	After a one-year hiatus, Shinpachi Shimura ret...	This is the second season of Gintama. In the e...	Spring 2011
7044	15417	Gintama: Enchousen	Gintama: Enchousen	銀魂 延長戦	['Gintama' (2012), 'Gintama' Overdrive', '...	TV Manga	['TV Tokyo', 'Dentsu', 'Shueisha', '...	['Action', 'Comedy', 'Historical', 'Parody', '...	[Sunrise]	...	71441.0	8.0	679.0	140930.0	1651.0	While Gintoki Sakata was away, the ...	This is a third season of Gintama. In the epis...	Fall 2012

Fig 7: Top 5 Ranked anime

A. Inferences

- Fig 1 - Higher the ranking higher is the score of the anime. Rank is inversely proportional to score
- Fig 2 - Most of the popular and favourite anime are densely present in the range of 0-20,000.
- Fig 3 - Most of the favorite anime are densely present with score range from 7-8.
- Fig 4 - The most dominant anime genres include Comedy, Action, Fantasy, Romance and least watched includes Cars, Dementia and martial arts from the Pie chart.
- Fig 5 - Shows the histogram range of anime genre same as Fig 4 explanation
- Fig 6 - Shows top 5 popular anime - Death note, one punch man, Shigeki no kyojin, sword art online and Full metal alchemist.
- Fig 7 - Shows top 5 popular anime

B. Step 3: Experimenting with different cases using SVM algorithms

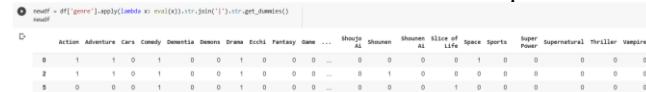
To create this program we first had to collect data to train the machine learning algorithm. We decided to collect data from a site called My Anime List (MAL) Kaggle because it is the largest databases and communities site for anime, manga and light novels. From this data, the machine learning algorithm will fit the data to calculate and compare accuracy for determining the most appropriate machine learning model through the implementation of support vector learning (SVM) and regression models.

We downloaded a CSV file that included anime_id, name, genre type, episodes, rating and number of members. However, this data was not usable as the genre types were not isolated in individual columns but instead a cluster of

strings. Pandas the data analysis and sorter does not take strings so a work around was needed.



We had to reformat the data so that genre was isolated and had an integer value attached such as the picture below [6]. When we reformatting the data we had to create multiple columns so excel could check each time if it had the string it was looking for. For example it would check if the anime had a romance tag it would put a 1 and everything else would be 0. The data would look like the example below.



This way we can actually feed these numeric values into Pandas to create data frames. The data frames are then used to train our model. For this project we decided to use Scikit-learn which is a python based machine learning algorithm and more specifically using Support Vector Machines to do classification. Our X axis is the genre and Y is the titles. From there we get a set of user inputs which will define a specific parameter and using linear regression predict or in this context suggest an anime to the user.

The below lines imports the necessary code libraries to function. Pandas is the data sorter and sklearn is the SVM machine learning code.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn import svm
```

In the experiment we wanted to determine the effect of specific rows used as training data and the accuracy of the SVM model associated with those rows. We performed a series of tests with and limited the test data to 100. The reason why we choose 100 rows as the size is because we are testing accuracy by checking how seminar our model's prediction is compared to the validation set (a set of data with known characteristics) however we do not have a validation set. We simply check the similarity between the predicted and the whole data set. This means that the larger the training set the more accurate it is going to be because the SVM would be based exactly on the test set

Case 1: Rows 0-100



Case 2: Rows 100-200



Case 2: Rows 200-300



Case 4: Rows 300-456



A comparison chart that shows the accuracy produced by the four methods is shown below:

Experiment	Accuracy
1 (Rows 0-100)	67%
2 (Rows 100-200)	50%
3 (Rows 200-300)	44%
4 (Rows 300-456)	42%

V. CONCLUSION

We found that 100 rows was not enough to trigger the aforementioned problem as our data set is about 456 rows long, and that it yielded realistic results. Our data are sorted by popularity and score meaning that “better”/ more desirable anime was at the top and less known/ lower score ones were on the bottom. On to the test itself we wanted to see if using high ranking shows would affect accuracy so we set the training data rows to 0 to 100, then 100-200 and finally 200-300 and here are the results. Surprisingly when using the top 100 anime in training the accuracy is more compared to the training with less popular anime. Rows 200-300 and Rows 300-400 almost yielded the same results. The reason why this might happen is because popular anime often come with sequels effectively doubling the effect on the model thus making it worse facing the full data set. Further research is needed.

Youtube link for the project - <https://youtu.be/3mkmgMgODEm>

REFERENCES

- [1] Anime4You: An intelligent analytical framework for anime recommendation ... (n.d.). Retrieved July 4, 2022, from https://www.researchgate.net/publication/357357234_Anime4You_An_Intelligent_Analytical_Framework_for_Anime_Recommendation_and_Personalization_using_AI_and_Big_Data_Analysis
- [2] *Cheahwen1997/a-data-analysis-on-anime-statistics*. Jovian. (n.d.). Retrieved July 4, 2022, from <https://jovian.ai/cheahwen1997/a-data-analysis-on-anime-statistics>