Gender voice recognition

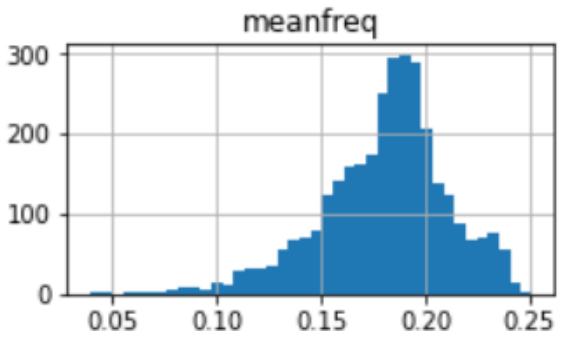
AUTORZY: MICHAŁ PIASECKI, BARTOSZ SIŃSKI

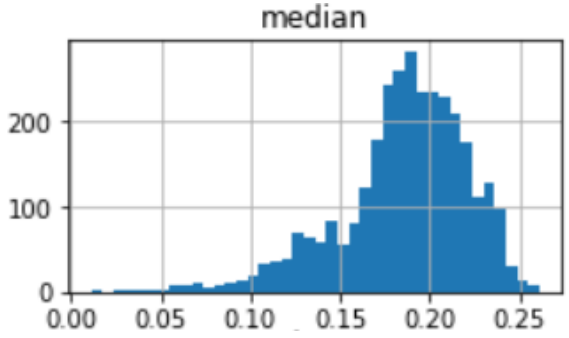
Eksploracja danych

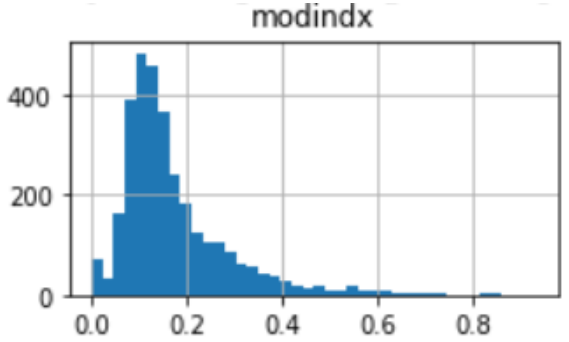
Wszystkie zmienne poza targetem są ilościowe.

Niektóre zmienne bardzo mocno skorelowane.

Podział danych na podstawie wartość zmiennej "meanfun" dał accuracy 0.95.

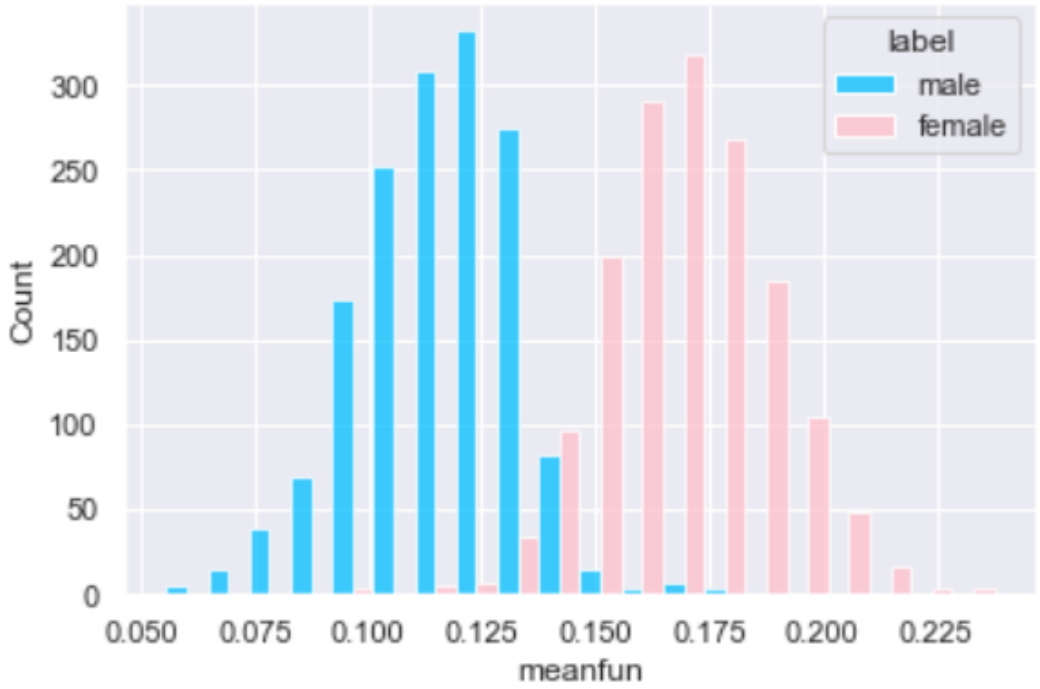






dfrange	maxdom	0.999
maxdom	dfrange	0.999
meanfreq	median	0.920
median	meanfreq	0.920
	centroid	0.920
centroid	median	0.920
meanfreq	Q25	0.906
Q25	meanfreq	0.906
	centroid	0.906
centroid	Q25	0.906
sp.ent	sfm	0.894
sfm	sp.ent	0.894
skew	kurt	0.888
kurt	skew	0.888
sd	sfm	0.879
sfm	sd	0.879
Q25	IQR	0.870
IQR	Q25	0.870
sp.ent	sd	0.861
sd	sp.ent	0.861
	IQR	0.859
IQR	sd	0.859
sd	Q25	0.831
Q25	sd	0.831
sfm	centroid	0.824

	name	type
0	meanfreq	float
1	sd	float
2	median	float
3	Q25	float
4	Q75	float
5	IQR	float
6	skew	float
7	kurt	float
8	sp.ent	float
9	sfm	float
10	mode	float
11	centroid	float
12	meanfun	float
13	minfun	float
14	maxfun	float
15	meandom	float
16	mindom	float
17	maxdom	float
18	dfrange	float
19	modindx	float
20	label	string



Feature Engineering

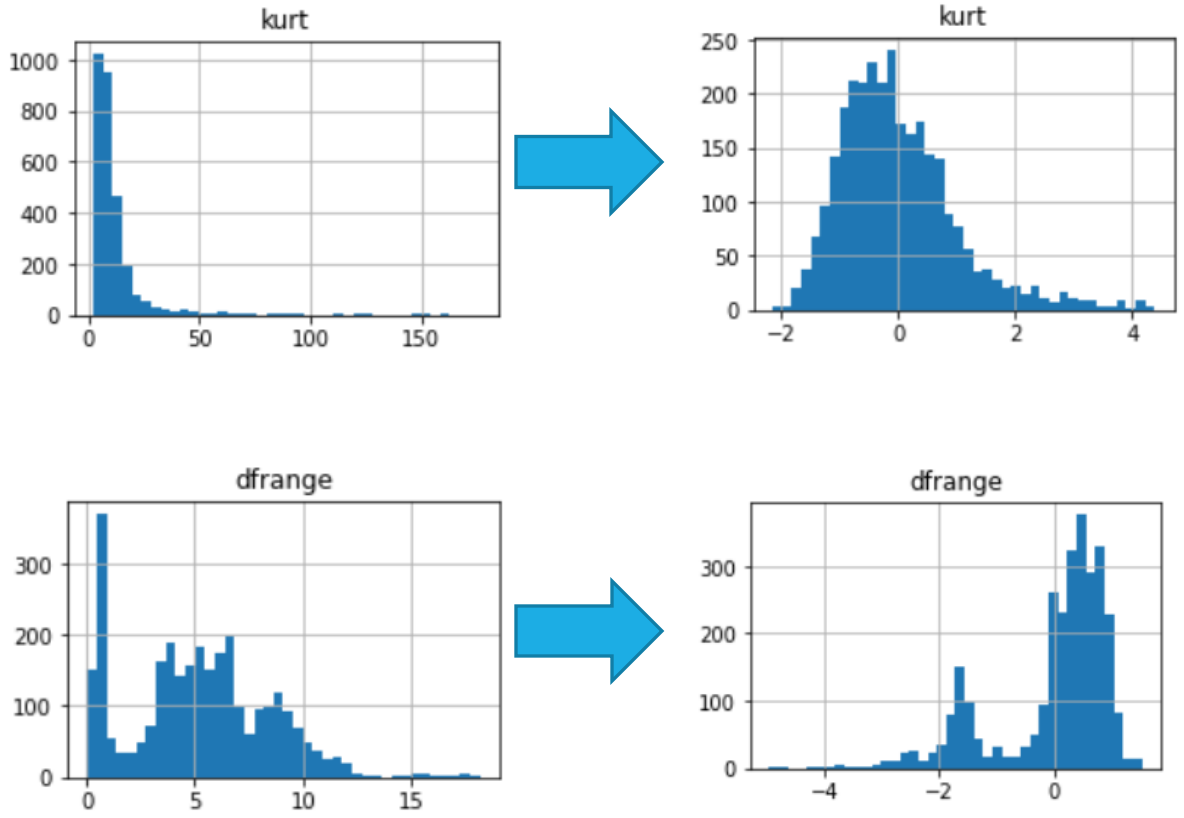
Usunięcie 219 outlierów.

Ręczne usunięcie 7 najbardziej skorelowanych kolumn.

Usunięcie 10 zmiennych za pomocą SelectKBest().

Standaryzacja danych.

Baseline model z 0.98 accuracy.



Trenowanie modelu

Klasyfikator SVM bez wyboru hiperparametrów miał 0.991 accuracy.

Tunning parametrów *C, kernel i gamma* podniosły accuracy o 0.002.

Zastostowanie Polymonial features pogorszyło wynik.

Najlepsze parametry:

```
{'classifier__C': 8, 'classifier__gamma': 0.3, 'classifier__kernel': 'rbf'}
```

Wyniki

Accuracy: 0.993

Precision: 0.997

Recall: 0.998

ROC AUC: 0.993

Średnie accuracy na CV : 0.968

	Actual positives	Actual negatives
Positive predictions	354	4
Negative predictions	1	379