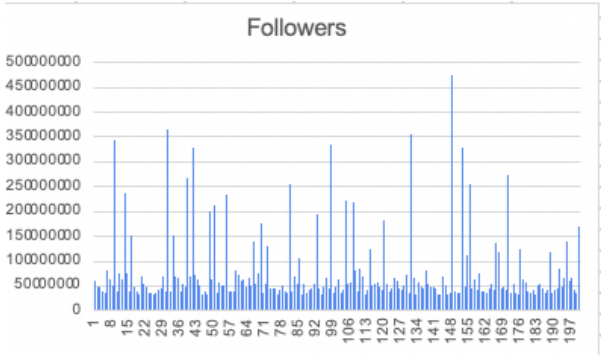
Karla Hildebrand

Proposal

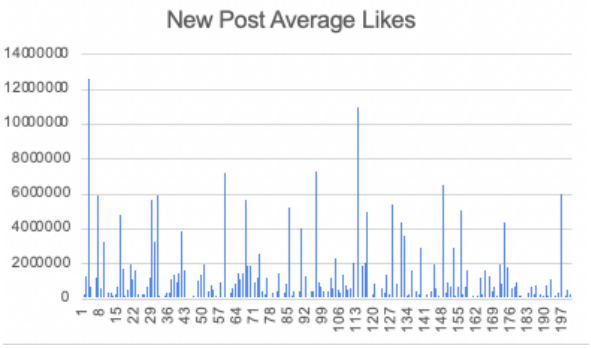
My data set describes the top influencers on Instagram. I will predict the average likes per post number of other variables as inputs. I retrieved the data from Kaggle at https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned. This data was recovered in August 2022. Each row in the data set is a given influencer on the Instagram platform, with 200 total observations. There is a ranking from number one, being the most followed, all the way to the 200th most followed. There are ten total variables in the data set. The input variables are user, rank, channel information, influencer score, followers, amount of posts, 60-day engagement rate, total likes, country of origin, and average likes per new post. The output variable is the total average likes. Most data is numeric except for the influencer's username and country because it is nominal. They are all individualized to that user. Rank is the rank of the influencer based on the number of followers they have.

Figure 1

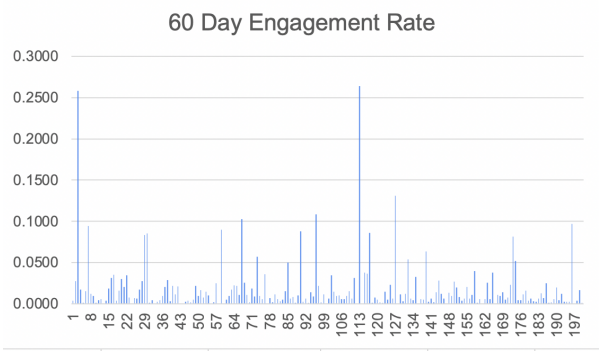
Variable Graphs and Stats



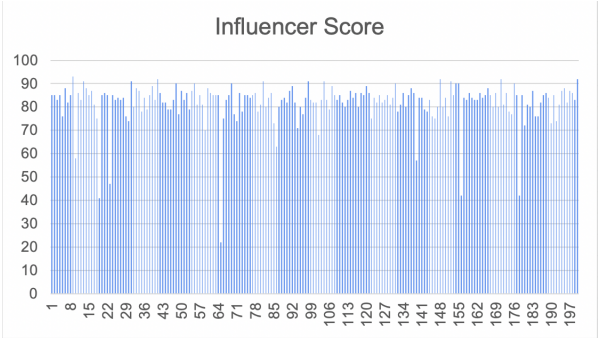
Standard Deviation	73,687,269.79
Minimum	32,800,000
Maximum	475,800,000
Mean	77,409,500
Median	50,050,000
Mode	33,200,000
Range	443,000,000



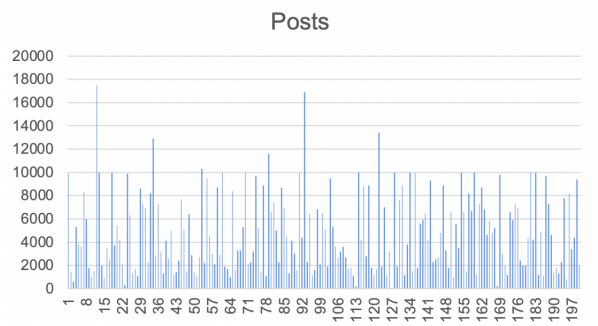
Standard Deviation	1,858,321.85
Minimum	0
Maximum	12,600,000
Mean	1,208,133
Median	532,150
Mode	0
Range	12,600,000



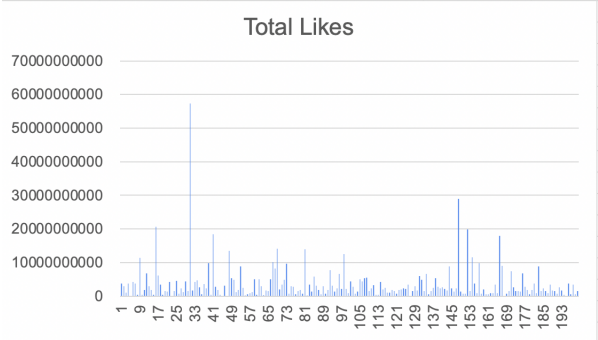
Standard Deviation	0.0332
Minimum	0.0001
Maximum	0.2641
Mean	0.0189
Median	0.0087
Mode	0.0002
Range	0.264



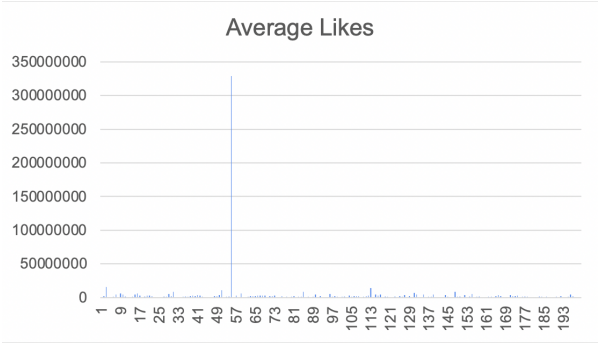
Standard Deviation	8.88
Minimum	22
Maximum	93
Mean	81.82
Median	84
Mode	85
Range	71



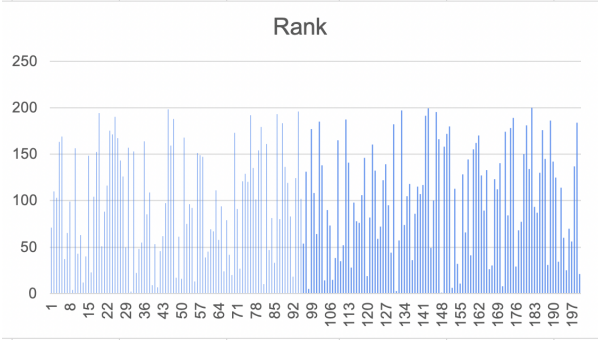
Standard Deviation	3,389.82
Minimum	100
Maximum	17,500
Mean	4,768
Median	3,800
Mode	10,000
Range	17,400



Standard Deviation	5,561,938,629
Minimum	18,300,000
Maximum	57,400,000,000
Mean	3,658,112,500
Median	2,000,000,000
Mode	1,600,000,000
Range	57,381,700,000



Standard Deviation	23,240,428.35
Minimum	65,100
Maximum	329,000,000
Mean	3430459
Median	1,100,000
Mode	1,100,000
Range	328,934,900



Standard Deviation	57.88
Minimum	1
Maximum	200
Range	199
Mean	100.5
Median	100.5

Influence score is the rating method for mentions that groups them according to importance and appeal. Depending on your subscription, a number with a range of 0 to 100 will show in various places on your platform. The score is retrieved by a company named The ROI Influencer. It is a new technology that utilizes proprietary algorithms to measure and rank the most effective influencers on Facebook, Twitter, and Instagram (Convergence, n.d.*)*. Followers are the number of followers of the user. Posts total are the number of posts they have made since their account has been active. The last 60 days' engagement rate metrics track how actively involved a user's content is with their audience/followers (Engagement, 2022). The level of interaction an influencer typically obtains on their content is gauged by the engagement rate in influencer marketing. It is just the proportion of an influencer's audience that engages with their material. To calculate an influencer's engagement rate, total the engagements (likes + comments + retweets, etc.) across all of the influencers posts on a particular profile (e.g., Instagram profile), then divide by the total number of followers and divide that by the number of posts (to get the

per post average) (Lewis, 2022). Total likes are the number of likes the user has accumulated on their posts.

The user's country or region of origin is where they are from. Lastly is the variable I am predicting, the average likes on Instagram posts. Below is a sample of my data and how it is set up in the table.

Table 1

Example Data

Rank	User	Influencer Score	Posts	Followers	60 Day Eng Rate	New Post Avg Likes	Total Likes	Country	Avg Likes
71	manchesterunited	85	9900	59600000	0.0041	241600	3800000000	United Kingdom	381800
110	blackpinkofficial	85	1400	48200000	0.0277	1300000	3000000000	South Korea	2100000
103	thv	83	600	49300000	0.2580	12600000	987400000	South Korea	15400000
163	anushkasen0408	85	5300	38300000	0.0175	658300	3700000000	India	695400
169	laudyacynthiabella	76	3800	36900000	0.0012	43400	66700000	Indonesia	172800
37	nasa	88	3600	81300000	0.0153	1200000	4200000000	United States	1200000
65	sooyaaa__	82	8300	62900000	0.0943	5900000	3800000000	South Korea	4500000
99	michelleobama	85	6000	50700000	0.0122	611200	421700000	United States	700500
4	selenagomez	93	1800	342700000	0.0097	3300000	11500000000	United States	6200000
156	mahi7781	58	1000	39100000	0.0017	0	439400000	India	4100000

When I retrieved my data, it was ranked in order. Therefore, I had to randomize it through Excel to get a better representative sample of training and test sets in Weka. In table 1, One of the users listed above has zero new post average likes, and eight users have zero in their new likes per post. I believe it is because that user does not actively post on the platform. Therefore, they cannot calculate how many likes the user post gets because they do not often post at all. A lag occurred between the date it was taken and their posting time. The influencer might be active on other platforms besides Instagram to have so many followers on Instagram.

Each variable, including total number of posts, average likes per post, country of origin, rank, username, 60-day engagement rate, influencer score, and total likes, will be a source to help predict the average like per Instagram influencer. For instance, the amount of total likes correlates to how long the user has been active. The number of posts can also indicate how long

the user has been on the platform. The more posts the user has on their page could indicate the amount of time the user has been on Instagram. Engagement rating and influencer score can help show how much the user has been on the platform, alluring their audience and putting out good content that makes other users like their posts. New post average likes will indicate whether one user has more likes than another because of how many likes they get on more recent posts. Total likes contribute by being an indicator of who has the most likes and probably gets the most likes on their average post. Lastly is the user's country of origin. This can help because there are different time zones throughout these countries which can be why they have more or fewer likes on them. Most of the top influencers live in the United States.

Some input variables not included in the data set are the fastest-growing user throughout their time on Instagram, users with a stagnant number of followers, or users losing followers. The findings from my predictions could be interesting to see how many people have active followers on their page and how many people have “ghost” followers, or followers who do not like the user's post but still follow the account. This would also be an interesting way to see how the Instagram algorithms are set up for what the followers see, how the platform predicts what the user would prefer to see more than others, and what people are interested in seeing.

The models I will use to make my predictions are decision trees and K-nearest neighbor. A decision tree will help predict the outcome based on the splits of the input variable. The k-nearest neighbor will assist in locating other individuals who have similar input attributes and then predicting average likes based on those neighbors.

There were missing values in my data under the country of origin. I had to fill in the country of origin based on where that user was from, not where that user currently lives. Many

users are from different countries but live in the United States. However, the users are not initially from the United States.

I processed the data in other ways too. The numeric data is not a precise number because of how it was retrieved from Kaggle; all my inputs end in zeros. Weka was reading all of my data as nominal, not numeric. There were percentage signs that I had to convert into decimals, and some of the data included K for thousands, M for millions, and B for billions, and only to the nearest whole number. The data only provided a one-decimal-place rounding. Therefore, I had to go back via Excel and insert the true zeros. I converted them to numeric data instead of nominal data because Weka was not reading it as numeric data because of the letters inside the values. Here is an example of the data before.

Table 2

Table Before Adjustments

Rank	User	Influencer Score	Posts	followers	Avg Likes	60 Day Eng Rate	New post avg like	Total Likes
13	Nike	90	0.95k	234.1m	329.0k	0.08%	181.8k	313.6m
14	taylorswift	91	0.53k	222.2m	2.4m	1.01%	2.3m	1.3b
15	jlo	89	3.2k	220.4m	1.7m	0.62%	1.4m	5.3b
14	virat.kohli	87	1.4k	11.8m	3.5m	0.96%	2.0m	4.9b

Table 3

Here is my data after making the necessary adjustments.

Rank	User	Influencer Score	Posts	followers	60 Day Eng	New post avg	Total Likes	Avg Likes

						like		
13	nike	90	9500	234100000	0.0008	181800	313600000	3290000
14	taylorswift	91	5300	222200000	0.0101	2300000	1300000000	2400000
15	jlo	89	3200	220400000	0.0062	1400000	5300000000	1700000
14	virat.kohli	87	1400	21180000	0.0096	2000000	4900000000	3500000

Since my data did come in order, I had to randomize the data and upload it to Weka to get a more accurate prediction. Because I used the 66% split, the training and test set observations would not be random. I had to switch the order of the variables in the original data set to put average total likes at the end to make Weka calculate my predictive variable. I assumed that the significant variables for predicting the average likes are ranking, followers, total likes, and influencer scores.

Before running the models, I found the mean for the average likes, which is 3,430,459. I found the mean absolute error for the test set, which is 5,850,940.72. This will be the baseline for the models.

Using the K- nearest neighbor, I uploaded the data to Weka and ran it through the “Lazy IBk” classifier. I kept all the same default settings. I ran the data using the K- nearest neighbor 1-19 only using the odd numbers. When I finished running through all the numbers, I went back through them and found the best prediction of average total likes. I used the mean absolute value (MAV) to determine the lowest possible prediction since I am working with numeric data. One of the methods most frequently used to assess the accuracy of forecasts is the MAV. It illustrates the Euclidean distance between measured true values and forecasts. I had to find the number with the lowest MAV.

Table 4

K- Nearest Neighbor Outcome for Five

Mean Absolute Error	1,071,681.79

In Table 4, the MAE had a higher accuracy than the others. I chose five as my value for K because it had the lowest MAE.

I did a 66% split with the decision tree model and kept the default settings. I had to take out the usernames for it to run because the usernames were working when I tried to run the tree. The top node was new post-average-likes. The tree was too big to visualize on this paper. However, the MAE from my original tree is shown in Table 5.

Table 5

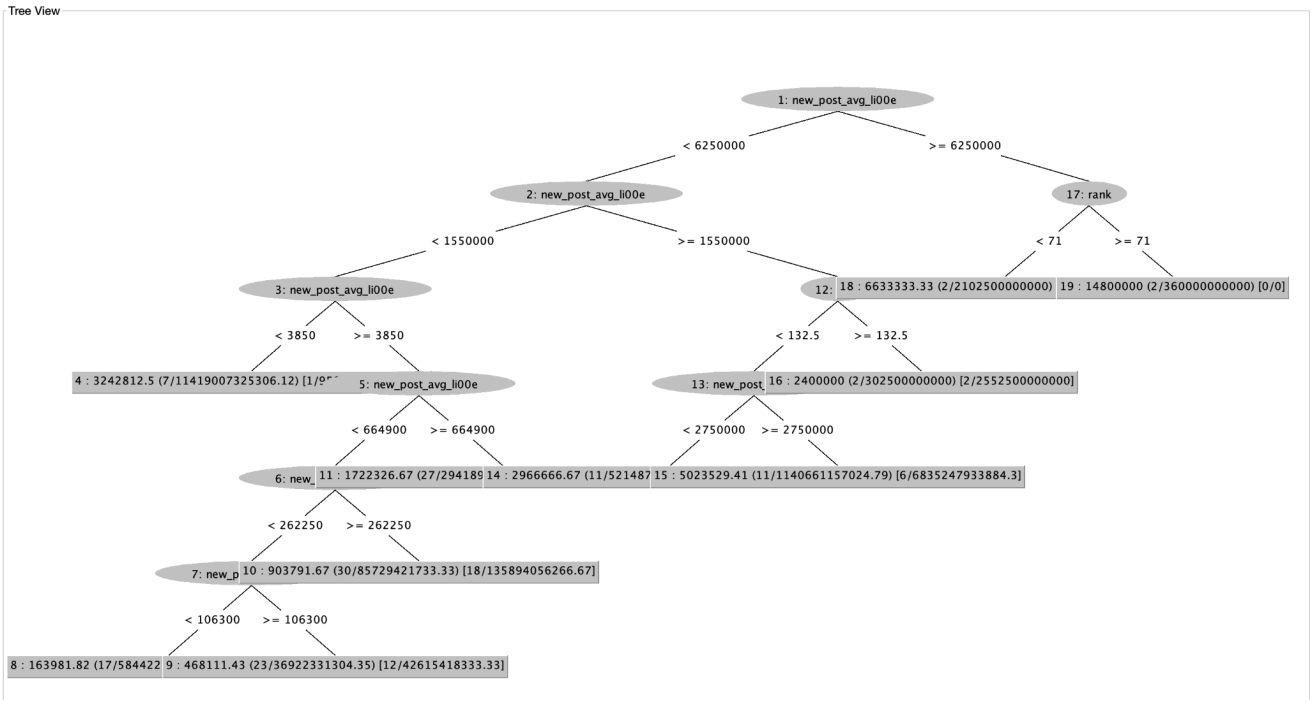
First Ran Tree Outcome

Mean Absolute Error	1,330,216.97

Then I removed the country name and reran the model. I got a tree visualization and a lower MAV, but the top node was still the same.

Figure 2

Second Tree Without Country of Origin Variable



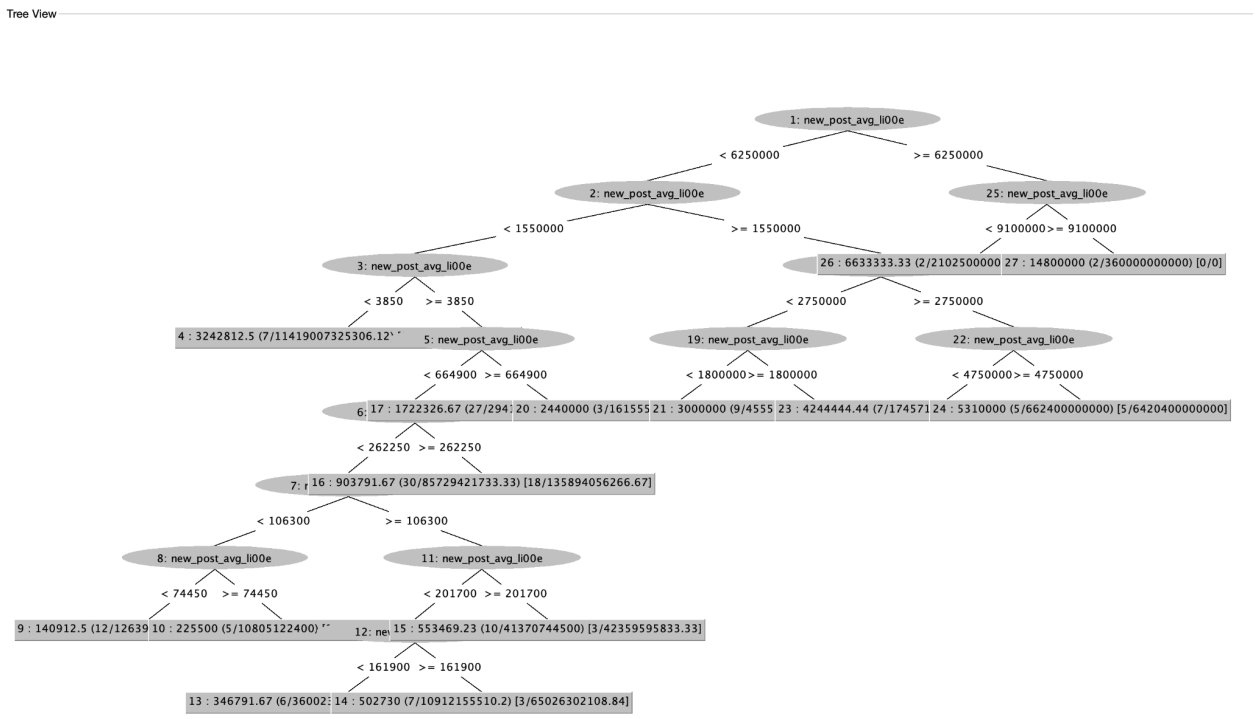
Mean Absolute Error	1,013,060.02

This interested me initially, so I took out another input variable to see if it would drop the MAV again. I took out the rank to see if it would make a difference, and sure enough, it did. Then I took out the 60-day engagement and ran it, and MAE lowered again. Then I took out the influential score. I ran it, and it lowered again. There were too many input variables to keep up with because there were nine input variables, and Weka was trying to calculate them simultaneously, which caused the calculations to be inaccurate.

I finally used the following variables in Figure 3: followers, new posts average likes, and average likes per post. When running the model, I received a lower MAV than any of the other previous trees ran.

Figure 3

Tree with Followers, Few Posts Average Likes, and Average Likes Per Post



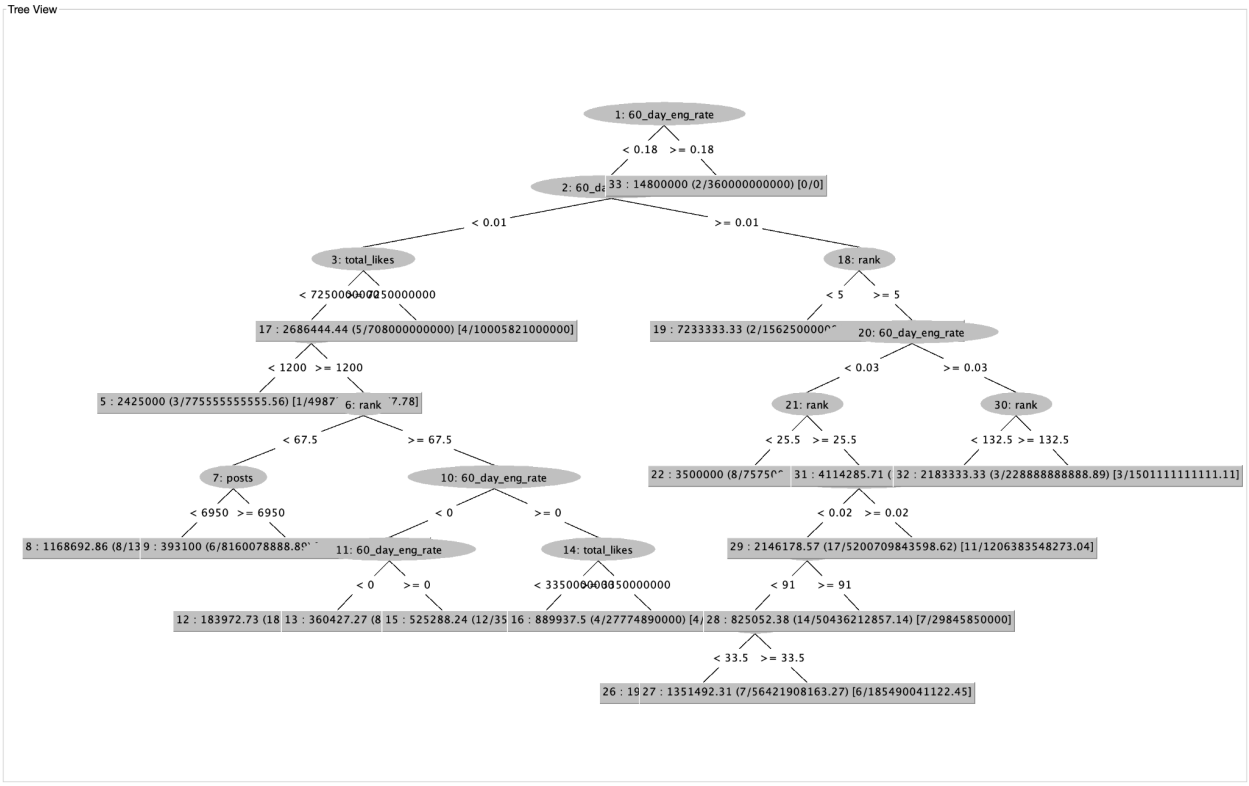
Mean Absolute Error	777,862.84

After the calculations were finalized, I realized that Weka was predicting average likes per post to average likes which honestly is essentially the same thing. Therefore, I decided to re-run the whole tree taking out average likes per new post because it was an inadequate prediction of what I was trying to predict. Then, I ran the model without the total average likes to see how the other variables would perform in predicting the total average likes in general.

I made a new tree again taking out the country and the usernames of the influencers. Then I started narrowing down more variables to take out and improve the accuracy of the tree. I believe with nine variables the tree tried to use all of them when there were only a handful that were more dependent on the prediction than others.

Figure 4

Tree Without Average Likes Per New Post



Mean Absolute Error	1,161,798.18

Weka calculated the following variables: posts, followers, 60-day engagement rate, and total likes. These were suitable to predict the outcome. The most dependable variable, the 60-day engagement rating, is located at the top node of the tree. It was a surprising result to see that the 60-day engagement rating was the most important rather than the other variables. This variable does not appear significant when compared to other well-known variables, such as the number of posts, total likes, and followers.

The best model to predict average likes is the K-nearest neighbor. It was the lowest by 90,116.39 less error than the decision tree, comparable to the baseline absolute mean error of 5,850,940.72, with a difference of 4,779,258.93.

Doing further studies on the 60-day engagement rating would be what many influencers need to gravitate toward rather than having more likes, followers, and posts. Continuous activity on accounts creates a great number of likes per post, and users become active followers engaged in the influencer's content. This also can be applicable to other platforms as well, such as Tik-Tok, Twitter, or Facebook. The more engaged the user is, the more activity the page will gain.

The studies in this area could help young entrepreneurs who post their products on social media sites. Realizing that 60-day engagement ratings are necessary for users to influence followers, market researchers could use similar data to increase followers and likes for specific products.

References

Engagement rate. Sprout Social. (2022, June 29).*Engagement rate*. https://sproutsocial.com/glossary/engagement-rate/

How does mention calculate the influence score? Mention Help Center. (n.d.). https://en.support.mention.com/en/articles/2046054-how-does-mention-calculate-the-influence-score#:~:text=The%20Influence%20Score%20is%20Mention's,differenttly%20depending%20on%20the%20Source.

JMP. (n.d.). *Correlation coefficient*. https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-correlation/correlation-coefficient.html

Lewis, D. (2022a, August 26). *How to measure an influencer's engagement rate (a scientific approach)*. Scrunch. https://scrunch.com/blog/measuring-an-influencers-engagement-rate/

Lewis, D. (2022b, August 26). *What is engagement rate, and why is it important in influencer marketing?* Scrunch. https://scrunch.com/blog/influencer-engagement-rates/

ROI Influencer Media the ROI Influencer Score. (n.d.). *the convergence of top influencers, technology, and data*. https://roiinfluencer.com/roi-influencer-score/