CHAPTER 4

INITIAL RESULTS

4.1 Overview

This chapter discusses the results and sentiment analysis of the free meal program. This chapter begins with the identification of the data set, and continues with the results of calculating the proportion of data, creating models and implementing models using machine learning techniques. The machine learning techniques used are K-nearest neighbors (KNN), Naive Bayes and Support Vector Machine (SVM). Based on the results of the implementation of these machine learning techniques, it was found that the KKN and Naive Bayes techniques had a higher percentage of accuracy and classification results compared to SVM. Details of the results and analysis are presented in the following subsections.

4.2 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is very important to do before the modeling stage. Exploratory Data Analysis (EDA) can be briefly interpreted as a process of understanding data to obtain as much information as possible. In addition, EDA can also be done to understand data patterns. The full_text column describes the public's reaction on social media X to the free meal program. Then the reaction will be analyzed to obtain the results of sentiment analysis of the program whether it is positive, negative or neutral.

index	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_t
0	1741196178123817043	Sat Dec 30 20:34:53 +0000 2023	0	sindir keras mahfud md soal program makan siang gratis prabowogibran prospek apa	1741196178123817043	NaN	NaN
1	1741191640012804207	Sat Dec 30 20:16:51 +0000 2023	0	sindir keras mahfud md soal program makan siang gratis prabowogibran prospek apa	1741191640012804207	NaN	NaN
2	1741190272342471020	Sat Dec 30 20:11:25 +0000 2023	0	sindir keras mahfud md soal program makan siang gratis prabowogibran prospek apa	1741190272342471020	NaN	NaN
3	1741190158580371915	Sat Dec 30 20:10:58 +0000 2023	0	sindir keras mahfud md soal program makan siang gratis prabowogibran prospek apa	1741190158580371915	NaN	NaN
4	1741189136109785350	Sat Dec 30 20:06:54 +0000 2023	0	sindir keras mahfud md soal program makan siang gratis prabowogibran prospek apa	1741189136109785350	NaN	NaN
5	1741180898555633979	Sat Dec 30 19:34:10 +0000 2023	4	sindir keras mahfud md soal program makan siang gratis prabowogibran	1741180898555633979	NaN	NaN

Figure 4.1 Dataset

	...	...	...	...	...			
1337	1851865142977310806	Thu Oct 31 05:53:50 +0000 2024	0	saltingan banget dia kenapa sihhh dari dulu ga...	1851865142977310806	NaN	NaN	in
1338	1851865037457231952	Thu Oct 31 05:53:25 +0000 2024	2	hi apa mas pd kenal ak abis ganti avaa	1851865037457231952	NaN	NaN	in
1339	1851806115144798370	Thu Oct 31 05:53:08 +0000 2024	0	banyak yang belum sadar bahwa presiden saat in...	1851864964790919221	NaN	abu_waras	in
1340	1851834028288057434	Thu Oct 31 05:52:57 +0000 2024	0	kamu udah makan udah sayang aku belum	1851864921203462351	NaN	jaeminmna	in
1341	1851864884654326214	Thu Oct 31 05:52:49 +0000 2024	0	keluarga besar lapas kelas iib brebes siap duk...	1851864884654326214	https://pbs.twimg.com/media/GbMlOEwakAI_aR7.jpg	NaN	in

1701 rows × 15 columns

Figure 4.2 Dataset

In Figures 4.1 and 4.2 above, it is a picture of the dataset owned after the data merging process from 2023 and 2024. The total data rows are 1701 and the columns are 15.

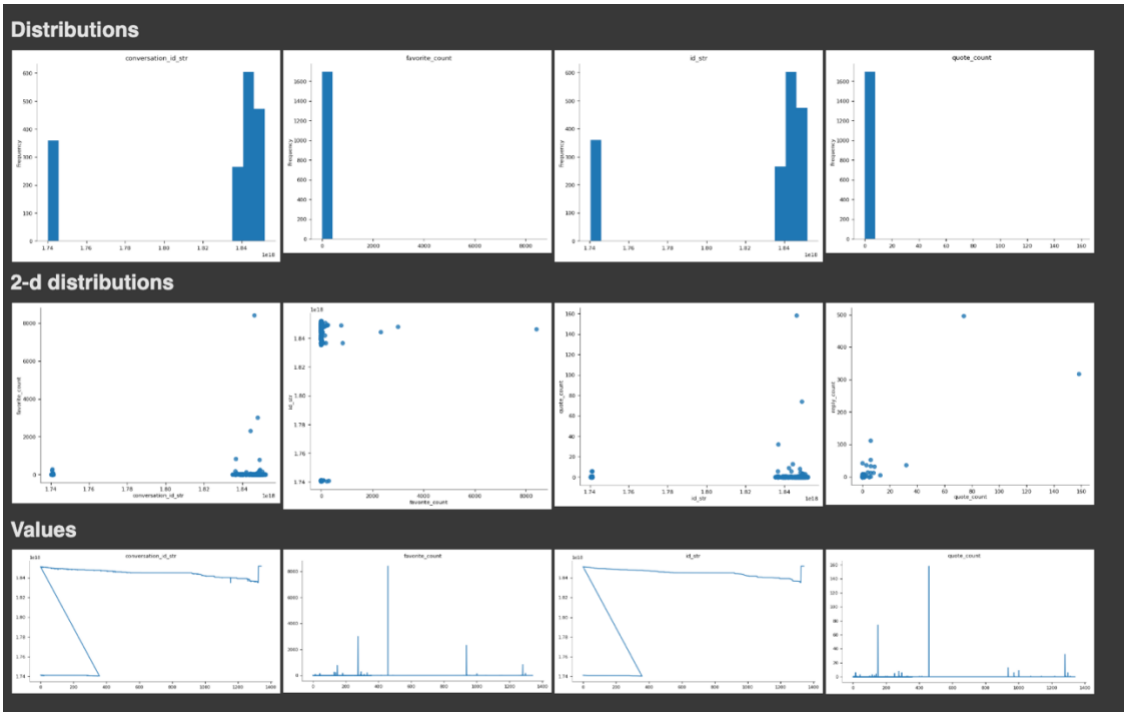

Figure 4.3 Distribution of data in each column in the dataset

The following is an explanation of the data distribution in the free meal program dataset as shown in Figure 4.3.

Distributions This section shows the distribution of values for each column of the dataset.	
conversation_id_str	This distribution shows unique conversation IDs that are mostly distributed in a certain range. Large ID values indicate that this is data taken from Twitter, as IDs are usually long numbers.
favorite_count	Distribution of the number of "likes" or "favorites" on tweets. Most tweets have a low "like" value (close to zero), indicating that many tweets receive little attention or interaction.
id_str	Like conversation_id_str, this is a unique ID for a tweet. Its distribution follows a similar long ID pattern.
quote_count	Distribution of the number of "quote retweets". Most of the data has a value of zero, indicating that most tweets are not quoted by other users. However, there are some extreme values with higher "quote" numbers.
2-d Distributions	

This section shows the relationship between variables with a 2-dimensional distribution.	
favorite_count vs conversation_id_str	This graph shows that the number of "likes" is sporadically distributed across the conversation IDs. Most of the "like" values are low, with a few outliers having high "like" counts.
favorite_count vs id_str	Similar to the previous relationship, but focused on the unique ID of each tweet. The pattern is similar, with a few dots indicating popular tweets.
quote_count vs id_str	Most tweets have a low quote value, but there are a few outliers where tweets have a significant number of quotes. This suggests that only a small number of tweets attract the attention of other users to re-comment.
Values This section visualizes the distribution of values in the form of a line:	
conversation_id_str	The lines indicate sequential IDs. This confirms that the data may have been collected chronologically.
favorite_count	The distribution pattern shows that most values are close to zero with a few peaks (outliers).
quote_count	Most of the values are close to zero, indicating tweets that are rarely requoted, but there are a few peaks with higher values.

Table 4.1 Analysis of each column in the dataset

In Figure 4.5 below are each column in the dataset and also the data type used. It can be seen that all columns are non-null, consisting of 8 objects and 7 int64.

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1701 entries, 0 to 1341
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   conversation_id_str     1701 non-null   int64
 1   created_at              1701 non-null   object
 2   favorite_count          1701 non-null   int64
 3   full_text               1701 non-null   object
 4   id_str                  1701 non-null   int64
 5   image_url               418 non-null    object
 6   in_reply_to_screen_name 338 non-null    object
 7   lang                    1701 non-null   object
 8   location                727 non-null    object
 9   quote_count             1701 non-null   int64
 10  reply_count             1701 non-null   int64
 11  retweet_count           1701 non-null   int64
 12  tweet_url               1701 non-null   object
 13  user_id_str             1701 non-null   int64
 14  username                1701 non-null   object
dtypes: int64(7), object(8)
memory usage: 277.2+ KB
```

```
data.columns
```

```
Index(['conversation_id_str', 'created_at', 'favorite_count', 'full_text',
       'id_str', 'image_url', 'in_reply_to_screen_name', 'lang', 'location',
       'quote_count', 'reply_count', 'retweet_count', 'tweet_url',
       'user_id_str', 'username'],
      dtype='object')
```

Figure 4.5 Dataset Information

```
data.describe()
```

	conversation_id_str	favorite_count	id_str	quote_count	reply_count	retweet_count	user_id_str
count	1.701000e+03	1701.000000	1.701000e+03	1701.000000	1701.000000	1701.000000	1.701000e+03
mean	1.823011e+18	10.801881	1.823043e+18	0.212228	0.887713	2.796002	1.365280e+18
std	4.260335e+16	225.938409	4.261119e+16	4.336194	14.754930	74.176484	6.082749e+17
min	1.740213e+18	0.000000	1.740399e+18	0.000000	0.000000	0.000000	1.538445e+07
25%	1.838814e+18	0.000000	1.838905e+18	0.000000	0.000000	0.000000	1.356878e+18
50%	1.845105e+18	0.000000	1.845106e+18	0.000000	0.000000	0.000000	1.684758e+18
75%	1.846905e+18	0.000000	1.847102e+18	0.000000	0.000000	0.000000	1.699318e+18
max	1.851873e+18	8417.000000	1.851873e+18	158.000000	497.000000	2966.000000	1.844122e+18

Figure 4.6 Dataset Description

In Figure 4.5 Dataset Description, there are extreme values (outliers) in favorite_count, quote_count, reply_count, and retweet_count indicating that some tweets are very viral, which may be caused by content factors or accounts with many followers.