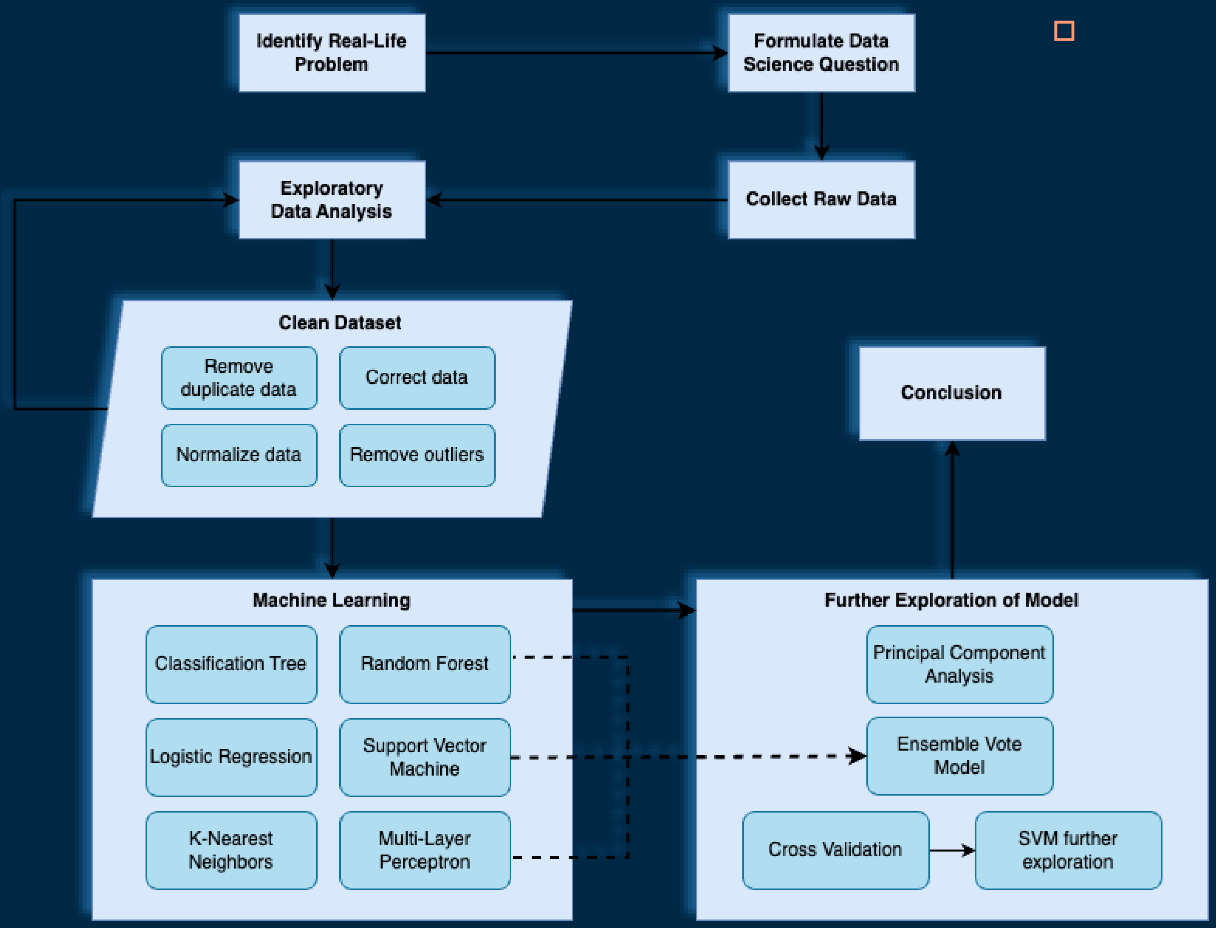
Gender Recognition by Voice

PU FANYI U2220175K

JIANG JINYI U2220259H

SHAN YI U2222846C

Table of Content

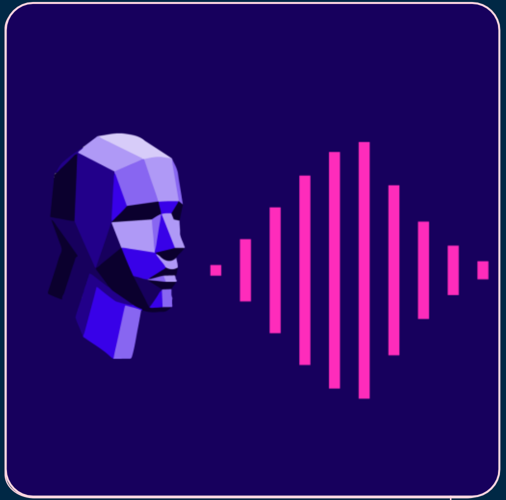

Identify Real-Life Problem
Formulate Data Science Question
Collect Raw Data
Exploratory Data Analysis
Clean Dataset
Remove duplicate data
Correct data
Normalize data
Remove outliers
Conclusion
Machine Learning
Classification Tree
Random Forest
Logistic Regression
Support Vector Machine
K-Nearest Neighbors
Multi-Layer Perceptron
Further Exploration of Model
Principal Component Analysis
Ensemble Vote Model
Cross Validation
SVM further exploration

Motivation & Problem Formulation

01

Motivation

- Gender is strongly associated with voice.
- The relationship between gender and voice has potential implications for various fields.
- By learning and comparing various models, we can identify their strengths and weaknesses and potentially develop more accurate and reliable algorithms.



Problem Formulation

Main Problem

Classify the gender of a speaker based on their voice characteristics

Sub-Problems

01

What are the key features to classify the gender of a speaker through their voice?

02

Which model can predict the gender of a speaker with higher accuracy?

Sample Collection

https://www.kaggle.com/datasets/primaryobjects/voicegender

	meanfreq	sd	median	Q25	Q75	IQR	skew	kurt	sp.ent	sfm	...	centroid	meanfun	minfun	maxfun	meandom	mindom	maxdom	dfrange	modindx	label
0	0.059781	0.064241	0.032027	0.015071	0.090193	0.075122	12.863462	274.402906	0.893369	0.491918	...	0.059781	0.084279	0.015702	0.275862	0.007812	0.007812	0.007812	0.000000	0.000000	male
1	0.066009	0.067310	0.040229	0.019414	0.092666	0.073252	22.423285	634.613855	0.892193	0.513724	...	0.066009	0.107937	0.015826	0.250000	0.009014	0.007812	0.054688	0.046875	0.052632	male
2	0.077316	0.083829	0.036718	0.008701	0.131908	0.123207	30.757155	1024.927705	0.846389	0.478905	...	0.077316	0.098706	0.015656	0.271186	0.007990	0.007812	0.015625	0.007812	0.046512	male
3	0.151228	0.072111	0.158011	0.096582	0.207955	0.111374	1.232831	4.177296	0.963322	0.727232	...	0.151228	0.088965	0.017798	0.250000	0.201497	0.007812	0.562500	0.554688	0.247119	male
4	0.135120	0.079146	0.124656	0.078720	0.206045	0.127325	1.101174	4.333713	0.971955	0.783568	...	0.135120	0.106398	0.016931	0.266667	0.712812	0.007812	5.484375	5.476562	0.208274	male
5	0.132786	0.079557	0.119090	0.067958	0.209592	0.141634	1.932562	8.308895	0.963181	0.738307	...	0.132786	0.110132	0.017112	0.253968	0.298222	0.007812	2.726562	2.718750	0.125160	male
6	0.150762	0.074463	0.160106	0.092899	0.205718	0.112819	1.530643	5.987498	0.967573	0.762638	...	0.150762	0.105945	0.026230	0.266667	0.479620	0.007812	5.312500	5.304688	0.123992	male
7	0.160514	0.076767	0.144337	0.110532	0.231962	0.121430	1.397156	4.766611	0.959255	0.719858	...	0.160514	0.093052	0.017758	0.144144	0.301339	0.007812	0.539062	0.531250	0.283937	male
8	0.142239	0.078018	0.138587	0.088206	0.208587	0.120381	1.099746	4.070284	0.970723	0.770992	...	0.142239	0.096729	0.017957	0.250000	0.336476	0.007812	2.164062	2.156250	0.148272	male
9	0.134329	0.080350	0.121451	0.075580	0.201957	0.126377	1.190368	4.787310	0.975246	0.804505	...	0.134329	0.105881	0.019300	0.262295	0.340365	0.015625	4.695312	4.679688	0.089920	male
10	0.157021	0.071943	0.168160	0.101430	0.216740	0.115310	0.979442	3.974223	0.965249	0.733693	...	0.157021	0.088894	0.022069	0.117647	0.460227	0.007812	2.812500	2.804688	0.200000	male
11	0.138551	0.077054	0.127527	0.087314	0.202739	0.115426	1.626770	6.291365	0.966004	0.752042	...	0.138551	0.104199	0.019139	0.262295	0.246094	0.007812	2.718750	2.710938	0.132351	male
12	0.137343	0.080877	0.124263	0.083145	0.209227	0.126082	1.378728	5.008952	0.963514	0.736150	...	0.137343	0.092644	0.016789	0.213333	0.481671	0.015625	5.015625	5.000000	0.088500	male
13	0.181225	0.060042	0.190953	0.128839	0.229532	0.100693	1.369430	5.475600	0.937446	0.537080	...	0.181225	0.131504	0.025000	0.275862	1.277114	0.007812	2.804688	2.796875	0.416550	male
14	0.183115	0.066982	0.191233	0.129149	0.240152	0.111004	3.568104	35.384748	0.940333	0.571394	...	0.183115	0.102799	0.020833	0.275862	1.245739	0.203125	6.742188	6.539062	0.139332	male
15	0.174272	0.069411	0.190874	0.115602	0.228279	0.112677	4.485038	61.764908	0.950972	0.635199	...	0.174272	0.102046	0.018328	0.246154	1.621299	0.007812	7.000000	6.992188	0.209311	male
16	0.190846	0.065790	0.207951	0.132280	0.244357	0.112076	1.562304	7.834350	0.938546	0.538810	...	0.190846	0.113323	0.017544	0.275862	1.434115	0.007812	6.320312	6.312500	0.254780	male
17	0.171247	0.074872	0.152807	0.122391	0.243617	0.121227	3.207170	25.765565	0.936954	0.586420	...	0.171247	0.079718	0.015671	0.262295	0.106279	0.007812	0.570312	0.562500	0.138355	male
18	0.168346	0.074121	0.145618	0.115756	0.239824	0.124068	2.704335	18.484703	0.934523	0.559742	...	0.168346	0.083484	0.015717	0.231884	0.146563	0.007812	3.125000	3.117188	0.059537	male
19	0.173631	0.073352	0.153569	0.123680	0.244234	0.120554	2.804975	20.857543	0.930917	0.518269	...	0.173631	0.090130	0.015702	0.210526	0.193044	0.007812	2.820312	2.812500	0.068124	male
20	0.172754	0.076903	0.177736	0.120070	0.245368	0.125298	2.967765	20.078115	0.925539	0.523081	...	0.172754	0.093574	0.015764	0.200000	0.235877	0.007812	0.718750	0.710938	0.235069	male

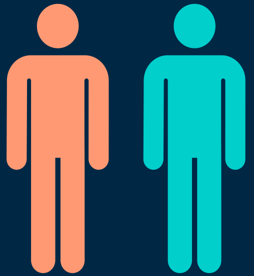
Data Preparation & Exploratory Analysis

02

Dataset Information

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3168 entries, 0 to 3167
Data columns (total 21 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   meanfreq  3168 non-null   float64
 1   sd        3168 non-null   float64
 2   median    3168 non-null   float64
 3   Q25       3168 non-null   float64
 4   Q75       3168 non-null   float64
 5   IQR       3168 non-null   float64
 6   skew      3168 non-null   float64
 7   kurt      3168 non-null   float64
 8   sp.ent    3168 non-null   float64
 9   sfm       3168 non-null   float64
 10  mode      3168 non-null   float64
 11  centroid  3168 non-null   float64
 12  meanfun   3168 non-null   float64
 13  minfun    3168 non-null   float64
 14  maxfun    3168 non-null   float64
 15  meandom   3168 non-null   float64
 16  mindom    3168 non-null   float64
 17  maxdom    3168 non-null   float64
 18  dfrange   3168 non-null   float64
 19  modindx   3168 non-null   float64
 20  label     3168 non-null   object 
dtypes: float64(20), object(1)
memory usage: 519.9+ KB
```

3168 voice data
21 features



Half male
Half female

Understanding Data

Central Tendency of Frequency

- meanfreq
- median
- Q25
- Q75
- centroid

Spread of Frequency

- IQR
- sd
- skew
- kurt

Spectral Features

- sp.ent
- sfm
- mode

Fundamental Frequency Features

- meanfun
- minfun
- maxfun
- modindx

Dominant Frequency Features

- meandom
- mindom
- maxdom
- dfrange

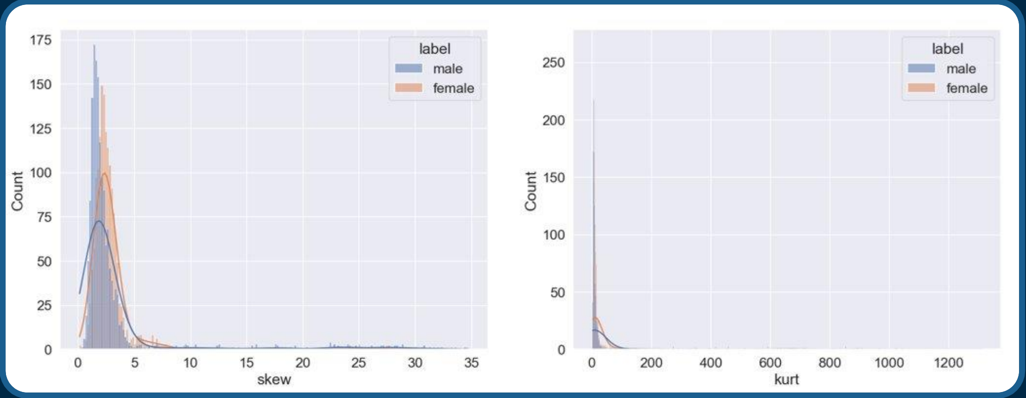
Remove Duplicate Data

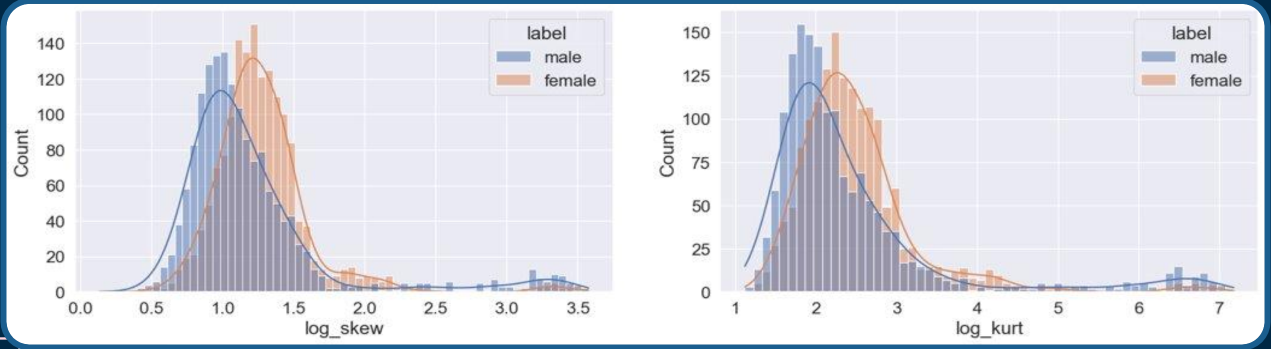
- **meanfreq**: mean frequency (in kHz)
- **centroid**: frequency centroid
- They have the same definition!

```python
assert (voice_data['meanfreq'] == voice_data['centroid']).all()
voice_data.drop('centroid', axis=1, inplace=True)
```

Data Correction

$$x' = \ln(x + 1)$$





Data Normalization

- Many models are sensitive to the scale of input features.
- Two points: $(0, 0), (1, 1000)$
- The main contributor to the distance is y.
- Ensure fair feature weighting.
- $x' = \frac{x - \overline{x}}{\sigma}$, where σ is the distance deviation of x.

Outliers Removal

```python
from sklearn.ensemble import IsolationForest

clf = IsolationForest(contamination=0.05)
clf.fit(X)
outlier_mask = clf.predict(X) == -1
X = X[~outlier_mask]
y = y[~outlier_mask]
```

Machine Learning

03

Machine Learning

Tree Based Algorithms

- Classification Tree
- Random Forest

Numerical Algorithms

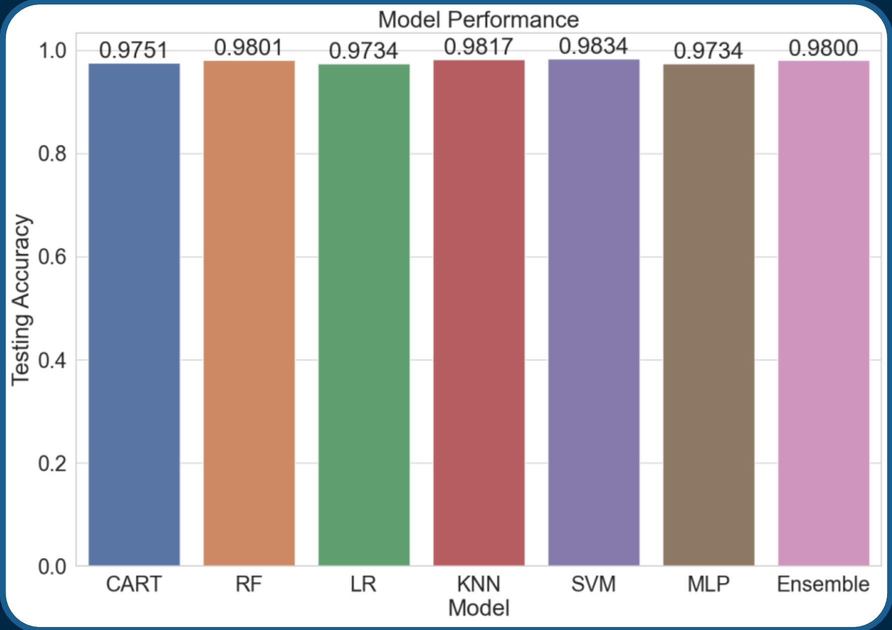
- Logistic Regression
- K Nearest Neighbour
- Support Vector_Machines
- Multi-Layer Perceptron

Further Exploration

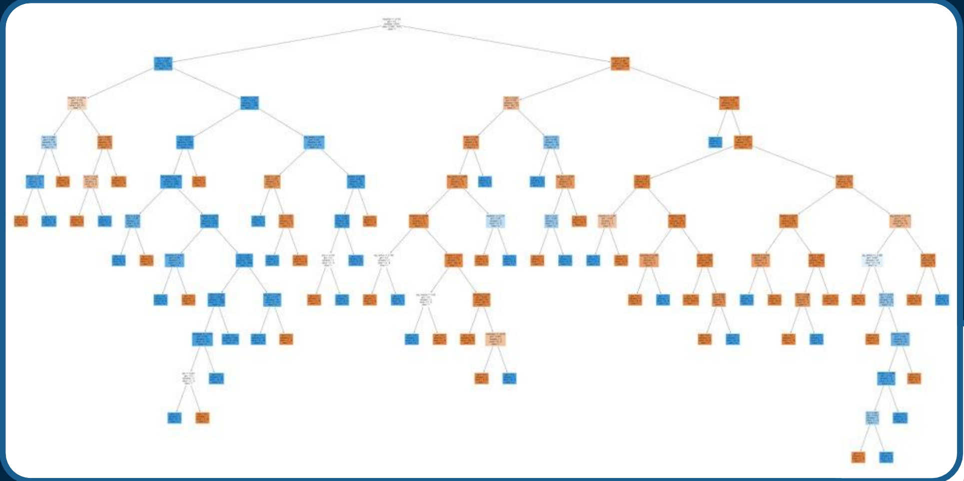
- Cross Validation
- Support Vector Machines
- Principal component analysis
- Ensemble Vote model

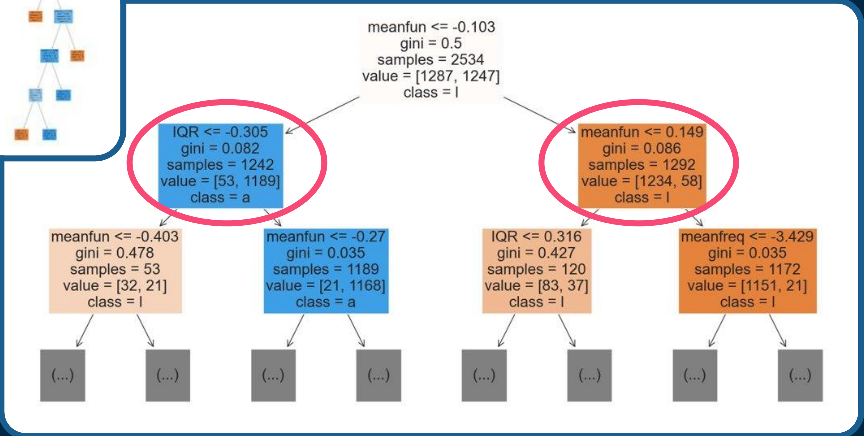
Machine Learning Overview

Model	Training	Testing
Classification Tree	1.0000	0.9751
Random Forest	1.0000	0.9801
Logistic Regression	0.9763	0.9734
K-Nearest Neighbors	1.0000	0.9817
Support Vector Machine	0.9896	0.9834
Multi-Layer Perceptron	1.0000	0.9734
Ensemble Vote	1.0000	0.9800



Classification Tree (CART)





The most important two variables:
IQR and meanfun

Logistic Regression

- Use regression to predict $p(y = \text{male} \mid X = x)$

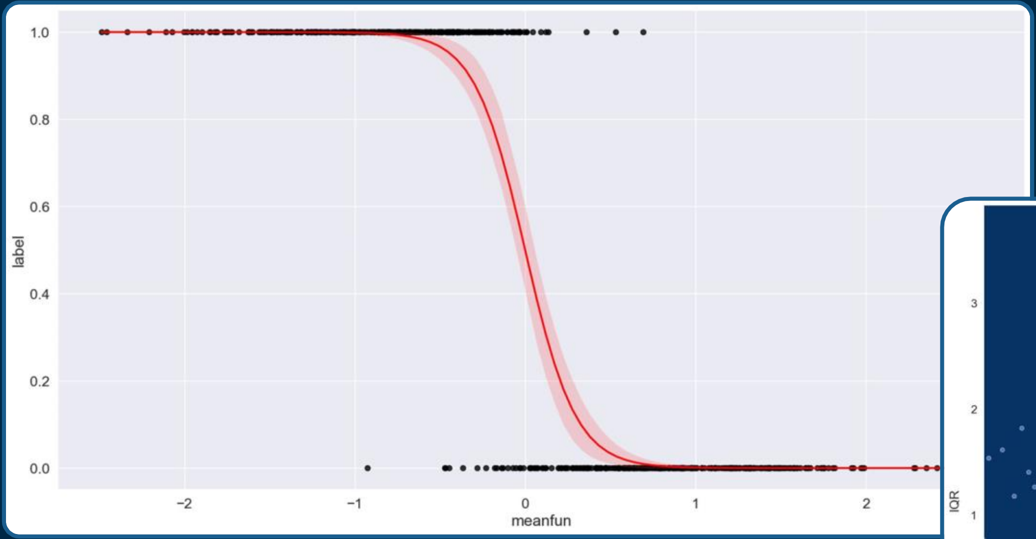
- Use the function:

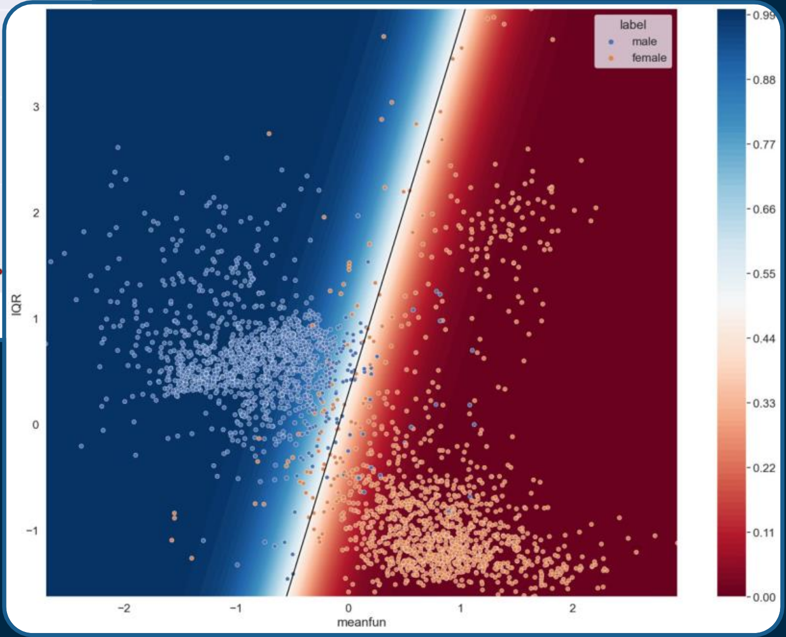
$$\hat{p}(y = male \mid X = x) = \frac{e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}} = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{i=1}^{n} \beta_i x_i\right)}}$$

- It is like linear regression:

$$\beta_0 + \beta^T x = \ln \frac{\hat{p}}{1 - \hat{p}}$$

Logistic Regression Demonstration



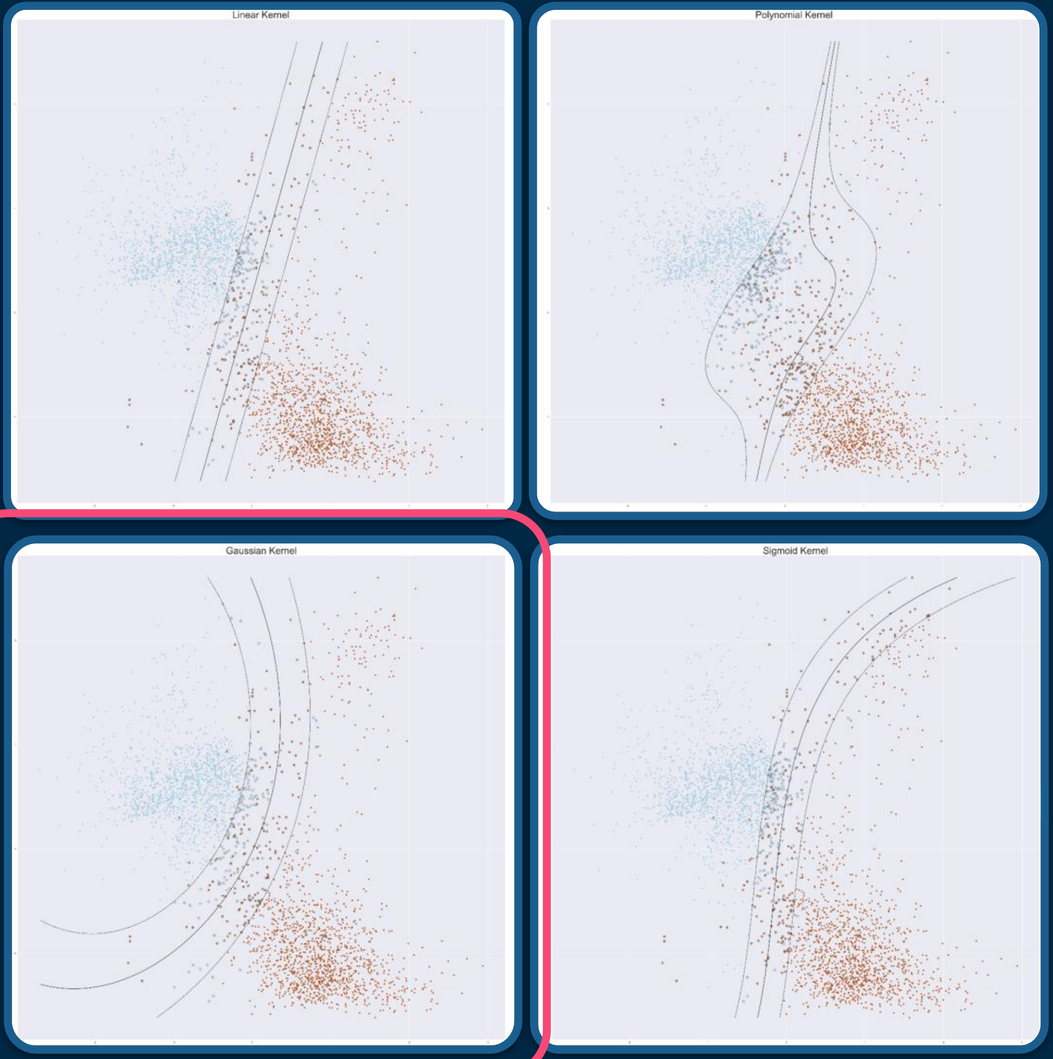


Cross Validation (CV)

- Evaluate, select models
- Splits data into folds
- Reliable performance assessment

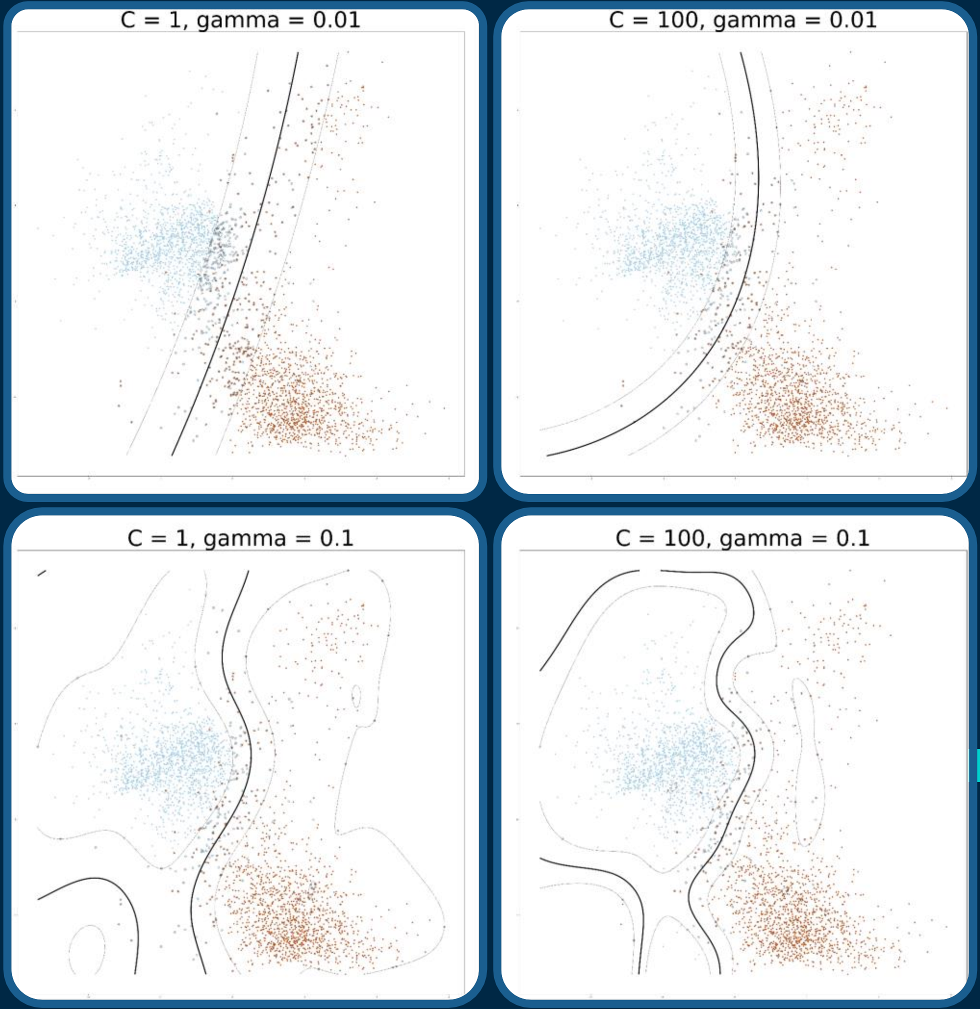
Support Vector Machines (SVM)

- **Main idea: hyperplane**
- **4 types of Kernel**
  - **Linear**: $x^T x'$
  - **Polynomial**: $\left(x^T x' + 1\right)^d$
  - **Gaussian**: $\exp\left(-\gamma |x - x'|^2\right)$
  - **Sigmoid**: $\tanh\left(\gamma x^T x' + r\right)$



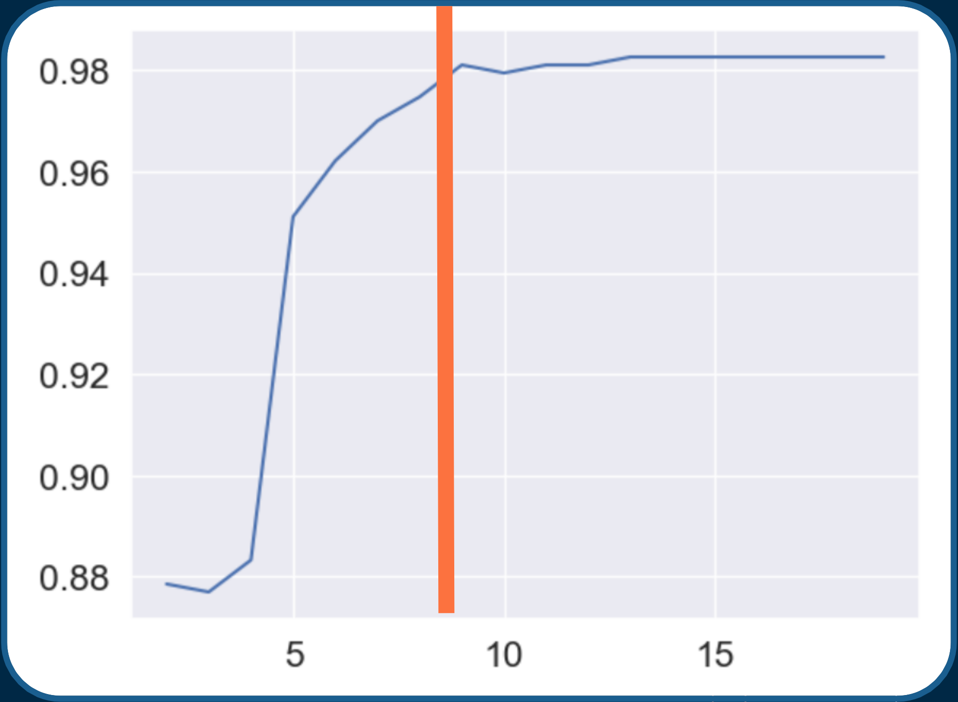
Support Vector Machines (SVM)

- γ: determines smoothness
- C: determines margin, misclassifications
- Hill Climbing
- $\gamma = 0.0026, C = 23.88$



Principal Component Analysis (PCA)

- **Feature Compression**
- **Improve Efficiency**
- **Balance Between Accuracy and Efficiency**



Ensemble Vote Model

- Something we developed
- Integrate the outputs of high-performing models and select the majority vote
- Not as ideal as we hoped

Outcome & Insights

04

Outcome and Data-Driven Insight

What are the key features to differentiate the gender?

- According to classification tree analysis, IQR and meanfun have been identified as the two main predictors for classifying male and female voices. A higher IQR and lower meanfun are more indicative of a male speaker.

Which models can better predict the gender of a speaker?

- Among the various models, we found that the SVM model with an RBF kernel achieved the highest accuracy, with a score of 0.9834.

Recommendation

Application

- Speech recognition
- Security systems

Limitation

- Non-acoustic factors
- Practicality

What we learned

- Importance of Data Preparation

  Data normalization → significant improvement in accuracy

- Exploring Various Machine Learning Models for Accurate Predictions
  - Supervised learning: CART, RF, LR, KNN, SVM
  - Unsupervised learning: PCA
  - Using CV to get the accuracy

- Ensemble Vote Model
  - Carefully select models based on their individual strengths and weaknesses
  - Consider the underlying assumptions and limitations of each model

Thank you