Leveraging AI to Analyze Factors Relating to Social Anxiety

Neha Krishnan

11/18/2023

Abstract

With our changing lifestyles and the rise of technology, there has been a steady rise of reported social anxiety cases throughout the years. Exploring how different factors contribute to social anxiety disorder may help people understand significant causes in a cost effective and relatively reliably way. Our research aims to leverage ML to understand what factors play the most critical role in the presence of social anxiety disorder in a person. Additionally, through training different models, we aim to determine the capability of AI in predicting one's social anxiety. To analyze the factors relating to social anxiety, we clustered our data to find relationships within the background data (e.g: family history, age, gender, symptoms) and one's social anxiety diagnosis. Our findings suggest that fears and physical conditions may contribute more to social anxiety than a person's history/background.

1. Introduction

With around 12.1% of US adults experiencing social anxiety disorder at some point in their lives, SAD seems to be a very prevalent issue in today's society. Modern innovations, such as social media, that contribute to lower self esteem and fewer social interactions along with a reduced social stigma may be contributors to the steady increase in reported cases of social anxiety. Furthermore, there is an increasing need for us to be able to identify key indicators of social anxiety and one's susceptibility to the disorder in an efficient and cost effective way. Because the goal of our research is to understand how AI can be utilized to analyze the relationships between SAD and certain contributing factors, along with considering how AI might be used to predict one's social anxiety, this classification problem involves both supervised and unsupervised learning.

2. Background

Approaches that have been used in the past used to evaluate social anxiety disorder include using multiple researchers to work together in an attempt to analyze the data physically. In the study "Factors related to the association of social anxiety disorder among adolescents: a systematic review," they used this tactic to conclude the contributing factors of female gender, peer acceptance, and the presence of other conditions (depression, OCD, etc.)[1]. In this paper, they used multiple examiners to help to analyze a large amount of sources (409). While the sheer amount of sources allows this approach to be very detailed and comprehensive, doing a lot of the work by hand necessitates a lot of time, especially with 409 sources. Furthermore, with the use

[1] Lima Dias da Cruz, Diniz de Carvalho Martins, Rejane Beserra Diniz. "Factors related to the association of social anxiety disorder and alcohol use among adolescents: a systematic review." Jornal de Pediatria. 93(5): 442-451, 2017.

of AI, my approach aims to help form many of these connections without the expensive resources and significant amount of time.

3. Dataset

We used the dataset “Data for: Development and use of a clinical decision support system for the diagnosis of social anxiety disorder” from Mendeley Data. This data consists of 240 samples, and is entirely numerical. Background information was collected for each sample regarding their age, education level (high school, diploma undergraduate, bachelor degree, masters degree, post-graduate), gender, family history of anxiety or depression, and occupation (student, faculty member, employee, self-employment, unemployed). We use this background information to find relationships with the symptoms/fears listed in the data and train our models.The dataset also includes information about whether individuals exhibit certain fears associated with social anxiety disorder, such as ATF (fear of being at the center of attention), EAF (fear of eating in front of another person), TKF (fear of speaking in public), CMT (fear of attending parties), DEF (fear of eating/drinking in public places), SMF (fear of contact with strangers), ERF (fear of getting into a room with other people), and DAF (fear of disagreements with strangers). Physical indications of social anxiety disorder are also included: HR (has heart palpitations), SW (has sweating), TR (has tremor), DR (has dry mouth), BR (has hard breathing), CK (has suffocating feeling), CP (has chest pain), NS (has gastrointestinal discomfort and nausea), DZ (has dizzy, weak, or sick feelings), UR (has feeling of beng unreal), UB (has fear of losing balance), MD (has fear of being crazy), and TG (has numbness or moaning). We used some of these fears or physical indications when training our models, depending on whether they increased or decrease the accuracy. At the end, it is stated whether or not the person is diagnosed with social anxiety disorder (hasSAD), along with the results of the Liebowitz Social Anxiety Scale questionnaire (LSAS) and the results of the Social Phobia Inventory questionnaire (SPIN). We used ‘hasSAD’ column as the value for the models to predict, as it was a simple binary optionTo better analyze relationships between age and other columns of the dataset, we categorized the column into four “bins”: Age_Group_teen (teens 13-18), Age_Group_young (young adults 19-34), Age_Group_middle (adults 35-50), and Age_Group_elder (elders 51+).

Table 1: Dataset Columns

	id	EducationLevel	Gender	HasFamilyHistory	Occupation	ATF	EAF	TKF	CMT	DEF	...	UB	MD	TG	hasSAD	SPIN	LSAS	Age_Group_teen	Age_Group_young	Age_Group_middle	Age_Group_elder
0	1	4	1	0	3	4	2	6	2	0	...	1	0	0	1	23	39.0	0	1	0	0
1	5	4	1	0	3	3	0	3	1	0	...	0	0	0	0	20	43.0	0	1	0	0
2	7	5	0	1	1	4	1	7	0	0	...	0	0	0	1	33	50.0	0	1	0	0
3	8	5	1	1	3	4	0	6	1	0	...	0	0	0	1	30	44.0	0	0	1	0
4	9	5	1	0	2	5	1	5	1	1	...	0	0	0	0	16	NaN	0	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
209	235	2	1	0	4	5	8	7	10	0	...	0	0	0	1	28	34.0	1	0	0	0
210	236	5	0	1	1	6	2	5	2	3	...	1	1	0	1	29	78.0	0	1	0	0
211	237	4	1	0	1	10	7	9	10	6	...	0	0	0	1	54	NaN	0	1	0	0
212	238	6	0	0	2	0	0	2	2	0	...	0	0	0	0	11	42.0	0	0	1	0
213	239	4	0	1	1	7	7	8	6	5	...	1	0	0	1	35	70.0	0	1	0	0

4. Methodology / Models

Classification Using Logistic Regression

Our logistic regression model predicts the probability of a sample having social anxiety disorder or not, which is later turned into 0 or 1 predictions for each individual. To create this model, we separated the dataset into 30 percent training and 70 percent testing data. We initially added random features to the training and testing data and used the model to predict the hasSAD column. In the end we found the optimal input data columns to be the columns ‘Age_Group_teen’, ‘Age_Group_young’, ‘Age_Group_middle’, ‘Age_Group_elder’, ‘Gender’, ‘EducationLevel’, ‘Occupation’, ‘ATF’, ‘EAF’, ‘ERF’, ‘DAF’, ‘HR’, ‘SW’, ‘DR’, ‘NS’, ‘UR’, ‘MD’, and ‘TG.’ We then initialized the model and fit it with the data, before creating predictions on the testing data. To stay consistent with the other models we used, we decided to use mean absolute error to measure accuracy.

Figure 1: Logistic Regression Implementation

```
X = ['Age_Group_teen','Age_Group_young', 'Age_Group_middle', 'Age_Group_elder', 'Gender', 'EducationLevel', 'Occupation', 'ATF', 'EAF', 'ERF', 'DAF', 'HR',

X_train, X_test = train_df[X], test_df[X]
y_train, y_test = train_df['hasSAD'], test_df['hasSAD']

model = linear_model.LogisticRegression()
model.fit(X_train, y_train)
preds = model.predict(X_test)
#print(r2_score(y_test,preds))
mae = mean_absolute_error(y_test, preds)
print(mae)
```

Predictions Using Random Forest Regressor

Our random forest regression model uses multiple decision trees to form a prediction about the hasSAD value of each sample. To create this model we follow a similar process as the logistic regression model: we first separate the dataset into 30 percent training and 70 percent testing data. In this model, however, we found the optimal input data columns to be ‘Age_Group_teen’, ‘Age_Group_young’, ‘Age_Group_middle’, ‘Age_Group_elder’, ‘Gender’, ‘HasFamilyHistory’, ‘EducationLevel’, ‘ATF’, ‘TKF’, ‘CMT’, ‘SMF’, ‘ERF’, ‘SW’, and ‘DR’. Then, we simply

initialize the model, fit the training data, and make predictions based on the testing inputs. Analogous to our logistic regression model, we again use mean absolute error to measure the accuracy.

Figure 2: Random Forest Implementation

```
#Random Forest Classifier
X = ['Age_Group_teen', 'Age_Group_young', 'Age_Group_middle', 'Age_Group_elder', 'Gender', 'HasFamilyHistory', 'EducationLevel', 'ATF', 'TKF', 'CMT', 'SMF',
X_train, X_test = train_df[X], test_df[X]
y_train, y_test = train_df['hasSAD'], test_df['hasSAD']


rf = RandomForestRegressor()
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)
mae = mean_absolute_error(y_test, y_pred)
#print(r2_score(y_test,preds))
#print(accuracy_score(y_test,y_pred))
print(mae)
```

Analyzing Columns Using KMeans Clustering

KMeans clustering is an unsupervised learning technique that clusters data based on defining centroids and assigning points to the closest centroid. Through repeatedly averaging the clusters to find the new centroid and reassigning the data until the centroid stays constant, this technique allows us to make inferences from the data based on the locations of each centroid. Similar to the previous models used, we made a list of columns to train the kmeans model with, and decided to use all of the columns except the last three, which function as scales to measure an individual's social anxiety. After training the model and fitting it with the columns, we decided to predict the 'ATF' column, because the hasSAD column was binary, and this column is very correlated with one's SAD. We used these predictions to create our cluster assignments and centroids, and plotted the points on a scatterplot.

Figure 3: KMeans Implementation

```
features = df_1[['Age', 'EducationLevel', 'Gender', 'HasFamilyHistory',
       'Occupation', 'ATF', 'EAF', 'TKF', 'CMT', 'DEF', 'SMF', 'ERF', 'DAF',
       'HR', 'SW', 'TR', 'DR', 'BR', 'CK', 'CP', 'NS', 'DZ', 'UR', 'UB', 'MD',
       'TG', 'hasSAD', 'SPIN']].values  # Adjust column names as needed
X = ['Age', 'EducationLevel', 'Gender', 'HasFamilyHistory',
       'Occupation', 'ATF', 'EAF', 'TKF', 'CMT', 'DEF', 'SMF', 'ERF', 'DAF',
       'HR', 'SW', 'TR', 'DR', 'BR', 'CK', 'CP', 'NS', 'DZ', 'UR', 'UB', 'MD',
       'TG']
X_train, X_test = train_df[X], test_df[X]
y_train, y_test = train_df['SPIN'], test_df['SPIN']

kmeans_model = KMeans(random_state=0, n_init="auto")
kmeans_model.fit(X_train, y_train)
label = kmeans_model.predict(X_test)
label.shape

# Get cluster assignments for each data point
labels = kmeans_model.labels_

#pca = PCA(n_components=2)
#pca.fit_transform(X_train)
#centers_2d = pca.transform(kmeans_model.cluster_centers_)
#plt.scatter(centers_2d[:, 0], centers_2d[:, 1], c='red', marker='X', s=200, label='Cluster Centers')

#plt.title('K-Means Clustering in 2D via PCA')
#plt.legend()
#plt.show()

# Get the cluster centers
centers = kmeans_model.cluster_centers_

# # Create a scatter plot
plt.scatter(features[:, 0], features[:, 5])

# # Plot the cluster centers as well
plt.scatter(centers[:, 0], centers[:, 5], c='red', marker='X', s=200, label='Cluster Centers')

plt.title('K-means Clustering')
plt.xlabel('Age')
plt.ylabel('SPIN')
plt.legend()
plt.show()
```

Analyzing Columns Using Hierarchical Clustering

Hierarchical clustering is an unsupervised learning technique that measures differences in data to create clusters, them constantly merging clusters to create a dendrogram. Through analyzing this dendrogram, we aim to reach further insights in our data that might not be as visible in our KMeans model. We first defined our linkage matrix to to use euclidean distance and averaging to form our clusters. We then defined our dendrogram from this linkage matrix, plotting our data points in the x axis against our distance in the y axis. Furthermore, we started with a specified threshold of 4, to mark the point in which to stop merging the clusters.

Figure 4: Hierarchical Clustering Implementation

```
linkage_matrix = linkage(df_1, method='average', metric='euclidean')
dendrogram(linkage_matrix, labels=df_1.index, leaf_rotation=90)
plt.xlabel('Data Points')
plt.ylabel('Distance')
plt.title('Hierarchical Clustering Dendrogram')
# Example: Cut the dendrogram at a specified threshold to form clusters
threshold = 4  # Adjust this threshold based on your dendrogram
clusters = fcluster(linkage_matrix, threshold, criterion='distance')
# Filter rows linked to cluster 7
cluster_7_rows = df_1[clusters == 7]
# Print the filtered DataFrame
print(cluster_7_rows)
print(clusters)
plt.show()
```

5. Results and Discussion

A First Look At the Data

At a first glance of our dataset, some columns that stood out as potentially contributing highly include Age, with elders likely being associated with not having SAD and young adults and teens having higher SAD. Additionally, we believed that hasFamilyHistory would have a high positive weightage. Some of the fears and physical conditions that we perceived to have the highest impact were SMF (fear of contact with strangers), ERF (fear of eating in front of another person), and BR (has hard breathing), simply because these conditions are often highly associated with having SAD. Because we wanted to observe relationships withins this data, and used models that allowed us to understand how they had reached their predictions, we used rainforest classifier and logistic regression rather than neural network models.

Rainforest Classifier vs Logistic Regression Accuracy,

We chose to measure our accuracy with 'mean squared error,' to stay consistent with models. While our logistic regression model achieved the lowest error of around 0.246, our random forest classifier achieved a higher error of around 0.298. Overall, these high errors suggests that these models are certainly not ready to be used to predict someone's social anxiety. We can, however, utilize these models to analyze our dataset and the most significant factors contributing to social anxiety. We chose to use Random Forest Classifier and Logistic Regression models, because of the different ways in why these models work to classify data, allowing us to reach more conclusions or support existing conclusions.

Logistic Regression Model Weights

```
Variable: Age_Group_teen, Coefficient: 0.1273
Variable: Age_Group_young, Coefficient: 0.2223
Variable: Age_Group_middle, Coefficient: 0.2898
Variable: Age_Group_elder, Coefficient: -0.6393
Variable: Gender, Coefficient: 0.1374
Variable: EducationLevel, Coefficient: -0.1288
Variable: Occupation, Coefficient: -0.3268
Variable: ATF, Coefficient: 0.2056
Variable: EAF, Coefficient: 0.1575
Variable: ERF, Coefficient: 0.3106
Variable: DAF, Coefficient: 0.4323
Variable: HR, Coefficient: 0.1655
Variable: SW, Coefficient: 0.3323
Variable: DR, Coefficient: -0.3012
Variable: NS, Coefficient: 0.0342
Variable: UR, Coefficient: 0.4088
Variable: MD, Coefficient: 0.0618
Variable: TG, Coefficient: -0.3242
0.24615384615384617
```

In our Logistic Regression model, the variables with the highest positive weights include UR (feeling of being unreal), ERF (fear of getting into a room with other people), SW (has sweating), and DAF (fear of disagreements with strangers). Above all, DAF and UR had the highest weights, being around 0.4, suggesting that the models emphasizes these specific weights more when determining that a person had social anxiety disorder. Furthermore, the feeling of being unreal seems like an extremely obscure variable to have this high weightage, perhaps implying that those who said yes to this question must definitely have SAD. Additionally, because the variables with the highest positive weights were all fears or physical indications may suggest that one's background (age, gender, etc.) may play a less critical role than I had once thought. The variables with the highest negative weights in this model include Age_Group_elder, Occupation, DR (has dry mouth), and TG (has numbness or moaning). These physical conditions having a high negative weight suggests that those who said yes often did not have social anxiety disorder, which may be because they had distinct health conditions can also make one have these physical characteristics. Additionally, the fact that 'Age_Group_elder' has a higher negative weightage suggests that the model associates elder people with a lower likelihood of SAD. Some variables with a low magnitude of weightage include NS (has gastrointestinal discomfort and nausea) and MD (has fear of being crazy). NS can clearly results from alternate conditions, which is likely why it had such a low weightage. MD, on the other hand, is a pretty vague indicator, so respondents may have simply been unsure of how to answer the question.

Rainforest Classifier Model Weights

```
Variable: Age_Group_teen, Importance: 0.0008
Variable: Age_Group_young, Importance: 0.0176
Variable: Age_Group_middle, Importance: 0.0243
Variable: Age_Group_elder, Importance: 0.0031
Variable: Gender, Importance: 0.0230
Variable: HasFamilyHistory, Importance: 0.0409
Variable: EducationLevel, Importance: 0.0606
Variable: ATF, Importance: 0.1621
Variable: TKF, Importance: 0.2082
Variable: CMT, Importance: 0.1032
Variable: SMF, Importance: 0.0870
Variable: ERF, Importance: 0.2339
Variable: SW, Importance: 0.0232
Variable: DR, Importance: 0.0123
0.29846153846153844
0.29846153846153844
```

In our rainforest model classifier, the variables that seemed to play the most importance include ‘EducationLevel’, ‘hasFamilyHistory’, and especially multiple of the fears like ‘ATF’ (fear of being at the center of attention), ‘TKF’ (fear of speaking in public), ‘CMT’ (fear of attending parties), ‘SMF’ (fear of contact with strangers), and ‘ERF’ (fear of getting into a room with other people). Because these fears had a higher weightage than essentially all of the background information (age, gender, etc.), these findings suggest that the Rainforest Classifier model perceives them as more significant indicators. Variables that had significantly lower weights include ‘Age_Group_teen’ and ‘Age_Group_elder’. Although one could argue that these results support the idea that these two age groups have less correlation with social anxiety, it is quite likely that the lack of samples including these two groups contributed to their low weightage. The fact that characteristics like gender, were not considered significant by either models, suggests a less emphasis than we had thought.

Analyzing Data with KMeans Clustering

Figure 5: KMeans Clustering with PCA

Figure 6: ATF vs Age

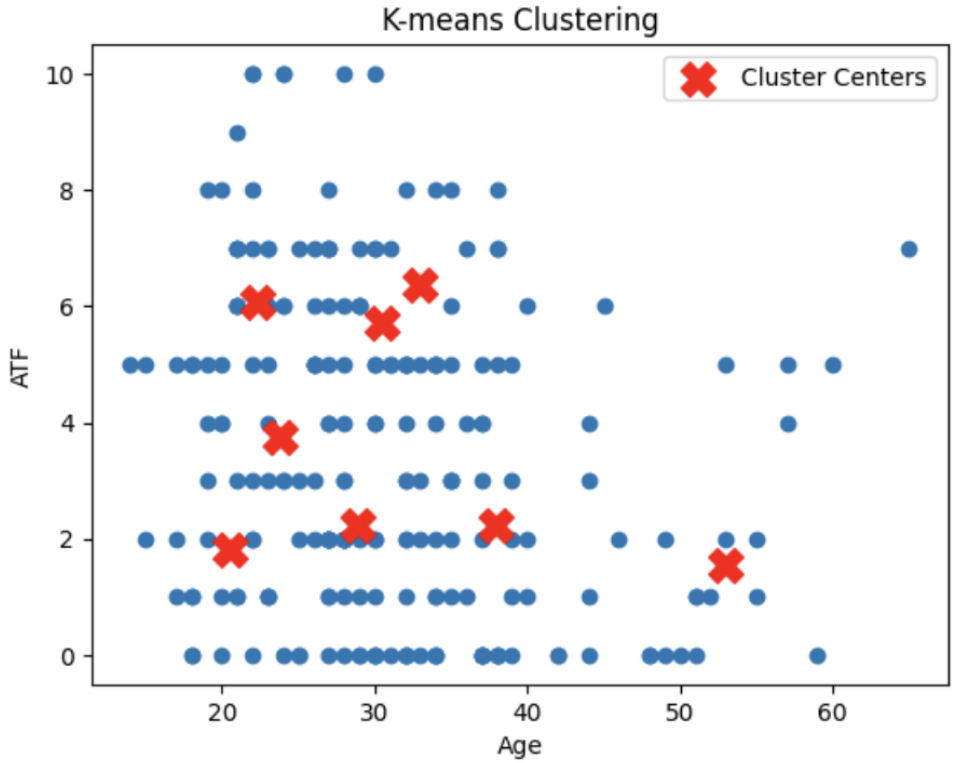


Figure 7: ATF vs Family History

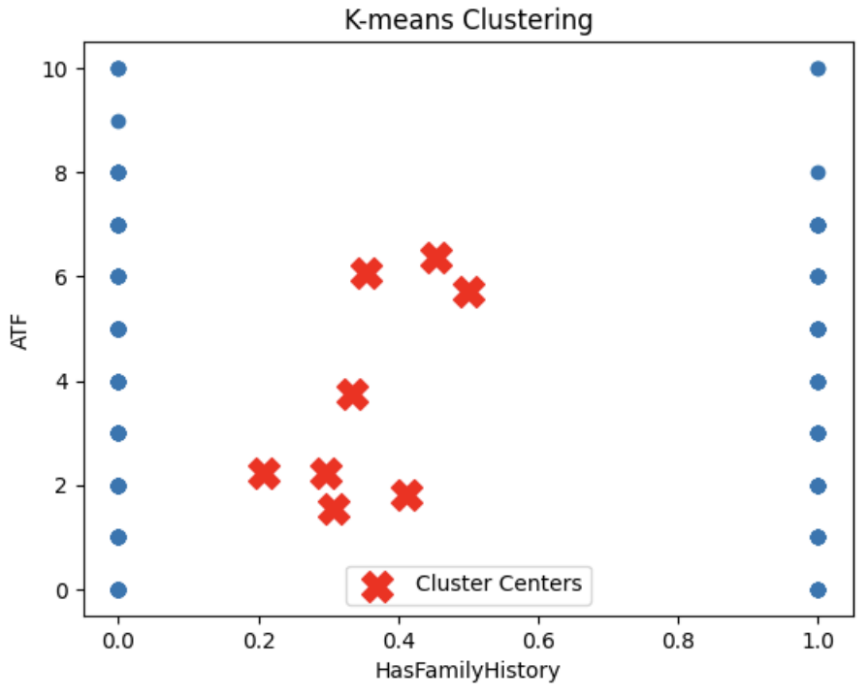


Figure 8: ATF vs Gender

Figure 9: ATF vs Education Level

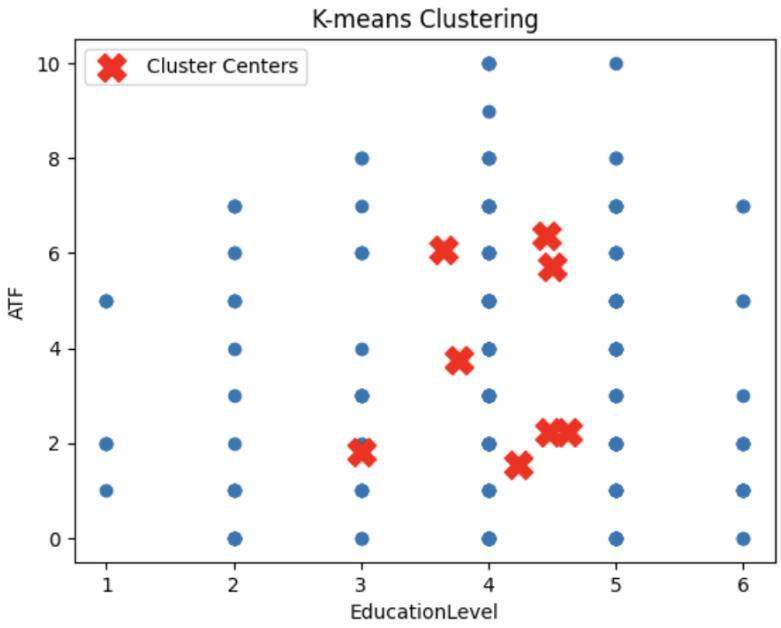
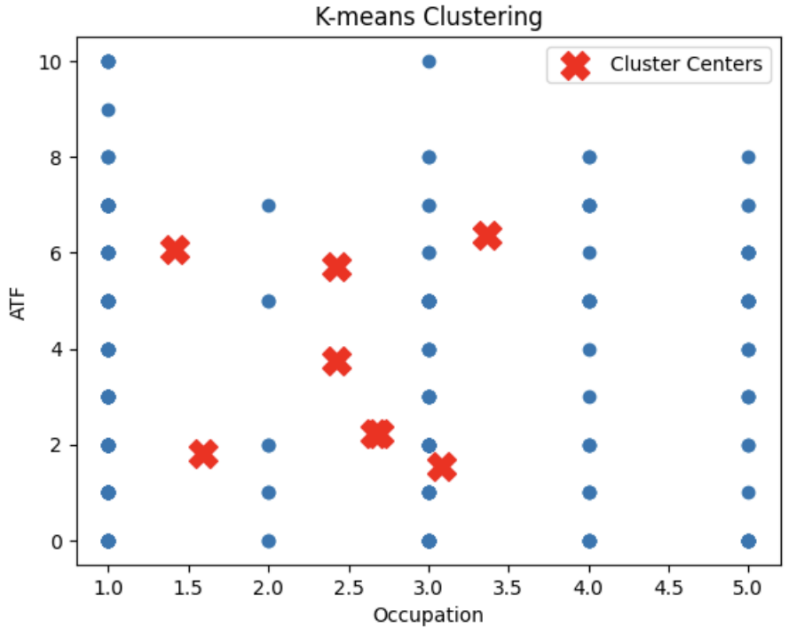


Figure 10: ATF vs Occupation



In our KMeans model, we decided to use PCA to form conclusions about the distribution of our data. Because the centroids are pretty evenly spaced, we can infer that our data is pretty well distributed. Furthermore, our scatter plots seem to support our prior conclusions about one's background being not seemingly too correlated with one's social anxiety. In HasFamilyHistory (1=yes, 0=no) and Gender (1=male, 0=female), all of the centroids seem to be bunched up in the middle, signifying that gender and family history may not be as correlated to social anxiety as we

once thought, also confirming our suspicions from our supervised learning models. In Occupation (1=student, 2=faculty member, 3=employee, 4=self-employment, 5=unemployed) and EducationLevel (1=high school, 2=diploma, 3=undergraduate, 4=bachelor degree, 5=master degree, 6=post-graduate), the centroids also do not indicate any clear relationship between this data and social anxiety. In the age, column, however, we may theorize that an older age may be correlated with lower social anxiety, but it is difficult for us to be sure of this idea without much supporting data.

6. Conclusions

Because there has been a rise in reported rates of social anxiety throughout the years, especially due to modern lifestyles and the rise of technology, the goal of my research paper was to leverage Ml to analyze contributing factors and predict social anxiety. In this research paper, I used both supervised learning (rainforest classifier and logistic regression models) along with unsupervised learning (k means clustering and hierarchical clustering) to evaluate how certain background information, physical conditions, and fears relate to one's SAD. I also leveraged machine learning to create models that predict whether a person has SAD. To train each of my models, I created a list of my columns based on which ones decreased my error (for optimal accuracy), and then created predictions based on the 'hasSAD' column. Overall, however, both of my models did not have high enough accuracy to be usable, likely because of the somewhat low number of data samples. To improve these models accuracy, I would try to get more samples to use with my model. Additionally, I would try to use different supervised and unsupervised learning techniques to see if I can reach a lower mean squared error and make more connections about my data.

Acknowledgments

I would like to thank Inspirit AI, my mentor Udgam Goyal, and my parents for giving me the resources to aid my research and helping me methodically analyzing my data.

References

- Lima Dias da Cruz, Diniz de Carvalho Martins, Rejane Beserra Diniz. "Factors related to the association of social anxiety disorder and alcohol use among adolescents: a systematic review." Jornal de Pediatria. 93(5): 442-451, 2017.
- National Institute of Mental Health (NIMH). "Social Anxiety Disorder." NIMH. https://www.nimh.nih.gov/health/statistics/social-anxiety-disorder (Accessed November 18, 2023).