Project E6: KAGGLE-GENDER-REC

Predicting Gender by Voice and Speech Analysis

Team members: Kaidi Tootmaa, Hanna Vallner

Goal 1: Create a model that predicts the gender of a person based on voice and speech analysis data from said measurements

Goal 2: Find out which features have the most impact on our prediction results

Repository: https://github.com/HannaVallner/gender-rec

Business understanding

- **Identifying your business goals**
    - Background

Clearly genderwise defined voices sound more natural to the human ear, this can benefit developing different virtual assistants and AI based NPCs, and also aid in choosing and/or checking voice actors' suitability for their roles.

- Business goals

Find out which speech features have the most impact on defining a person's voice as either feminine or masculine. Create intervals for both genders for such features, where the genders are clearly defined within said features.
Train a machine learning model, which could be used to predict the gender of a person based on voice and speech analysis, which the interested parties could use to verify the gender to match their desired expectations. If implemented correctly, the model can help companies increase their client base's overall satisfaction levels with their product(s).

- Business success criteria

Our business goals are achieved when our model and overall analysis gets beneficially used in a company's production.

- **Assessing your situation**
    - Inventory of resources

Our project will be conducted by the project team members, sourced on the data from Kaggle, analysed using Jupyter notebook and required Python modules (Pandas, Numpy etc).

- Requirements, assumptions, and constraints

The code will have to be ready for assessment by December the 12th, the presentation of outcomes by December the 15th. As the data is from an open source, we won't have any legal or security obligations to follow. The assumed workload for the project is > 60 hours.

- Risks and contingencies

The risks that could delay the completion of the project include finding suitable workhours for the project team, technical difficulties and dissatisfaction with the dataset. For the first risk, the solution is to divide and conquer - make agreements and compromises to match eachother's schedules or firmly divide some of the workload to fulfill individually. If we face technical difficulties, we have the opportunity to consult with the tech support provided by the university. And when facing problems with the dataset, we can explore the Internet for additional datasets and information, to expand our data.

- Terminology

The terms used in our dataset and their explanations:

- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness (measure of symmetry)
- kurt: kurtosis (measures the heaviness of a distribution's tail)
- sp.ent: spectral entropy (measure of signal irregularity)
- sfm: spectral flatness (measure used in digital signal processing to characterize an audio spectrum)
- mode: mode frequency
- centroid: frequency centroid
- peakf: peak frequency (frequency with highest energy)
- fundamental frequency: lowest frequency
- meanfun: average of fundamental frequency measured across acoustic signal
- minfun: minimum fundamental frequency measured across acoustic signal
- maxfun: maximum fundamental frequency measured across acoustic signal
- dominant frequency: the frequency that is most heard
- meandom: average of dominant frequency measured across acoustic signal
- mindom: minimum of dominant frequency measured across acoustic signal
- maxdom: maximum of dominant frequency measured across acoustic signal
- dfrange: range of dominant frequency measured across acoustic signal
- modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range

- Costs and benefits

The cost for the project is only the time spent on it. The benefit of the project for us is acquiring new knowledge on data science and its possible usages, and testing said knowledge by creating a larger data science project for the first time.

- **Defining your data-mining goals**
    - Data-mining goals

Our project will deliver a model that predicts the gender of a person based on voice and speech analysis data. We will also create a thorough analysis on which features play the biggest role in defining a person's voice as either feminine or masculine. To report our findings, we will give a poster presentation on all our findings.

- Data-mining success criteria

The project will be assessed by using test data and analysing the machine learning model's accuracy, recall, precision and F-measure. Our wish is to achieve an accuracy percentage higher than 95%. In the analysis, we will bring out specific quantitative measures for each of the attributes and related findings.

Data understanding

- **Gathering data**
    - Outline data requirements

For this project we will need a dataset, consisting of analysed measures for each voice recording (the features are listed above). As analysing voices does not require using recordings from different timelines (the voices stay the same), we do not have a required time range. But in order for us to draw valid conclusions, we need the data to have a significant amount of sample instances.

- Verify data availability

As the data we will be using is from an open source and has a substantial amount of instances, we will have access to the whole dataset, which should suffice. Therefore we will not need any additional or alternative data.

- Define selection criteria

Our data is sourced from Kaggle, which contains collective findings from four different databases (e.g The Harvard-Haskins Database of Regularly-Timed Speech). As we are planning on analysing each of the given measure's relevance to the outcome, we will initially need to use all of the features.

We have successfully obtained and accessed the data, and imported it to our data-mining platform. We did not run into any problems or limitations during said process.

- **Describing data**

Our data is sourced from Kaggle formatted in CSV (comma-separated values) file. The dataset has 3168 cases, with 21 fields for each of them.

Fields and their corresponding descriptions:

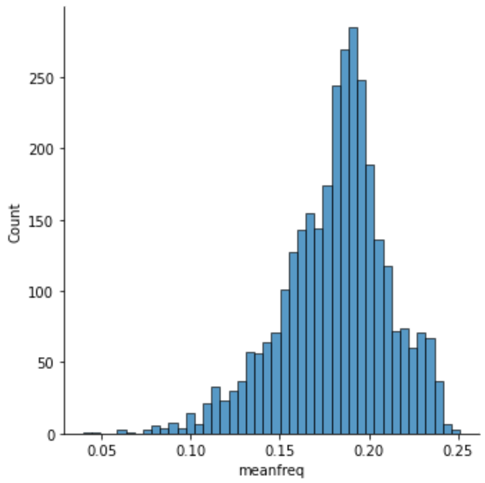
- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness (measure of symmetry)
- kurt: kurtosis (measures the heaviness of a distribution's tail)
- sp.ent: spectral entropy (measure of signal irregularity)
- sfm: spectral flatness (measure used in digital signal processing to characterize an audio spectrum)
- mode: mode frequency
- centroid: frequency centroid
- peakf: peak frequency (frequency with highest energy)
    - fundamental frequency: lowest frequency
- meanfun: average of fundamental frequency measured across acoustic signal
- minfun: minimum fundamental frequency measured across acoustic signal
- maxfun: maximum fundamental frequency measured across acoustic signal
    - dominant frequency: the frequency that is most heard
- meandom: average of dominant frequency measured across acoustic signal
- mindom: minimum of dominant frequency measured across acoustic signal
- maxdom: maximum of dominant frequency measured across acoustic signal
- dfrange: range of dominant frequency measured across acoustic signal
- modindx: modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
- label: female or male

As the dataset is rather large, it has enough findings for our project. The attributes brought out above are ideal for reaching our goals.

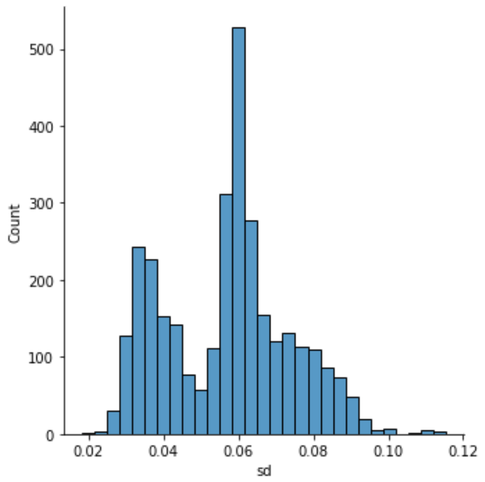
- **Exploring data**

The ranges and distributions of all the features in our dataset:

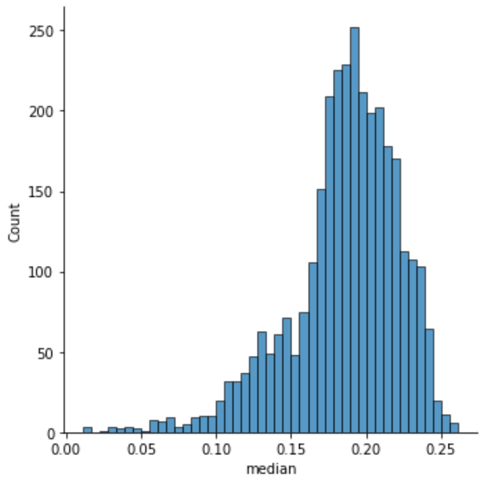
- Column: meanfreq – Min: 0.0393633425835608 – Max: 0.251123758720282



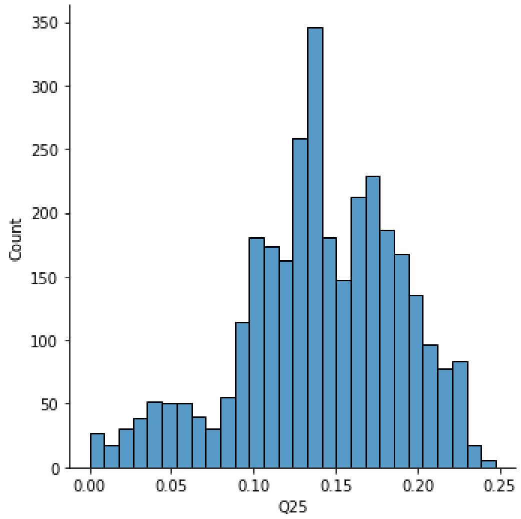
- Column: sd – Min: 0.018363242444455 – Max: 0.115273246743733



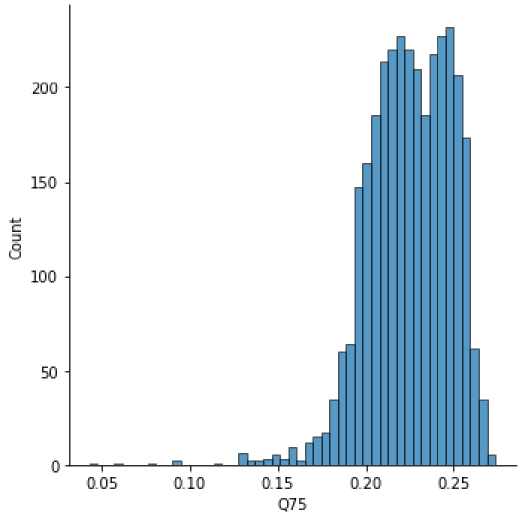
- Column: median – Min: 0.0109745762711864 – Max: 0.261224489795918



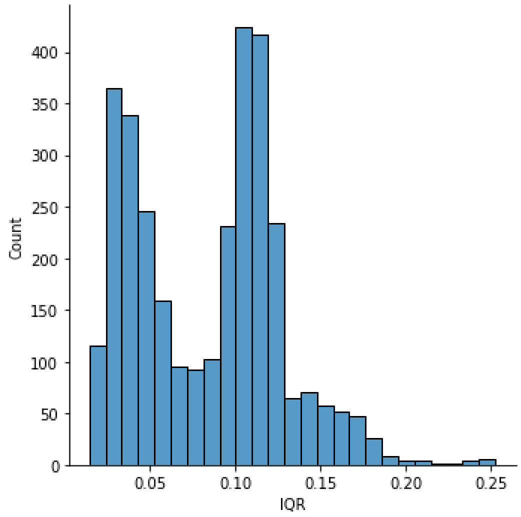
- Column: Q25 – Min: 0.0002287581699346 – Max: 0.24734693877551



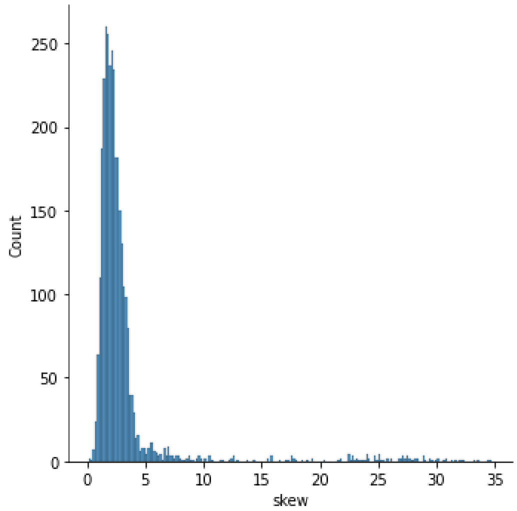
- Column: Q75 – Min: 0.042946273830156 – Max: 0.273469387755102



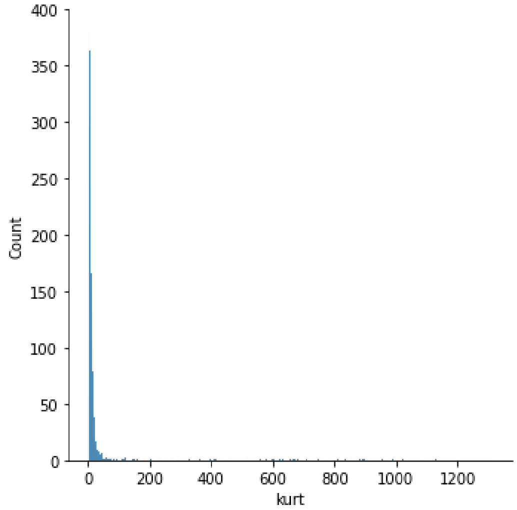
- Column: IQR – Min: 0.0145577312626604 – Max: 0.252225201072386



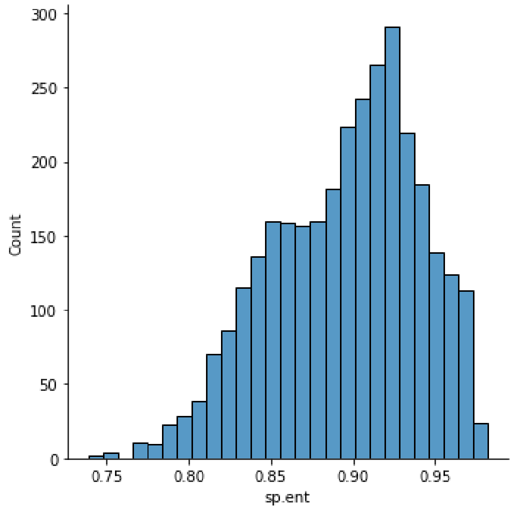
- Column: skew – Min: 0.141735424138914 – Max: 34.7254532660205



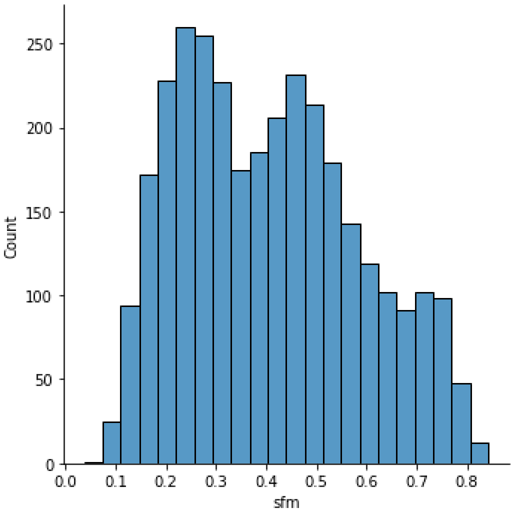
- Column: kurt – Min: 2.06845549084691 – Max: 1309.61288737064



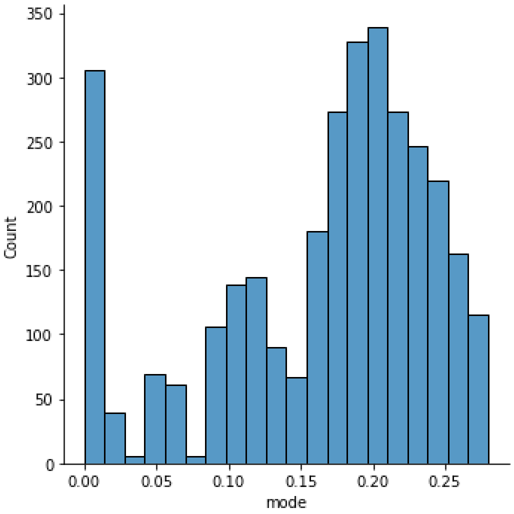
- Column: sp.ent – Min: 0.738650686223529 – Max: 0.98199658896419



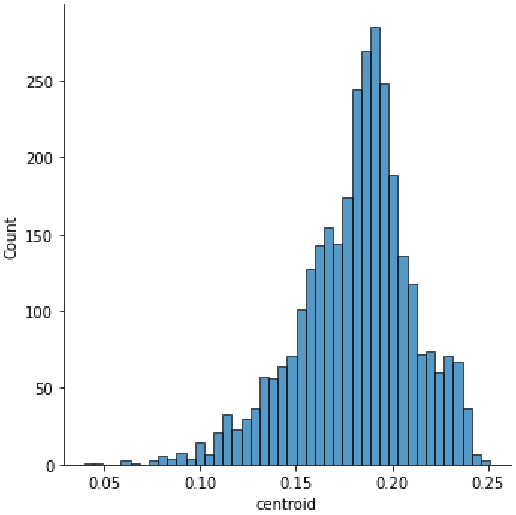
- Column: sfm – Min: 0.0368764745063272 – Max: 0.842935931446768



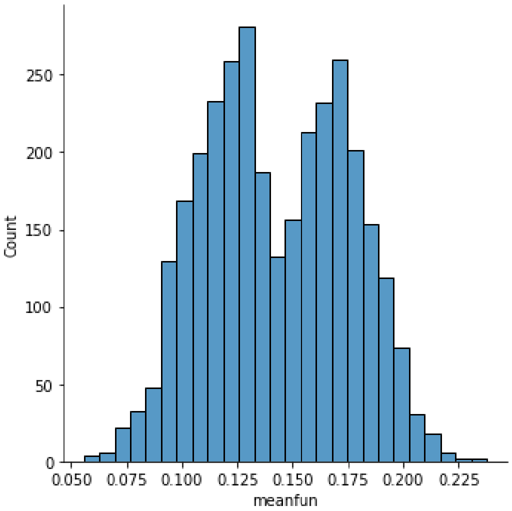
- Column: mode – Min: 0.0 – Max: 0.28



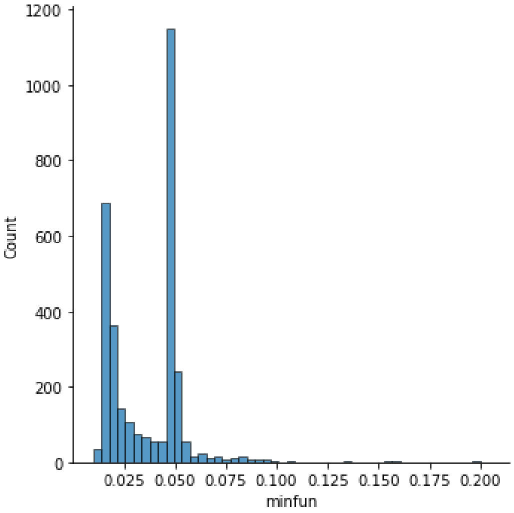
- Column: centroid – Min: 0.0393633425835608 – Max: 0.251123758720282



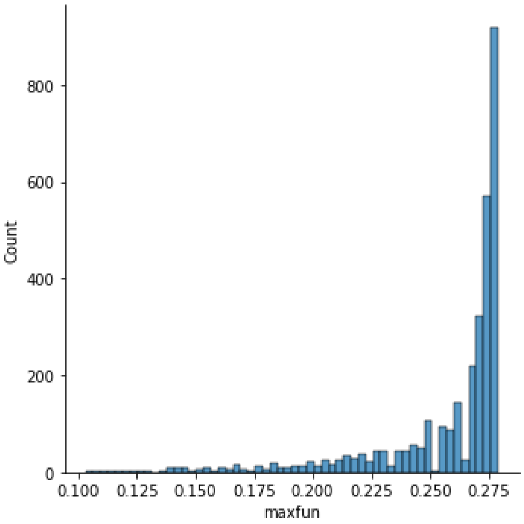
- Column: meanfun – Min: 0.0555653493134555 – Max: 0.237636387269209



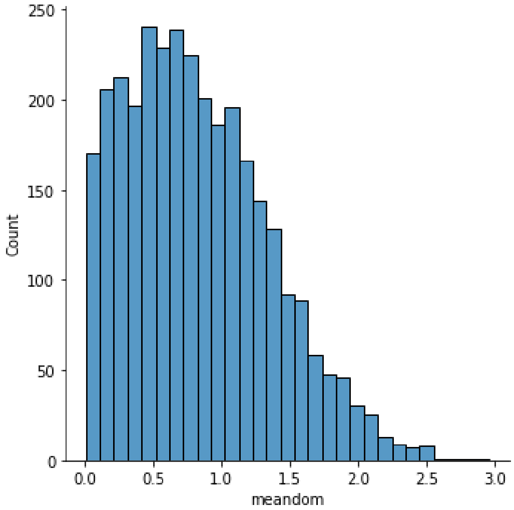
- Column: minfun – Min: 0.0097751710654936 – Max: 0.204081632653061



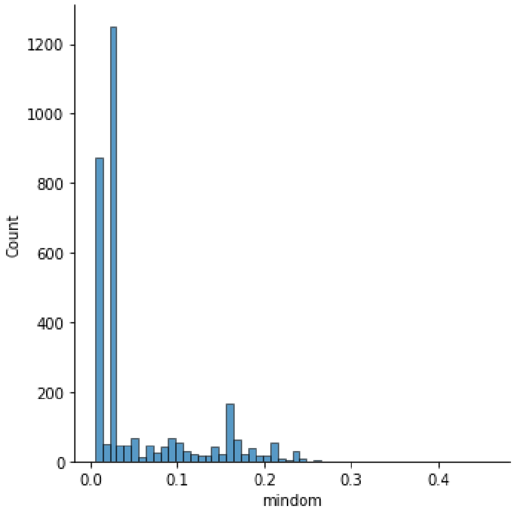
- Column: maxfun – Min: 0.103092783505155 – Max: 0.279113924050633



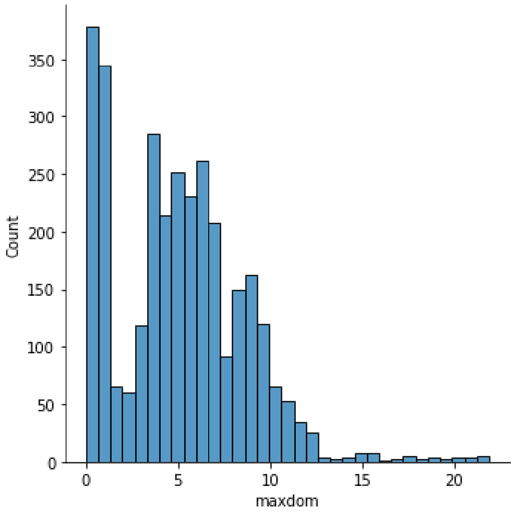
- Column: meandom – Min: 0.0078125 – Max: 2.95768229166667



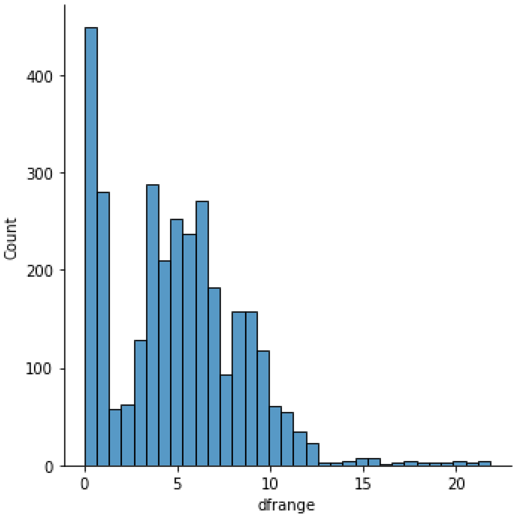
- Column: mindom – Min: 0.0048828125 – Max: 0.458984375



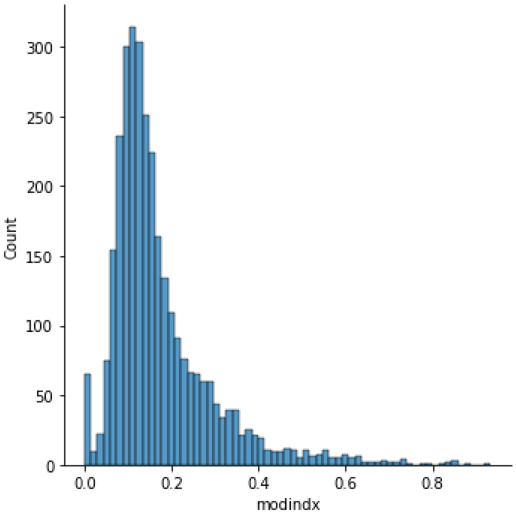
- Column: maxdom – Min: 0.0078125 – Max: 21.8671875



- Column: dfrange – Min: 0.0 – Max: 21.84375



- Column: modindx – Min: 0.0 – Max: 0.932374100719425



- Column: label – female – male

While some of the values do not have a clear distribution (e.g mindom, dfrange), most of them seem to be normal distributions, which is expected from natural phenomena. Some of the values are bimodal (have two peaks), therefore we are making an assumption that one of the peaks corresponds to the male values and the other to the female values.

The data seems to be of good quality, as is to be expected from a Kaggle dataset. We did not find any signs of data quality problems.

- Verifying data quality

As brought out above, our dataset doesn’t have any issues with data quality. The data we need for our project all exists within the dataset and is accessible with no restrictions.

Project plan

Make a detailed plan of your project with a list of tasks. There should be at least 5 tasks. Specify how many hours each team member is going to contribute to each task.
List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.

1. Data exploration (get more familiar with our dataset, find out its' potential concerns) – 4 hours each
2. (Split our data to test and training sets – doesn't take a lot of time, maybe 15 minutes, but is an important part of our project)
3. Decide on which models to use (explore different possibilities and their suitability with our project and test their initial outcomes) – 2 hours each
4. Find suitable parameter settings (we will try GridSearch and halving) – 4 hours each
5. Analyse models (which suit our dataset and goals the best, analyse based on accuracy, recall, precision and T-measure, while keeping in mind the possibility of under- and overfitting (we are going to use cross validation to avoid that)) – 8 hours each
6. Feature correlation analysis (analyse the impact that different measures have on the outcome) – 6 hours each
7. Summarize our findings, within doing so, prepare for the poster presentation – 10 hours each