PES
UNIVERSITY

Human Voice Analysis to Determine Age and Gender

Audio Technology Project

2 Credits

Team Composition

Bhushan K M	PES1UG19EC359	
Prem Kumar	PES1UG19EC215	
Adityaram	PES1UG19EC018	
K Abheeshta Kumar Jain	PES1UG19EC118	

Guide : Prof. Ashwini

Contents

- Problem Statement.
- Block Diagram.
- Literature Review.
- Results.
- Summary.
- References.
- Deliverables.
- Project timeline Gantt Chart (Aug-Dec 2022).
- Q & A.

Problem Statement

Age and Gender Detection using Audio data and various ML model.

Literature Review

S.No	Title:One Source to Detect them All: Gender, Age, and Emotion Detection from Voice
1.	Key Takeaway: ● One model to detect **Gender, age, and emotion** concurrently ● Audio files converted to **20 statistical features** ● Models used: 1) Random Forest, 2) CatBoost, 3)Gradient Boosting, 4)K-nearest neighbours(KNN), 5) XGBoost, 6) AdaBoost, 7) Decision Tree, 8) Artificial neural networks (ANN), 9)Naive Bayes, and 10) Support vector machine (SVM). **Total ten models**. ● Datasets used: Mozilla audio dataset (Common voice corpus), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) ● **Methodology**: All the audio files converted into WAV format required for **frequency spectrum analysis**. 20 features were extracted from each dataset for every target (age, gender, and emotion). 20 features include: mean frequency (meanfreq), standard deviation of frequency (sd), median frequency (median), first quartile (Q25), third quartile (Q75) etc. Models were trained using these. ● Confusion matrix was generated and performance metrics were calculated. ● Best models for concurrent age, gender, and emotion detection: **Cat Boost, Random Forest, and XGBoost**, respectively.

Literature Review

S.No	Title:Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca
2.	**Key Takeaway** ● **Two main objectives:** 1) To determine the influence of mother tongue versus lingua-franca languages on voice recognition(voice verification) 2) To evaluate how training sets with: (**a**) English language set of utterances (**b**) a native language set of utterances (**c**) and a mixture of native language set of utterances influence the **classification of gender from voice**. ● **Dataset**: A total of 3980 voice utterances from 520 Nigerians.Two Sessions(Between June and November 2018). 326 male and 194 female subjects with 1964 and 2016 voice utterances, respectively. Voice has 2 sets: English and native utterances datasets. ● **Methodology:** Mel frequency discrete wavelet cepstral coefficients were used (**MFDWC**). A CNN model with **VGG-16** architecture was used to extract the **MEL coefficients**(224).Gender was estimated from voice notes of samples of (1) English utterances made by all subjects (2) Native language utterances made by all subjects (3) English language utterances made by the dominant native tribe (4) Native language utterances made by the dominant native tribe**.4 Vgg-16 models** were used for these four different sample pools. ● **Overall accuracies** for the datasets on 4 models: **83.33%, 84.85%, 82.32%** and **85.1%** respectively. ● **Conclusion :** Result of gender classification from voice is optimal when training sets have a mother tongue language rather than a second language.

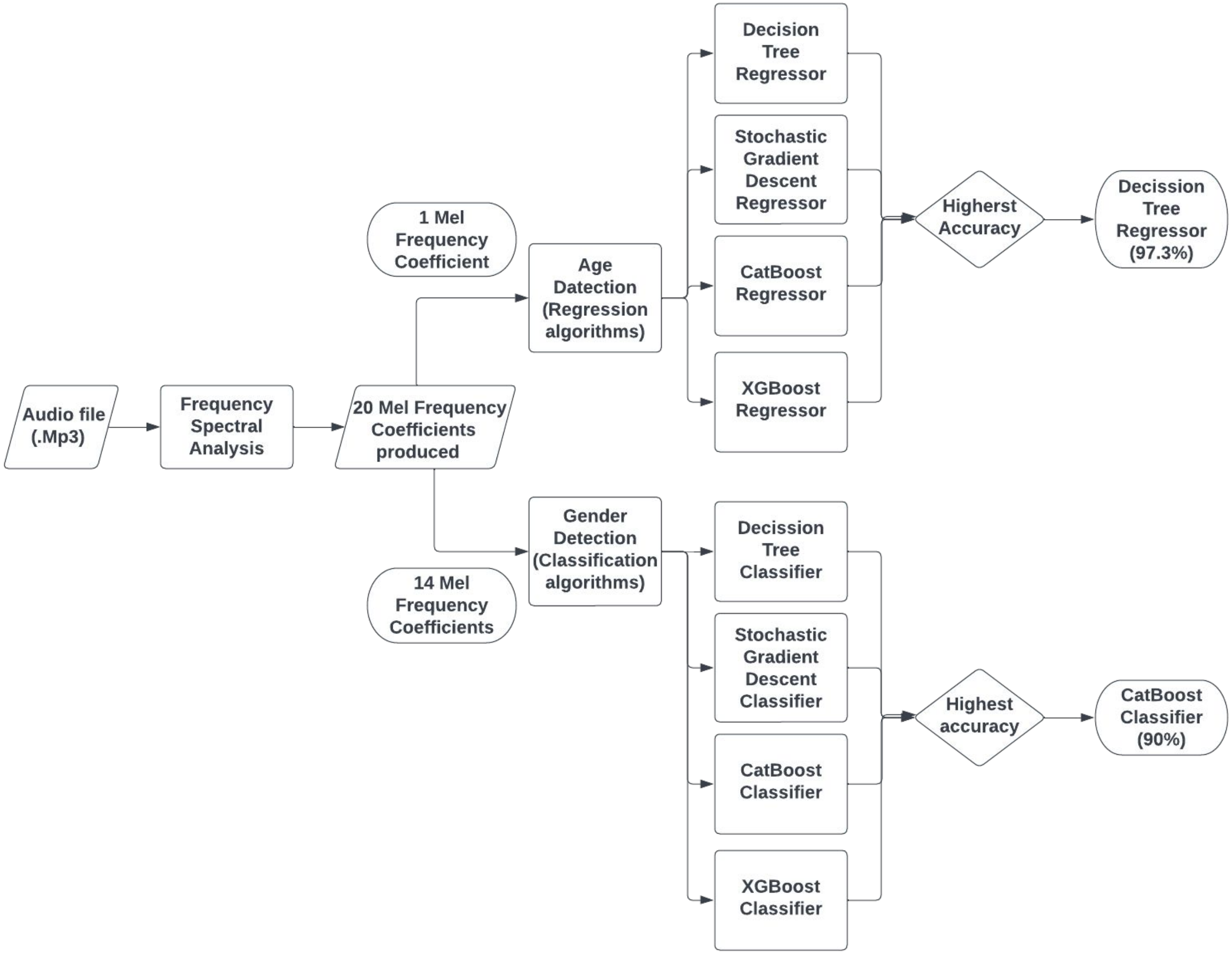
Literature Review

S.No	Title:Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks
3.	Key Takeaway ● They have applied various **Deep Neural Network-based embedded architectures** such as **x-vector and d-vector** to age estimation and gender classification tasks. Furthermore, they have applied a **transfer learning-based training scheme** with pre-training the embedder network for a speaker recognition task using the **Vox-Celeb1** dataset and then fine-tuning it for the joint age estimation and gender classification task ● In terms of age estimation, the proposed system with **two-staged transfer learning scheme and a QuartzNet embedder** achieved new state-of-the-art result on the **TIMIT dataset**, with a MAE of 5.12 years for male, 5.29 years for female speakers, and RMSE of 7.24 and 8.12 years for male and female speakers respectively ● In terms of gender classification, the d-vector-based system achieved a robustly high performance with accuracies varying from **96.8% to 99.6%**depending on the training and testing datasets. The highest result was achieved when the Common Voice dataset was used for training, the algorithm was further fine-tuned onTIMIT dataset, what enabled the classification accuracy at the level of **99.6%** for gender recognition

Literature Review

S.No	Title: A new approach for age estimation system based on speech signal
4.	**Objective:** To get higher efficiency from the voice features by using different feature extractors. **Dataset:**Mozilla open source dataset was used. (offered a variety of accents and includes speakers from different geographical regions.) majority of the speakers are in the agr group of 20s. **Methodology:** 1) Invoice detection: The silent and unvoiced audio parts are removed. Zero-Crossing Rate and Root mean square energy is used to perform these operation. 2) Feature extraction: instead of using only MFCC, Rasta-PLP and Chroma(features belonging to Pitch) was also used as a feature extraction method. Total of 450 features were extracted from the audio signal using these methods. 3) Dimension reduction: Random Forest Regression and PCA are used to reduce the dimensions of the dataset. With RFR, top 150 features were extracted and with PCA feature set of 150 was reduced to 100 by choosing 0.95 variance. 4)Multi layer Perceptron:Totally 14 layers are considered and learning rate is 0.0001. **Results:** The obtained feature set was classified with MLP algorithm and an accuracy rate of 94.34% was obtained.

Block Diagram


Audio file (.Mp3)
Frequency Spectral Analysis
20 Mel Frequency Coefficients produced
1 Mel Frequency Coefficient
Age Detection (Regression algorithms)
Decision Tree Regressor
Stochastic Gradient Descent Regressor
CatBoost Regressor
XGBoost Regressor
Higherst Accuracy
Decission Tree Regressor (97.3%)
14 Mel Frequency Coefficients
Gender Detection (Classification algorithms)
Decission Tree Classifier
Stochastic Gradient Descent Classifier
CatBoost Classifier
XGBoost Classifier
Highest accuracy
CatBoost Classifier (90%)

Methodology

- Procuring the Dataset
- Frequency Spectral Analysis
- Applying deep/machine learning models
- Output Prediction

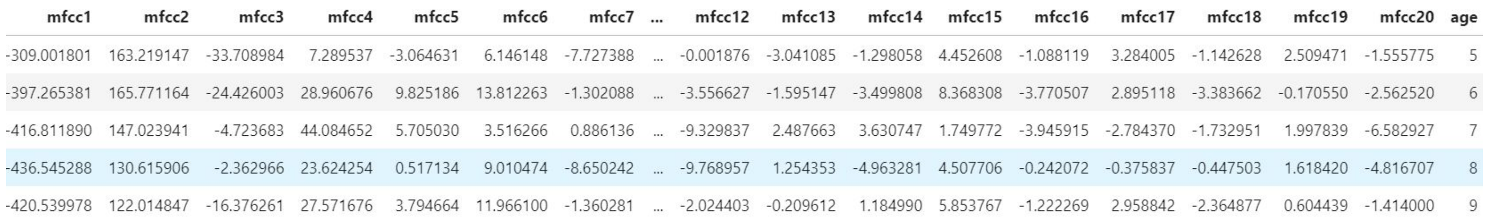
Dataset

The dataset used in this project is self-built. We collected audio samples from the web by extracting them from various online video footage. The samples comprise the voices of people, men and women, belonging to the age range of 5-75 years

MFCCs(Mel-frequency cepstral coefficients)

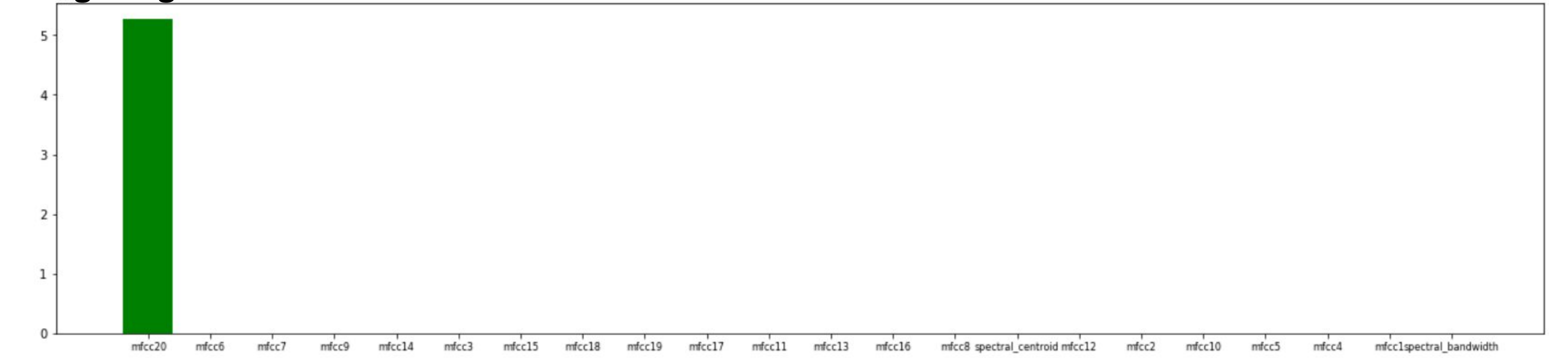
- Set the sampling frequency for the audio and divide it into equal sized frames(windowing).
- Take the DFT of the sampled signal.
- Compute the mel-spaced filterbank. Mel frequency scale maps the actual frequency to the frequency which we(humans) can perceive.
- $mel(f) = 1127 ln(1 + f/700)$
- Log transform for approximation.
- Convert the signal back to time domain using IDFT
- Log transform of the magnitude of the signal is called cepstrum
- The coefficients of the IDFT(first 12), energy of the signal(13th), and the first and second order derivatives of the signal together constitute the mel frequency cepstral coefficients(Total 39).
- We have used 21 MFCC coefficients in our project

Feature Extraction and Feature Importance

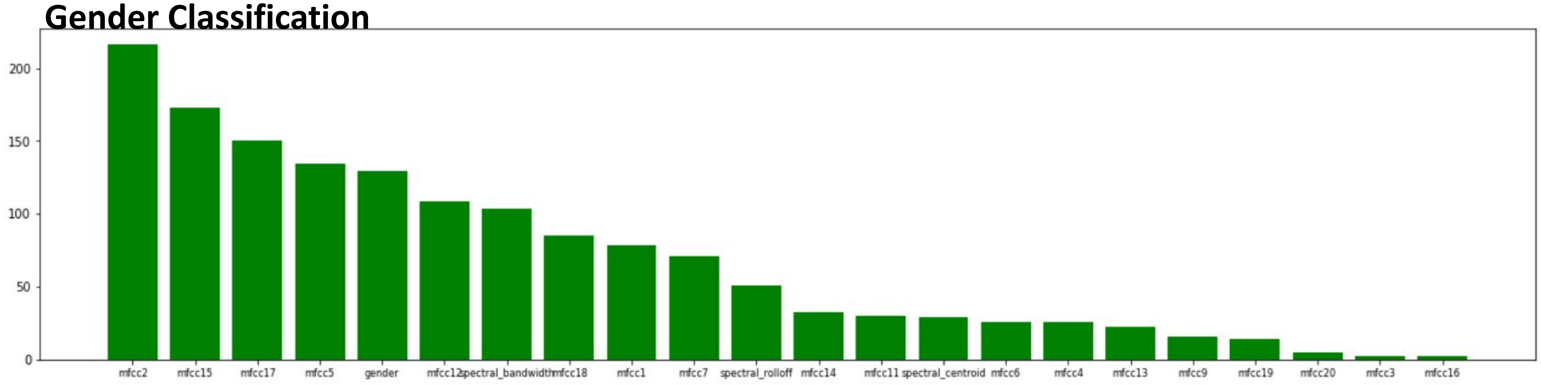


mfcc1	mfcc2	mfcc3	mfcc4	mfcc5	mfcc6	mfcc7	...	mfcc12	mfcc13	mfcc14	mfcc15	mfcc16	mfcc17	mfcc18	mfcc19	mfcc20	age
-309.001801	163.219147	-33.708984	7.289537	-3.064631	6.146148	-7.727388	...	-0.001876	-3.041085	-1.298058	4.452608	-1.088119	3.284005	-1.142628	2.509471	-1.555775	5
-397.265381	165.771164	-24.426003	28.960676	9.825186	13.812263	-1.302088	...	-3.556627	-1.595147	-3.499808	8.368308	-3.770507	2.895118	-3.383662	-0.170550	-2.562520	6
-416.811890	147.023941	-4.723683	44.084652	5.705030	3.516266	0.886136	...	-9.329837	2.487663	3.630747	1.749772	-3.945915	-2.784370	-1.732951	1.997839	-6.582927	7
-436.545288	130.615906	-2.362966	23.624254	0.517134	9.010474	-8.650242	...	-9.768957	1.254353	-4.963281	4.507706	-0.242072	-0.375837	-0.447503	1.618420	-4.816707	8
-420.539978	122.014847	-16.376261	27.571676	3.794664	11.966100	-1.360281	...	-2.024403	-0.209612	1.184990	5.853767	-1.222269	2.958842	-2.364877	0.604439	-1.414000	9

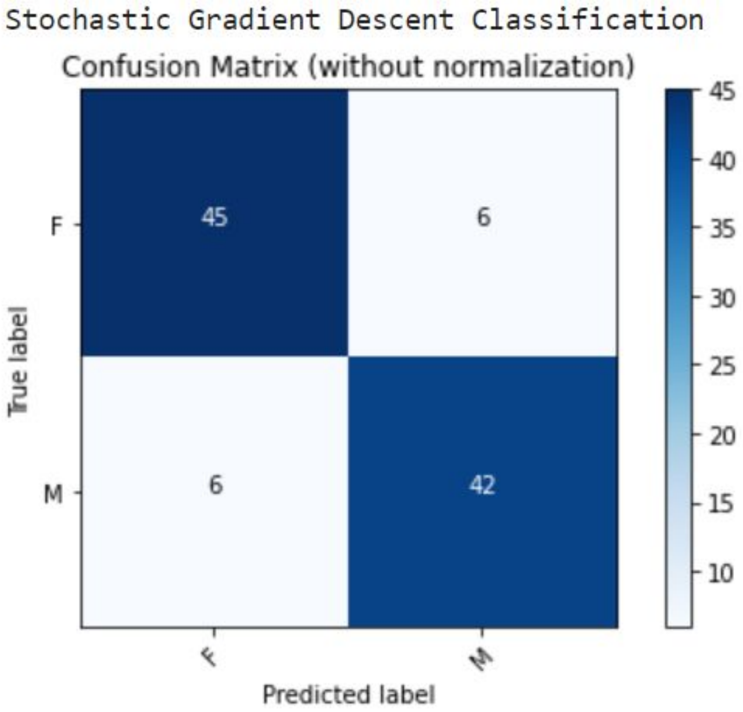
Age Regression



Gender Classification

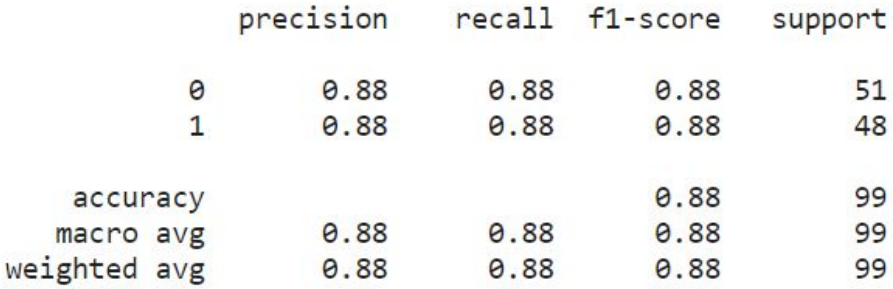


Results



Stochastic Gradient Descent Classification

Confusion Matrix (without normalization)

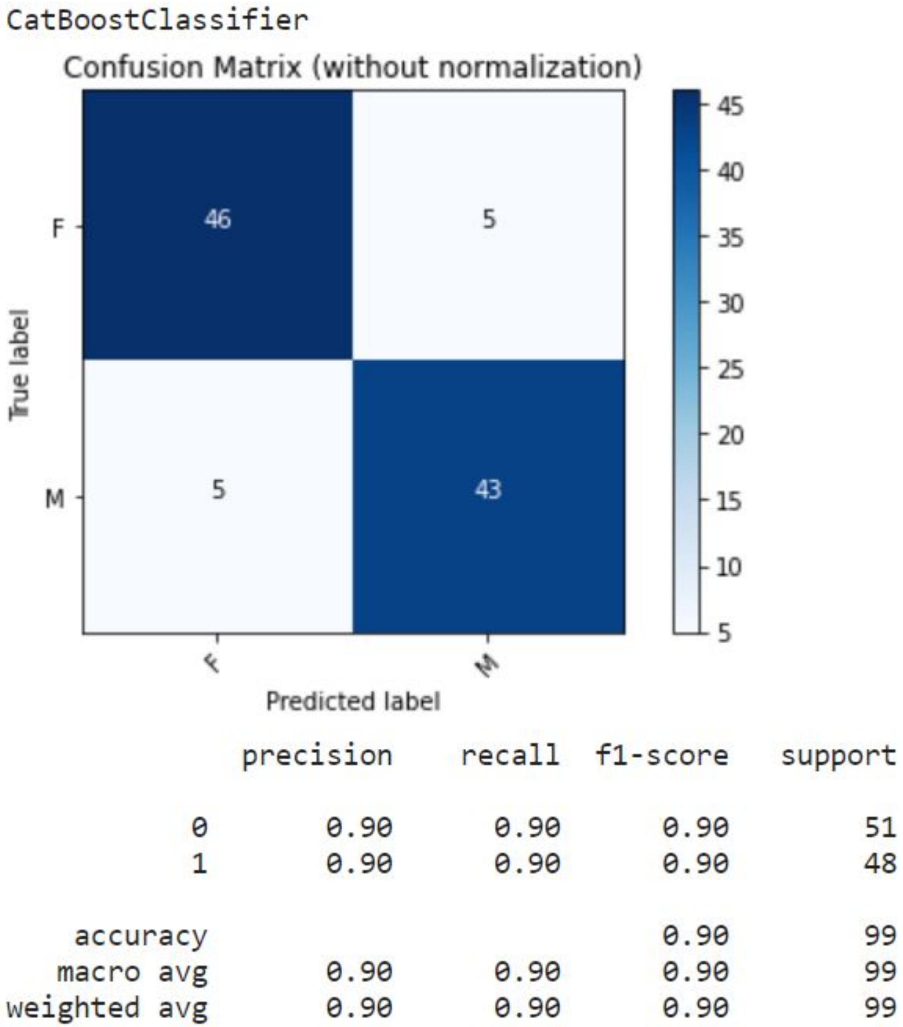


```
              precision    recall  f1-score   support

           0       0.88      0.88      0.88        51
           1       0.88      0.88      0.88        48

    accuracy                           0.88        99
   macro avg       0.88      0.88      0.88        99
weighted avg       0.88      0.88      0.88        99
```

Results



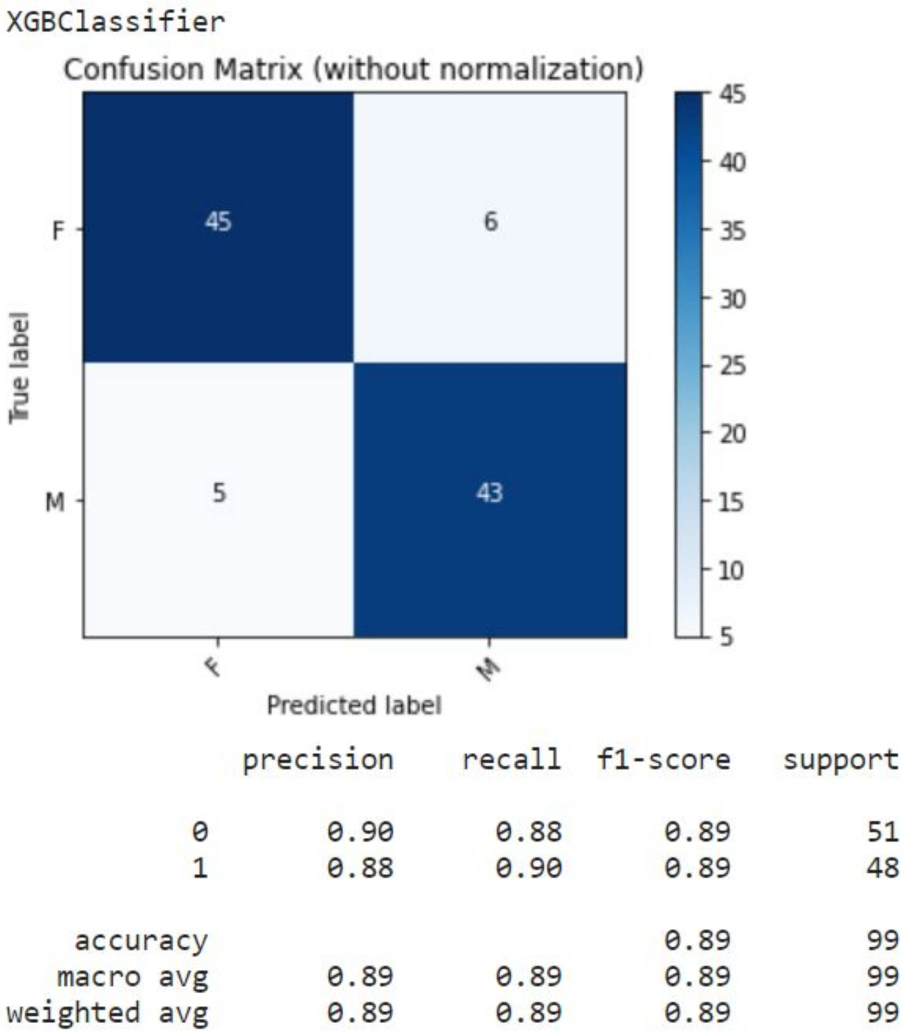
```
CatBoostClassifier
```

```
              precision    recall  f1-score   support

           0       0.90      0.90      0.90        51
           1       0.90      0.90      0.90        48

    accuracy                           0.90        99
   macro avg       0.90      0.90      0.90        99
weighted avg       0.90      0.90      0.90        99
```

Results



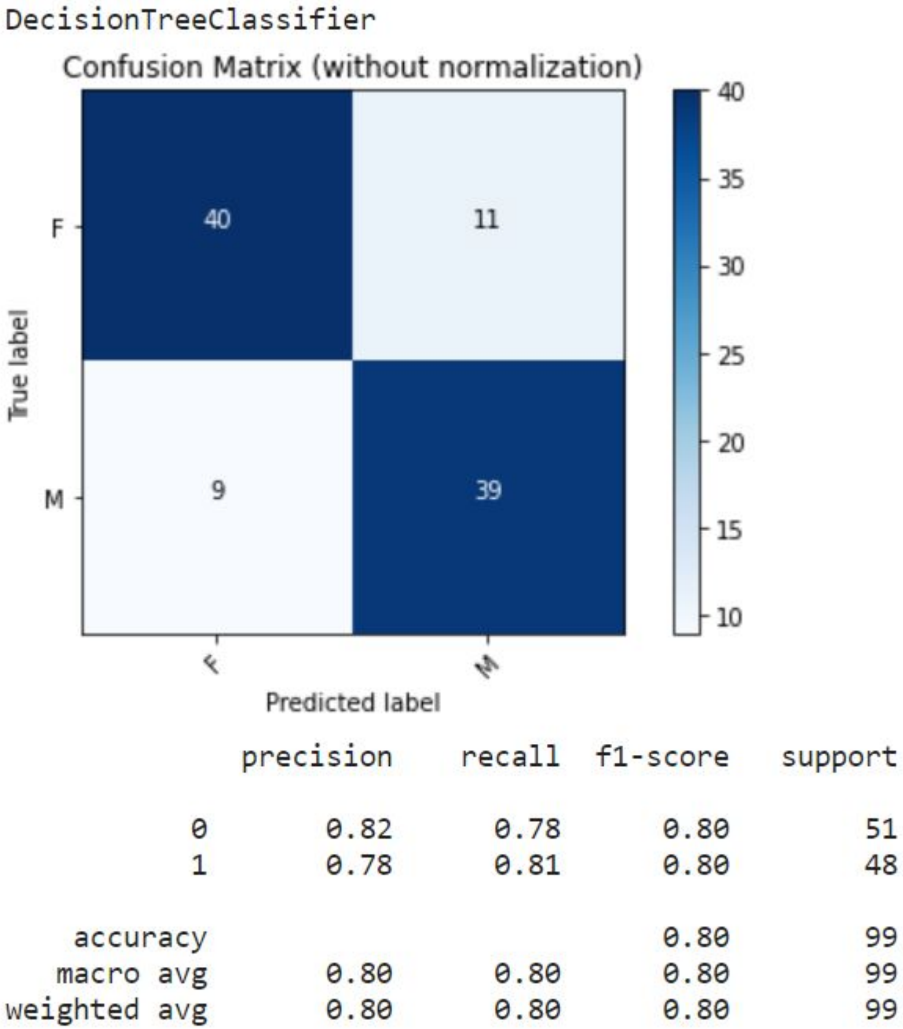
```
XGBClassifier
```

```
              precision    recall  f1-score   support

           0       0.90      0.88      0.89        51
           1       0.88      0.90      0.89        48

    accuracy                           0.89        99
   macro avg       0.89      0.89      0.89        99
weighted avg       0.89      0.89      0.89        99
```

Results



DecisionTreeClassifier

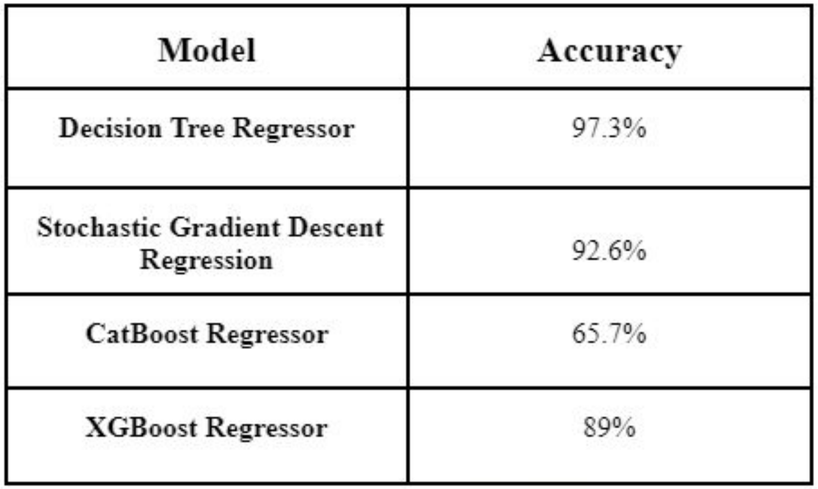
```
              precision    recall  f1-score   support

           0       0.82      0.78      0.80        51
           1       0.78      0.81      0.80        48

    accuracy                           0.80        99
   macro avg       0.80      0.80      0.80        99
weighted avg       0.80      0.80      0.80        99
```

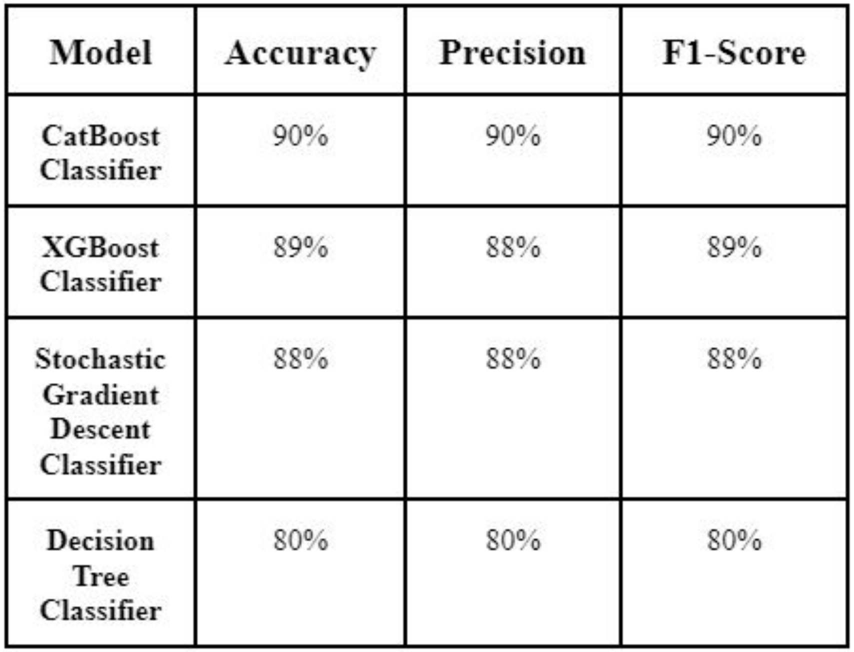
Summary

- Data Acquisition(Self-built Dataset).
- MFCC feature extraction algorithm was used and a total of 20 features were extracted.
- Different types of ML models were used and the final results were compared.



Model	Accuracy
Decision Tree Regressor	97.3%
Stochastic Gradient Descent Regression	92.6%
CatBoost Regressor	65.7%
XGBoost Regressor	89%

a. Results of Regression Models



Model	Accuracy	Precision	F1-Score
CatBoost Classifier	90%	90%	90%
XGBoost Classifier	89%	88%	89%
Stochastic Gradient Descent Classifier	88%	88%	88%
Decision Tree Classifier	80%	80%	80%

b. Results of Classification Models

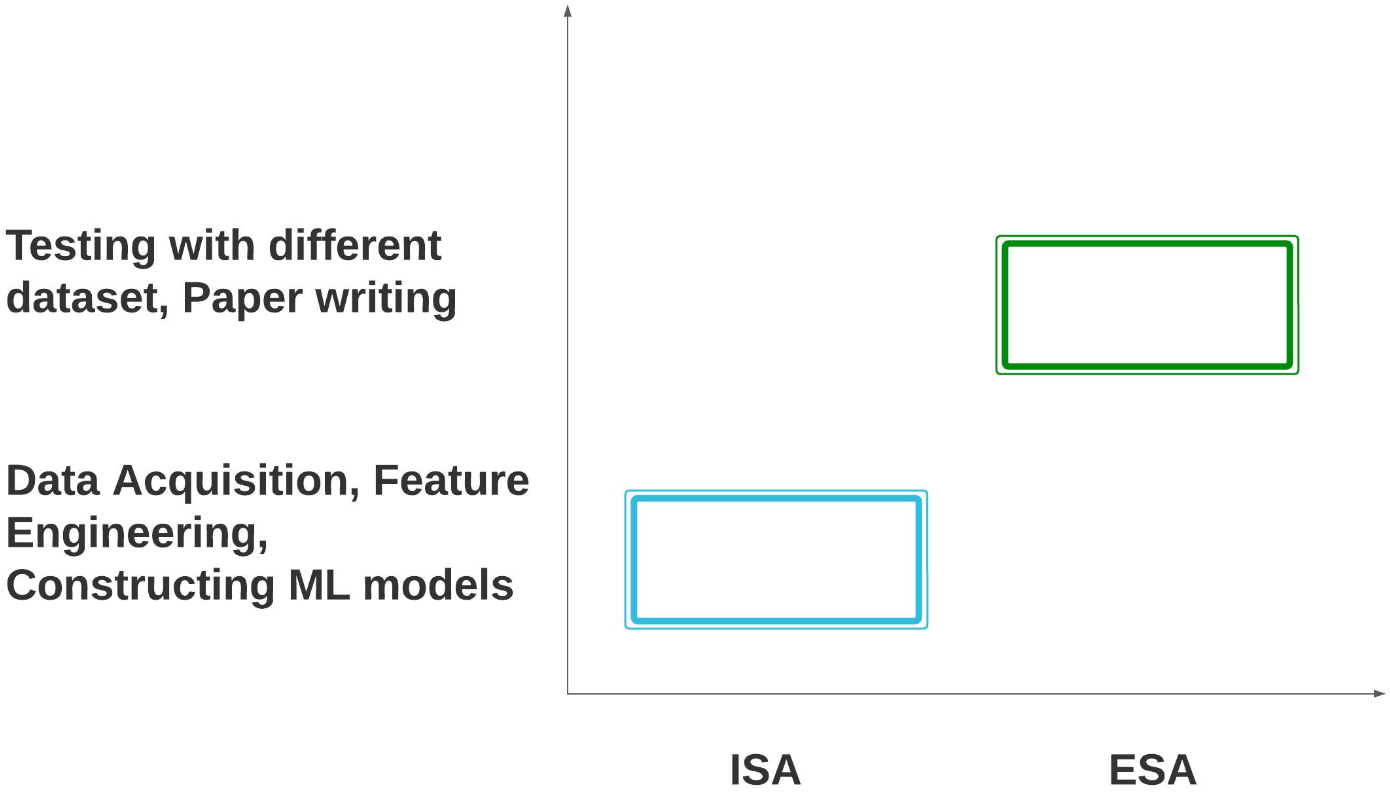
References

- Syed Rohit Zaman, Dipan Sadekeen, M Aqib Alfaz, Rifat Shahriyar, “One Source to Detect them All: Gender, Age, and Emotion Detection from Voice” , 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)
- Ogechukwu Iloanusi, Ugogbola Ejiogu, Ife-ebube Okoye, Ijeoma Ezika, Samuel Ezichi, Charles Osuagwu, “Voice Recognition and Gender Classification in the Context of Native Languages and Lingua Franca”, 2019 6th Intl. Conference on Soft Computing & Machine Intelligence
- Kwasny, D.; Hemmerling, D. “Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks”, Sensors 2021, 21, 4785. https://doi.org/10.3390/s21144785
- Armagan Fidan, Rabia Ozge Bircan, Saeid Karamzadeh, “A New Approach For Age Estimation System Based on Speech Signals”,2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) | 978-1-6654-4930-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISMSIT52890.2021.960461

Deliverables

- Project report
- Research Paper

Project Timeline



Testing with different dataset, Paper writing

Data Acquisition, Feature Engineering, Constructing ML models

ISA

ESA

Q & A

Thank You