DAB-304

HealthCare Analytics

Dyslexia: Stuttering severity Identification post speech fluency therapy

GROUP - 9

SAI ABHIRAM GP (779155)

ANIKETH REDDY (793954)

PRAGNA REDDY (792673)

MADHUSUDAN REDDY (785790)

Contents

1 Introduction 2

2 Related Work 4

2.1 UCLASS

2.2 Sep 28K

2.3 ML for Stuttering identification

3 Methods..... 6

3.1 Data Pre-Processing and Cleaning: 6

3.2 Importing Libraries: 6

3.3 EDA: 7

3.4 Evaluation and Success Criteria: 9

3.5 ML Models: 9

4 Results 10

4.1 Decision Tree 10

4.2 KNN 11

4.3 Logistic 12

5 Discussion 13

6 Conclusion 14

7 Contributions 15

8 References..... 16

9 Appendices 16

Tables and Figures

1. Sep 28K, UCLASS, MLSTI Accuracies..... 4

2. MSTI Decision Tree..... 5

3. Heat Map..... 7

4. SFPM vs Severity..... 8

5. MOT vs SeverityAfterTherapy..... 9

6. Decision Tree, KNN, Logistic Accuracies..... 13

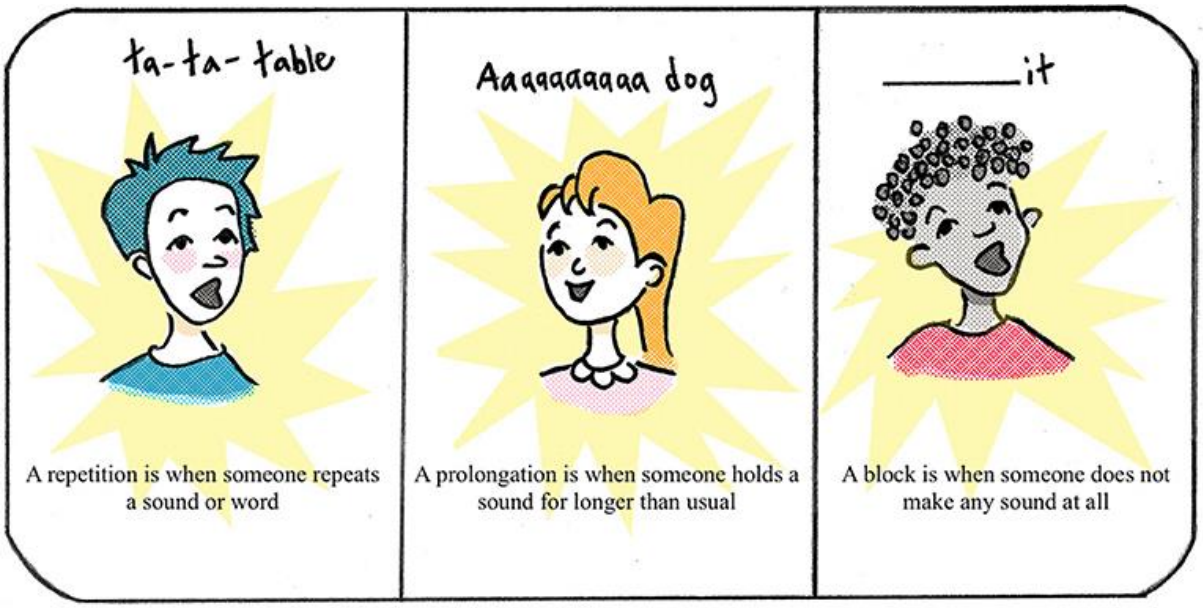
Introduction:

Speech is the most basic and effective forms of expression and communication. According to data from westutter.org, only *1% adults and 5% children* cannot achieve what a normal human being does without any effort i.e., speaking fluently. Stuttering is a disorder directly related to speech where the *fluency flow is interrupted by repetition of words, involuntary pauses and being stuck for a while before attempting to speak*. Dyslexia and stuttering, although the two disorders are unique, the underlying neural disorder process connects the two. Through the lens of dyslexia, we can have a perspective of reality and challenges concerned with stuttering. This how stuttering may look like –

My name is **"Sai A-A-A-A bhiram, I have Sta-sta-sta-tuttering disorder" –** sign of repetitions.

"_____pro-pro-ject is based on dyslexia" – sign of being stuck for a while before attempting to speak

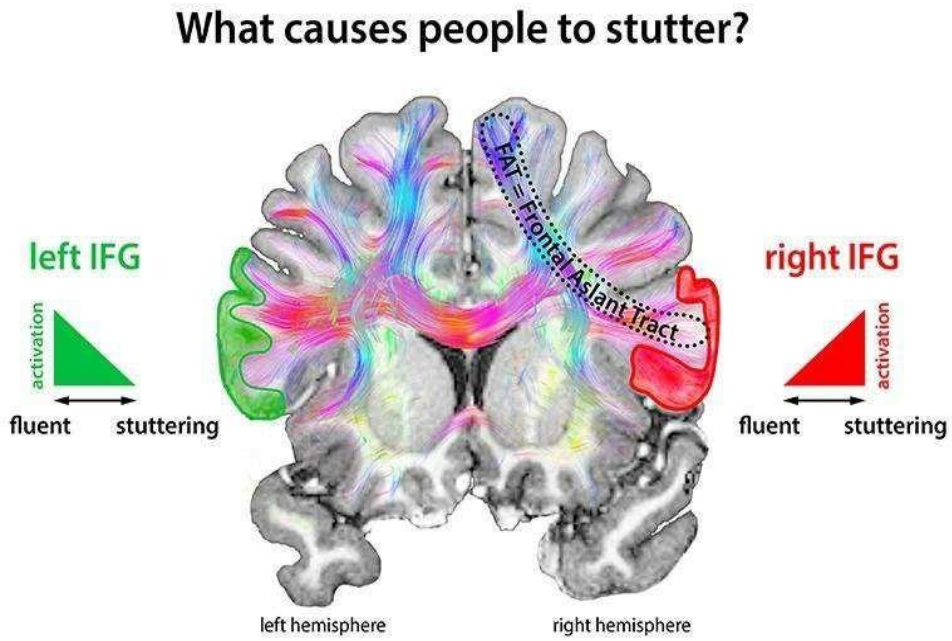
Types:



What neural process causes stuttering?

Nicole Neef, neuroscientist at MPI CBS – Max Planck Institute for Human Cognitive & Brain Sciences in Leipzig and University of Medical Centre Gottingen have gained crucial insights in regard of stuttering. According to their insights,

- The *hyperactivity in the regions of right hemisphere of brain is the central cause for stuttering.*
- Parts of *right inferior frontal gyrus (IFG)* are active when we stop actions such as hand or speech movements. If this region is overactive, it hinders other brain areas that are responsible for speech.



- In the above figure, the *right inferior frontal gyrus (IFG)* restricts the flow of speech while the *left IFG* supports it but in people who stutter, the both areas are conversely activated. Right IFG gets overactive and leads to tightened connections with *frontal aslant tract (FAT)*, which further is sign of strengthened movement inhibition causing interruptions in flow of speech and might inhibit activity in left IFG. The *left IFG processes the planning of speech movements* while *left motor cortex controls speech movements*. "If these two processes are sporadically inhibited, stuttering results in the affected person."

2. Related Work:

2.1 UCLASS:

University College London Stuttering Archive (UCLASS) public data set (albeit very small) [73] is the most widely used in the stuttering research community. UCLASS is offered in his two editions from UCL's Department of Psychology and Language Sciences. It contains monologues, dialogues and readings, for a total of 457 voice recordings. Some of these records contain transcriptions, including orthographic, phonetic, and canonical transcriptions. Of these, the orthographic is ideal for stuttering lettering. UCLASS3 version 1 contains 139 monologue samples from 81 PWS aged 5 to 47 years.

2.2 SEP- 28K:

The public dataset on stuttering is too small to make a sufficiently generalizable ASIS. To address this, Colin et al. [34] Recently curated a public version of a stuttering event Podcast (SEP-28k) dataset. This dataset contains a total of 28,177 samples. or
The SEP-28k dataset is the first published annotated dataset with deranged labels.

2.3 Machine Learning for Stuttering Identification by Shakeel Sheik :

Stuttering identity is an interdisciplinary studies trouble wherein a myriad range of studies works (in phrases of acoustic characteristic extraction and type. This work focuses on detection of stuttering type.

Method	Accuracy					Tot. Acc.	MCC.
	R	P	B	F	In		
Dataset: UCLASS							
Resnet+BiLSTM [29]	20.39	**23.17**	**53.33**	55.00	NA	46.10	0.20
StutterNet *(Baseline)*	**27.88**	17.13	42.43	66.63	NA	49.26	0.21
StutterNet **(Optimized)**	23.98	12.96	47.14	**69.69**	NA	**50.79**	**0.23**
Dataset: SEP-28k							
StutterNet [30]	21.99	27.78	1.98	88.18	49.99	60.33	NA
MB *StutterNet* [31]	28.70	37.89	9.58	74.43	57.65	57.04	NA
Sheikh *et al* [35]	**46.79**	**40.79**	**23.86**	84.32	**69.54**	**68.35**	NA

Table 1 SEP-28K, UCLASS, MLSTI ACCURACIES

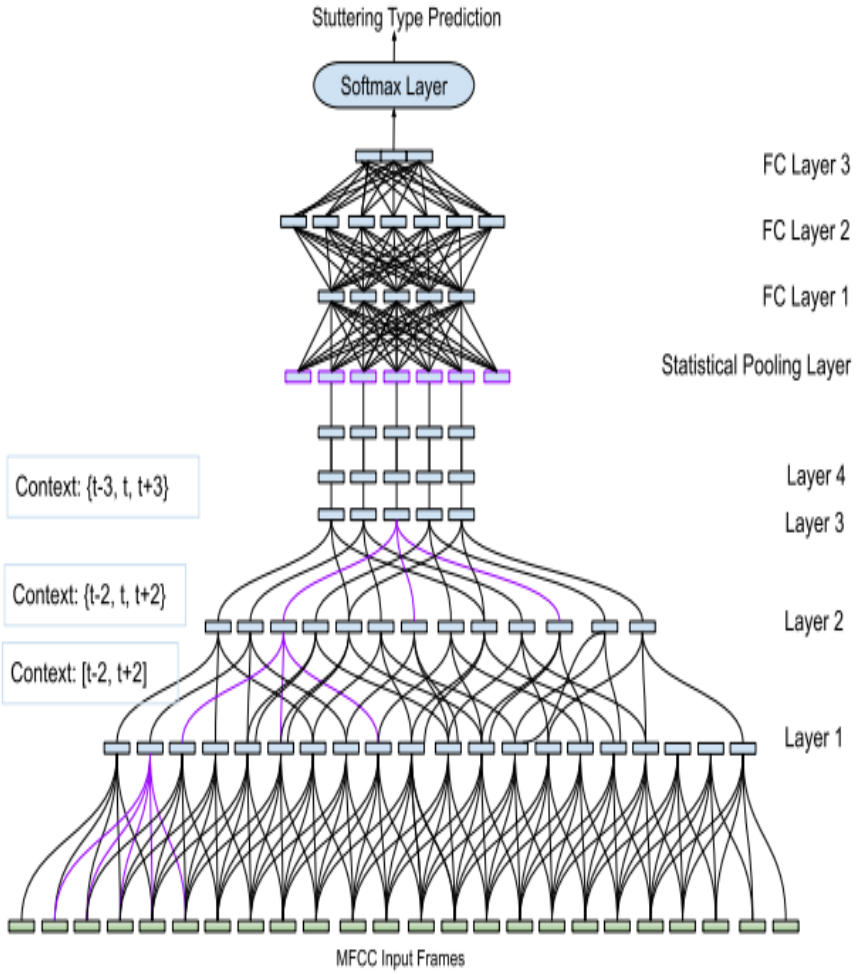


FIG 1 MLSTI DECISION TREE

- The above table and figure represents the Machine Learning Models utilized by different group, the accuracies and scores achieved.

- The above table and decision tree fig. can be used as a reference to check the accuracy of the model. Although I used some different algorithms than the one used above. Based on this, we can compare the accuracy, recall scores, and confusion matrices collected for our machine learning models.

3. Methods

This segment affords short data approximately the manner or technique we intend to apply to our dataset, exploratory information analysis, pre-processing techniques, machine learning models, and score metrics.

3.1 Data pre-processing and cleaning

Data preprocessing is a way that's used to convert the statistics right into a usable and comprehensible version. The uncooked statistics is continually incomplete and can motive mistakes later, that's why it's far continually an amazing preprocess the statistics. Data is wiped clean in order that handiest essential facts is retained and the identical may be fed into the system gaining knowledge of fashions for in addition processing.

We commenced making ready the statistics via way of means of checking the wide variety of null values or lacking values for every column in our dataset. Based at the output from data cleaning using Excel, we discovered that there has been no null values in any of the column.

Based at the preprocessed dataset, there has been no irregularities determined and the entirety become searching good. Which manner that our dataset is ready for in addition processing while fed into the Machine Learning Models.

Features Information

Patient ID: Unique value assigned to each patient

Age: of the patient

Gender: Male/Female/Prefer not to say

Profession

Anxiety Level: On a scale of 1-3 (1- Low, 2-Medium, 3-High)

SFPM: Stuttering Frequency Per Minute

Severity: On a scale of 1-3 (0-nNill, 1- Low, 2-Medium, 3-High)

SSI-3: Stuttering Severity Instrument (stuttering measurement technique)

MOT: Months on Therapy (Maximum of 9 months)

Severity After Therapy – (0 – Recovered, 1 – Low)

FPMAT: Frequency/Minute After Therapy

3.2 Importing Libraries and Dataset

We imported the specified libraries and applications that is vital to run a selected phase or all of the sections of the code withinside the python notebook (ipynb file). Then we had imported our dataset and saved it in a statistics body for similarly processing.

3.3 EDA

We performed data exploration by looking at the contents of the data frame and understanding the variables stored in the data frame. This will help yus understand what kind of columns you are dealing with and which specific columns are useful for prediction. Understand data characteristics, data types, non-zero counts, sums, averages, standard deviations, minimum, maximum, and specific quartiles using basic functions such as Head, Tail, Description. We also analyzed the distribution of numerical and categorical features in the dataset to gain a visual understanding of the features.

- We then created a correlation matrix heat map to summarize the data.

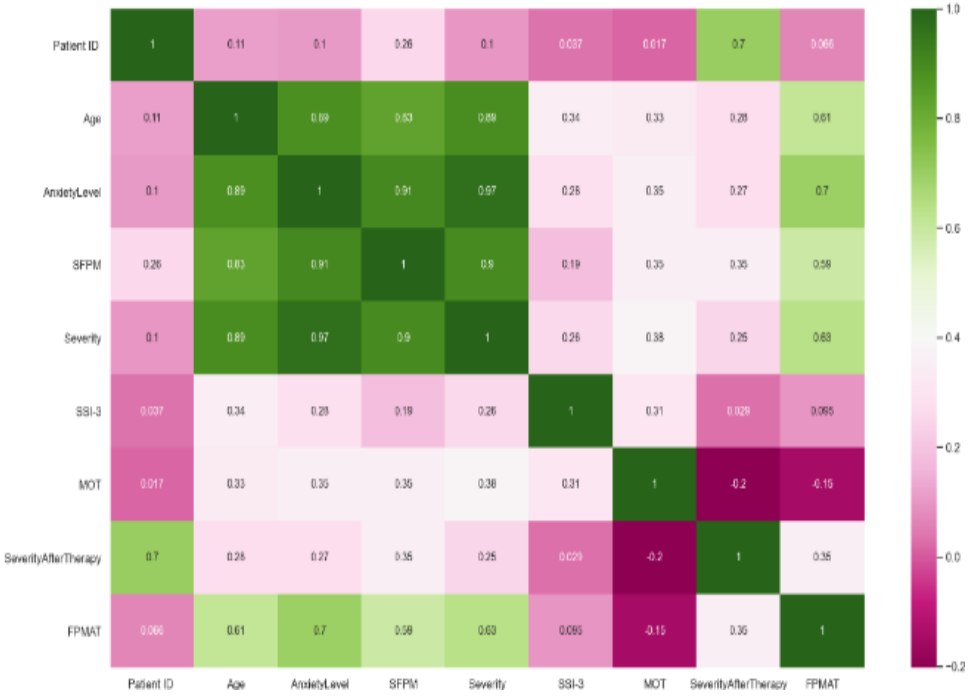


FIG 2 HEATMAP

SFPM vs Severity

From the graph it is clear that, Severity increases with increase in SFPM Score

- For SFPM score between 1 – 3, severity is Low
- For SFPM scores between 4-6, severity score is Medium
- For SFPM score above 6, severity is High

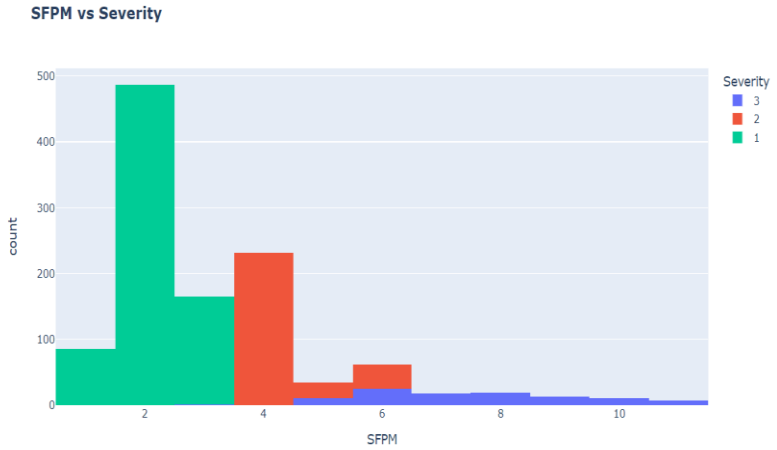


FIG 3 SFPM vs Severity

- **Months of Therapy vs Severity After Therapy**

It is observable that, as a patient undergoes longer months of therapy, severity reduces.

- For MOT between 0 – 4, decrease in severity is less
- For MOT above 6, we can see the rapid fall of severity

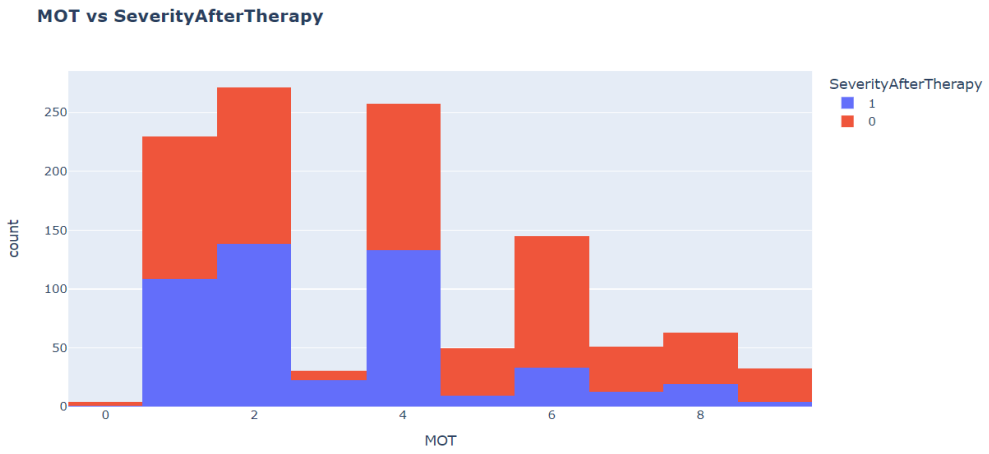


FIG 4 MOT vs SeverityAfterTherapy

Similarly, the patterns for other variables can be found in the **Appendices** section.

3.4 Evaluation and success criteria

Metrics are used to define model performance. Using these metrics is important. These metrics help you better understand how each model is performing at its peak and what you can improve later to improve your score. Here are the metrics we used:

o Data gathering through group research and online sources and verifying the gathered data.

o Data Cleaning and transformation using **MS Excel** and **SQL**. Performing **Exploratory Data Analysis (EDA)** on the developed data using Python.

o Analysing the patterns and corelating factors like **Patient ID, Gender, Age, Profession, Stutter Type, Stutter Frequency, Number of Repetitions, Severity After Therapy** using Python.

o Performing **classification tests** and building a model using packages like **sklearn** and train the same to predict the severity before and after speech fluency therapy

o Performing visual analysis using **Tableau**.

By performing and working on all the above processes and by measuring the result to be above 70% for severity reduction after therapy from the model, we will consider the project to be successful. As stuttering is a disorder without any cure, if we are able to correlate the factors using analytical process and prove that speech fluency therapy with respect to individuals personal and mental conditions improve the stuttering condition, then it is a straight communication success.

3.5 Machine Learning Models

Machine learning is the study of algorithms that can be improved over time. This is an area of artificial intelligence used to process sample data, called training data, to train models and create models that can make better and more accurate predictions on test data in the future. is.

- Decision Tree
- KNN
- Logistic Classification

4. Results

4.1 Decision Tree

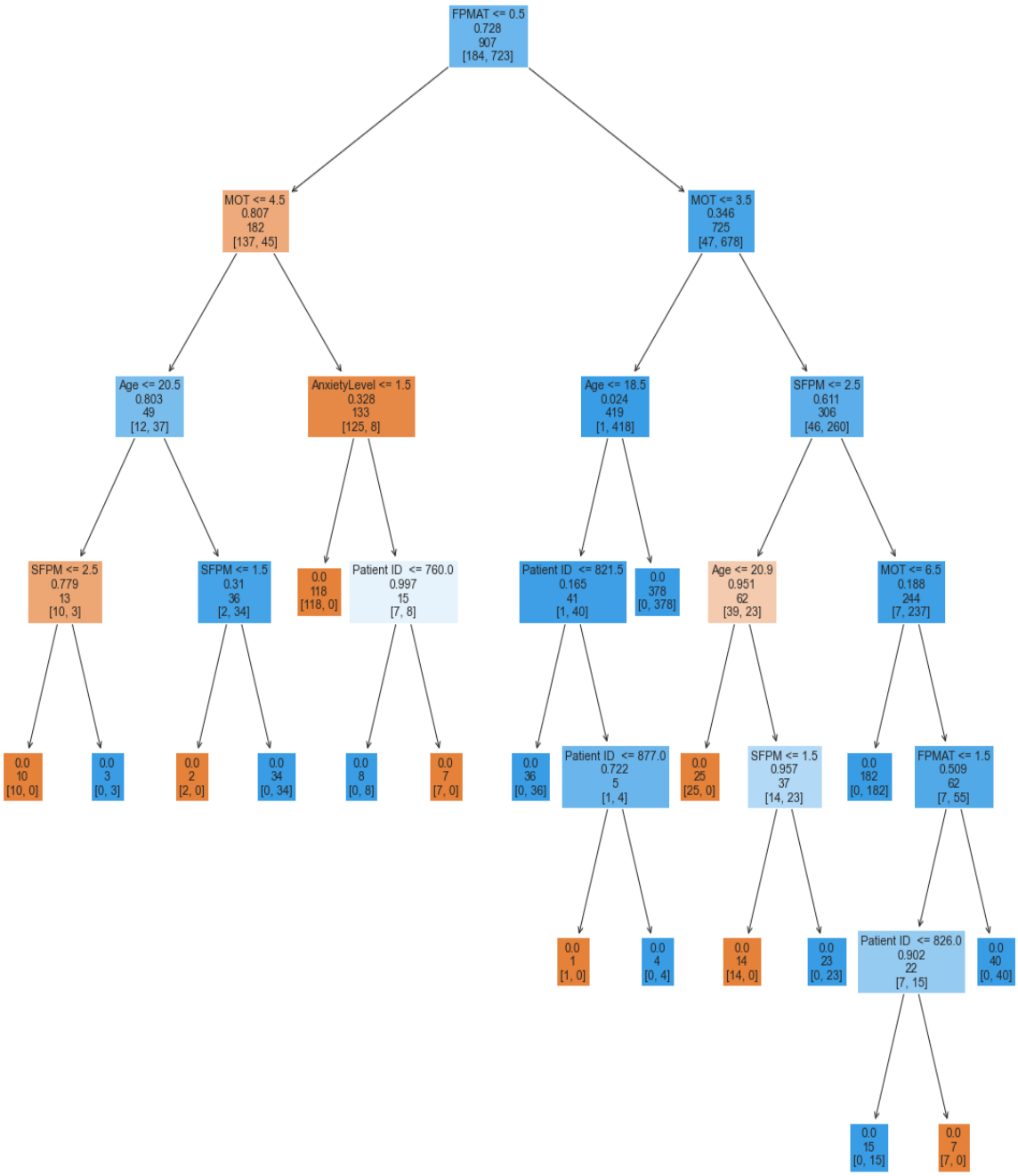
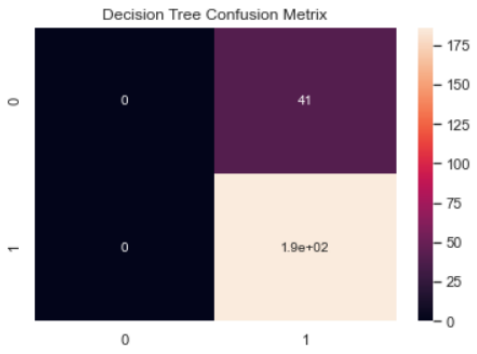


FIG 5 Decision Tree

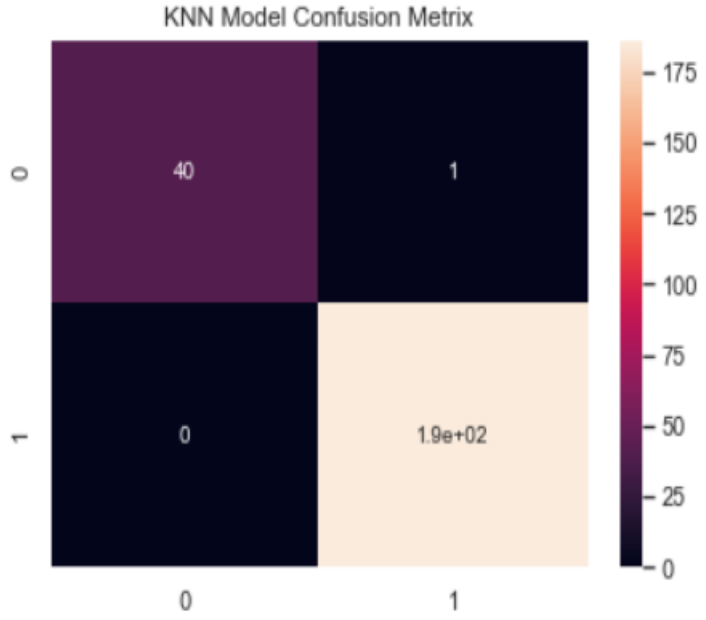



```
In [675]: print(classification_report(y_test, y_pred_d))
```

	precision	recall	f1-score	support
0	1.00	0.98	0.99	41
1	0.99	1.00	1.00	186
accuracy			1.00	227
macro avg	1.00	0.99	0.99	227
weighted avg	1.00	1.00	1.00	227

The accuracy with 0.99 and f1 score of 1.00 states that the the model is overfitting

4.2 **KNN**



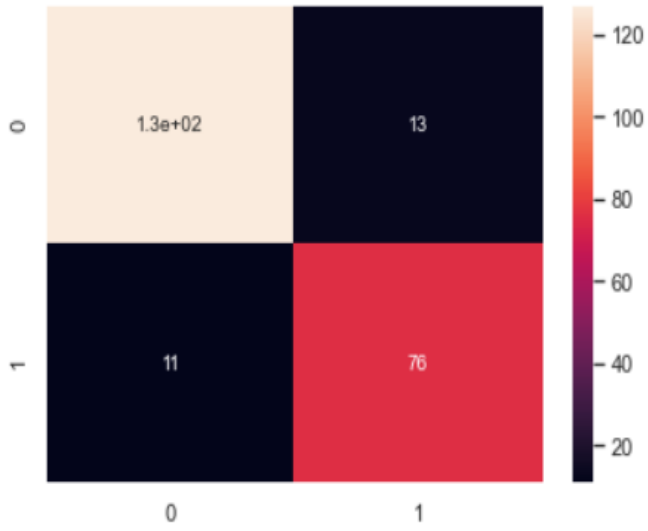

```
accuracy = accuracy_score(y_test, y_pred)
print(accuracy)
```

```
0.8193832599118943
```

-

On evaluating the model using **KNN** classification, we gained an accuracy of **81.9%**

4.3 Logistic classification




```
In [644]: cr = classification_report(y_test, y_pred)
          print(cr)
```

```
              precision    recall  f1-score   support

           0       0.92      0.91      0.91       140
           1       0.85      0.87      0.86        87

    accuracy                           0.89       227
   macro avg       0.89      0.89      0.89       227
weighted avg       0.89      0.89      0.89       227
```

On evaluating the model using **Logistic** classification, we gained an accuracy of **89%**

5. Discussion

Data Gathering and Verifying

- Since stuttering is a rare condition, we had to particularly extract data from two or more resources and further verify them.
- We had created new variables from scratch and fill the rows with verified data.

Converting SFPM audio file to text

- Converting audio files to text based to detect the stutter frequency per minute

Model Evaluation

- Decision tree classification – Our model was overfitting and hence had to choose other techniques like KNN and Logistic classification
- We gained better results with KNN and Logistic classification as shown above in the results section.

Model	Accuracy
Decision Tree	99% (overfit)
KNN	81%
Logistic	89%

TABLE 2 Decision Tree, KNN, Logistic Accuracies

6. Conclusion

We as a group strived to work on unique real life health scenarios that any of our group member is directly facing. Since one of our group members is an active stutterer, we chose to explore this area. One must be able to speak efficiently and read efficiently to be successful in most careers and to live life completely. It can be daunting to grasp steps leading to success but fall repeatedly. Stutter might always feel like “I’m living in a world that is not meant for me because of who I am.” Stutterers always suffer with problems like thoughts like depression, anxiety, stage fear, stress, loneliness, acting like introverts and others.

Since only 1% of adults and 5% of children are affected with this disorder and is a rare case, there is not much research done on this field of healthcare leading to data restrictions. Considering all the challenges, we found it very motivating and interesting to work and explore this area of healthcare.

Stuttering identification is an interesting interdisciplinary research problem involving pathology and psychology making it hard and complicated to detect. By working on projects like this we will learn with (mainly Pathologists) and from a person stuttering which can be a bridge for bringing in ideas, therapies and techniques to reduce stuttering and create awareness

Based on the outcome of the implementation of various Machine Learning models namely Decision Tree, KNN, Logistic Classification and and from their individual accuracy scores, it is very evident that the two best performing algorithms turn out to be KNN and Logistic classification.

Hence Fluency Therapy can inspire dreams and when a dream has truly been embraced, it fulfils the heart of man or a kid. Sometimes it takes something or someone to help light a spark and that spark in this case might be fluency therapy by a pathologist. Fluency therapies include speech exercises, meditation and yoga and other techniques which reduce stuttering by lowering disfluency affecting factors like stress, depression, fear, and others.

7. Contributions

Name	Contribution
Sai Abhiram GP, 779155	- Research and study on stuttering disorder - Data Collection, developing data and Verifying data. - EDA - Correlation between factors, classification test, Logistic Regression, building a decision tree model using scikit learn, Tableau Visualisation.
Aniketh Reddy, 793954	- Study on brain and neuro-disorders related to stuttering - Developing new numerical feature related to frequency of stuttering - Exploratory Data Analysis and statistics - Visualizations with respect to factors.
Pragna Reddy, 792673	- Data exploration and gathering using resources like UCLASS and SEP-28K - Study on the cause of stuttering. - Developing new categorical and numerical features like Profession, Number of Repetitions. - Data cleaning using MS Excel. - Supporting in EDA and classification test. - Tableau visual analysis.
Madhusudan Reddy, 785790	- Creating new categorical feature based on severity i.e., Low, Moderate, High - Data Transformation using Microsoft SQL server. - Supporting in classification tests and model building using python

8. References

1. https://vkc.vumc.org/childhoodstuttering/forscientists_datasets.html - Partial data collection.
2. Understanding Stuttering through Dyslexia: Finding Hope and Inspiring Dreams – International Stuttering Awareness Day (isad.live) – study on importance of speech therapy.
3. Machine Learning for Stuttering Identification: Review, Challenges & Future Directions (archives-ouvertes.fr) - Further study on types of stuttering, severity and factors.
4. Scientists Image the Brains of Stutterers to Find Cause | Technology Networks – learning the cause of stuttering.
5. (https://stammer.in) – TISA, Categorical features collection.
6. Release 1 (ucl.ac.uk) – Partial numerical features collection.
7. Machine Learning for Stuttering Identification: Review, Challenges & Future Directions (archives-ouvertes.fr) – Related work.

9. Appendices

https://colab.research.google.com/drive/1EtwncwSNKRUIvFizo84RxthOjTUtTqze?usp=sharing

Contains all the code for our task inclusive of importing, pre-processing, exploratory information analysis, algorithms and metrics.