Progress Report

Gabriel Hili (13502H)

Gabriel Hili
University of Malta
L-Università ta' Malta
Msida, Malta
gabriel.hili.20@um.edu.mt

1. INTRODUCTION

Machine learning is a method of teaching computers to learn and make decisions based on data, without explicitly programming them. It involves training a model on a large dataset, and allowing the model to make predictions or decisions based on this data. Machine learning has been successfully applied to a wide range of problems in many different fields, namely sex prediction by voice. In this project five different supervised machine learning techniques will be applied on a speaker dataset in order to predict the sex of the speaker.

Firstly, the raw data is preprocessed and uninformative features are dropped in order to reduce training complexity. This is done by plotting box-plots and kde-plots of all 20 features of the original dataset and then gauging how distinct the distributions are between male and females. Then, the data is passed to the machine learning algorithms, namely Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbour (KNN), Decision Tree, and Logistic Regression. The hyperparameters of the algorithms are varied in order to observe which hyperparameters perform the best on the current dataset. Lastly, the results are aggregated, evaluated, and concluded.

2. DATA PREPROCESSING

Initially, the data had 20 features excluding the label. Each feature was a statistical measure on the frequency (kHz) of the sound sample, such as the mean, median, mode, kurtosis, skewness, and IQR. The main approach of the preprocessing was to only keep the features that showed a clear distinction between male and female. To do this, the data was visualised using boxen plots and KDE plots.

- **IQR** - The interquartile range, Q25 and Q75, are calculated on the frequencies sampled from the voice recording.
- `sp.ent` - Spectral Entropy. The sepctral power distribution along with forecastability of time. Spectral Power is the amplitude of the different frequency components (pure sine waves) that make up the entire signal.
- `sfm` - Spectral Flatness. Quantifies how much a sound resembles a pure tone, as opposed to being noise-like.
- **Fundamental Frequency** - The lowest frequency of a sinusodial waveform making up our sound. This term encapsulates the variables `meanfun`, `minfun`, and `maxfun`
- **Dominant Frequency** - The frequency that is the most heard, and is always a multiple of the fundamental frequency. The dominant frequency may be equal to the fundamental frequency. This term encapsulates the variables `meandom`, `mindom`, `maxdom`, `dfrange`
- `modindx` - Modulation Index. Defined as the ratio of the peak frequency deviation of the carrier wave to the frequency of the modulating sine wave.

After preprocessing, only 9 features (excluding the label) were kept, namely `sd, Q25, IQR, sp.ent, sfm, meanfun, mindom, dfrange, label`.

2.1 Intitial Observations

Firstly it was discovered that the **first 50% of the data was male, and the other 50% was female**. Moreover, it was also discovered that **the feature columns `centroid` and `meanfreq` were identical**. Hence the data was shuffled in order to create a representative train-test split when passing to machine learning models, and `centroid` was dropped in order to remove redundant data, which adds extra complexity while training.

2.2 Data Visualisation

Visualisation on the data made it obvious which features where informative on distinguishing between male or female. Each feature was visualised and ranked on its ability to categorise the target label. Figure 1 show the boxen plot distribution of the informative features and **Table 1 shows how useful all the features are for sex classification, based on how dissimilar the distributions in Fig. 1 looked when split between male and female**.

2.3 Transformations

Following the findings from Section 2.1 and Section 2.2, the following transformations will be applied on the dataset:

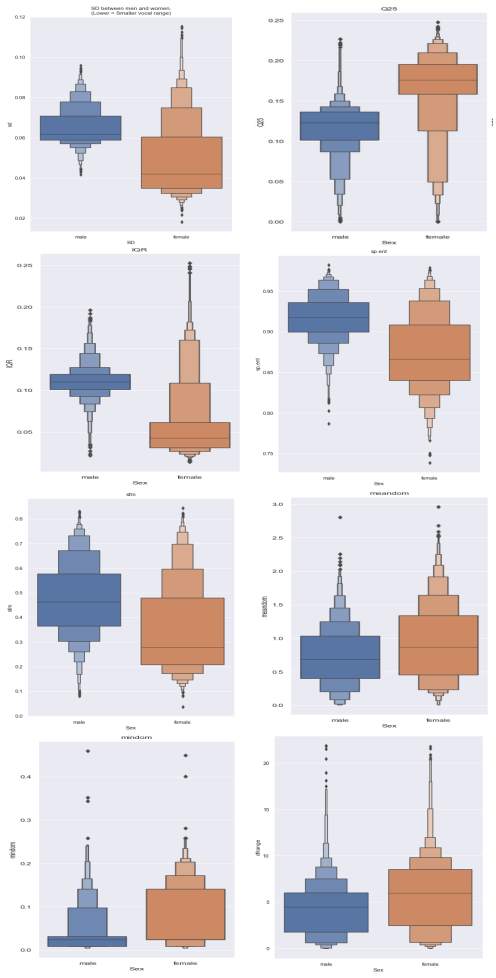

Figure 1: The boxenplot distributions of the features selected as being informative. In Left-to-Right, Top-to- Bottom Order: SD, Q25, IQR, sp.ent, sfm, meandom, mindom, dfrange

Table 1: Table showing how informative a feature is in predicted the sex of the voice.

Features	Information
`meanfreq`	LOW
`median`	LOW
`mode`	LOW
`centroid`	NULL
`sd`	HIGH
`skew`	LOW
`kurt`	LOW
`Q25`	HIGH
`Q75`	LOW
`IQR`	HIGH
`minfun`	LOW
`maxfun`	LOW
`meanfun`	HIGH
`mindom`	MID
`maxdom`	LOW
`meandom`	LOW
`sp.ent`	HIGH
`sfm`	MID
`dfrange`	MID
`modindx`	LOW

1. Numerically encode `label`. (Label Encoding)
2. Remove `centroid`
3. Standardise all features by removing the mean and scaling to unit variance (Standard Scaler)
4. Randomley Shuffle dataset
5. Remove *LOW* features (See Table 1)

3. ARTIFICIAL NEURAL NETWORKS

The ANN was constructed using PyTorch, and optimised using ADAM. Early-stopping was implemented by evaluating the performance under the test set every set epochs. It seems that ANNs are suitable for sex prediction on the dataset. During this experiment, hyper-parameters such as the hidden layers and the activation function were changed. As Table 2 depicts, ANNs managed to receive consistently good results under a number of different hyper-parameters, even performing well with no hidden layers. Although the difference between the variations is minimal (with one exception), using an ANN with 1 hidden layer of 6 neurons on the filtered dataset produced the best results, with an accuracy of 99.1%. However training was distorted when using 2 hidden layers of 1 neuron each, producing the only bad result in the model with an accuracy of 49.4%. Because of the small size of the dataset, the numbers varied ever so slightly when re-training the model, and when reproducing the results, a different variation might give the best results.

4. SUPPORT VECTOR MACHINES

SVMs produced good results on the preprocessed data, achieving accuracies above 95%. The best kernel to use was RBF

Variation	Accuracy	Precision	Recall	F1
0H + ReLU	0.978	0.971	0.984	0.977
1H (8) + ReLU	0.986	0.987	0.984	**0.986**
1H (6) + ReLU	**0.991**	**0.994**	**0.987**	0.990
1H (4) + ReLU	0.975	0.971	0.977	0.974
1H (2) + ReLU	0.978	0.971	0.984	0.977
1H (1) + ReLU	0.976	0.968	0.984	0.976
2H (8x8) + ReLU	0.986	0.987	0.984	**0.986**
2H (8x6) + ReLU	0.986	0.984	0.987	**0.986**
2H (6x6) + ReLU	0.976	0.968	0.984	0.976
1H (1x1) + ReLU	**0.494**	**1.000**	**0.494**	**0.661**
0H + L_ReLU	0.978	0.971	0.984	0.977
0H + Sigmoid	0.978	0.971	0.984	0.977
0H + TanH	0.978	0.971	0.984	0.977
1H (8) + ReLU	0.986	0.987	0.984	**0.986**
1H (8) + L_ReLU	0.984	0.987	0.981	0.984
1H (8) + Sigmoid	0.984	0.984	0.984	0.984
1H (8) + TanH	0.981	0.978	0.984	0.981
2H (8x4) + ReLU	0.981	0.981	0.981	0.981
2H (8x4) + L_ReLU	0.981	0.984	0.978	0.981
2H (8x4) + Sigmoid	0.976	0.974	0.978	0.976
2H (8x4) + TanH	0.978	0.978	0.978	0.978

Table 2: Evaluating ANNs under different hyper-parameter conditions. XH denotes the number of hidden layers used. Figures of interest are emphasised.

Kernel	Accuracy	Precision	Recall	F1
Linear	0.978	0.971	0.984	0.977
Cubic	0.964	0.939	**0.987**	0.962
RBF	**0.979**	**0.981**	0.978	**0.979**
Sigmoid	0.826	0.843	0.812	0.828

Table 3: Evaluation of SVM using different kernels.

achieving an accuracy of 97%, whilst the worst one was Sigmoid, with an accuracy of 82% (See Table 3). When classifying with polynomial kernels, different degrees for the polynomial were chosen. As the reader can appreciate in Fig. 2, the best polynomial degree was cubic, with the performance tapering off as the degrees increase. Moreover, the regularisation parameter `c` of the SVM was also experimented with. For the given dataset, the best value of `c` was between 3 and 6 (See Fig. 3).

5. KNN

KNN consistently provided high values across all metrics as can be seen in Table 4. The classifier performed the best when using Euclidean Distance. Moreover, as the power parameter for the Minkowski distance grows, the performance tapers off and ultimately converges as can be seen by Fig. 4. Moreover, the best values for `k` when using Euclidean Distance are 2 or 4 as can be seen by Fig. 5.

6. DECISION TREE

Decision Trees performed well on the data across all differ-

Metric	Value
Accuracy	0.984
Precision	0.974
Recall	0.993
F1	0.984

Table 4: Performance of KNN classifier.

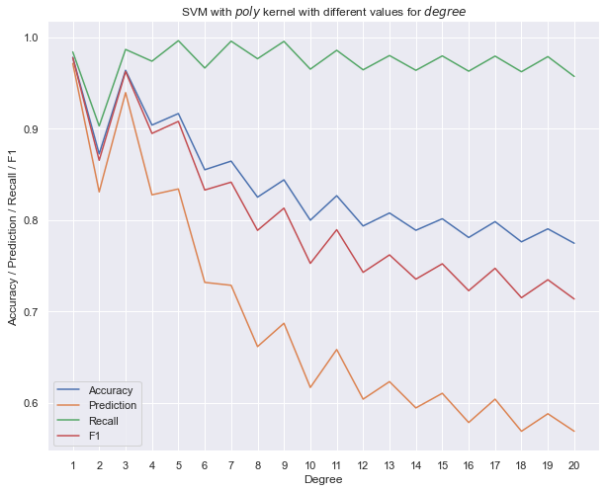

Figure 2: Graph showing how the performance of the SVM changes with different degrees for a polynomial kernel.

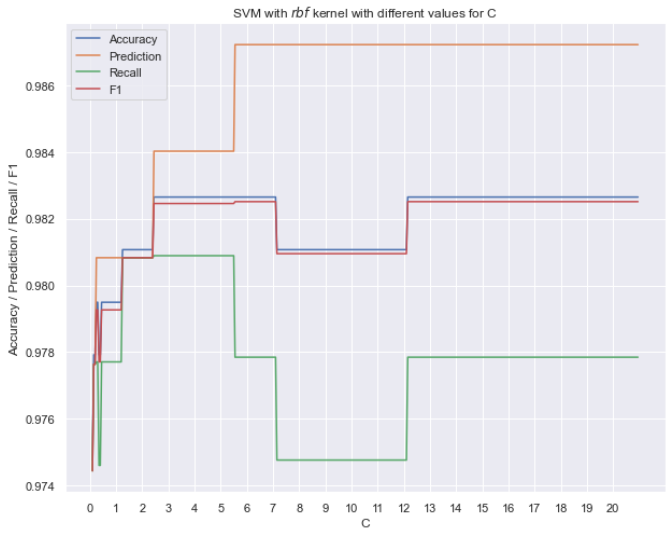

Figure 3: Graph showing how the performance of the SVM with an `rbf` kernel changes with a varying regularisation parameter `c`.

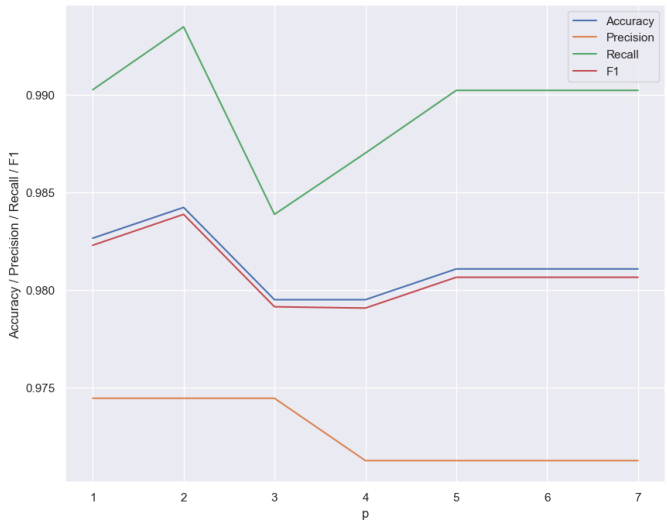

Figure 4: Varying the power of the Minkowski distance p in a KNN classifier. When $p = 1$, Manhattan Distance is used. When $p = 2$, Euclidean Distance is used.

Figure 5: Varying the hyperparameter `k` in K-Nearest Neighbours

Splitter	MM	MF	FM	FF
Best	314	7	10	303
Random	309	12	12	301

Table 5: Confusion Matrix Tree Classifier using different split selection criteria.

Quality of Split Heuristic	MM	MF	FM	FF
Gini Impurity	314	7	10	303
Shannon Information Gain	311	10	11	302

Table 6: Confusion Matrix of a Decision Tree Classifier using different split evaluation heuristic functions.

ent combinations of hyper-parameters. As can be seen by Table 5, when choosing the best split over a random split, 5 additional instances were not misclassified. Moreover, when using the Gini Impurity over Shannon Information Gain in order to determine the quality of a split, an additional 4 instances were correctly classified (See Table 6). Variables such as the maximum depth d, and the minimum number of samples needed to perform a split m were varied. As can be seen by both Figure 6 and 7, the best values for d and m are around 5 and 8 respectively.

7. LOGISTIC REGRESSION

Logistic Regression gave consistently good accuracy 97% across many different combinations of hyper-parameters. We experimented with creating a 3D surface-plot showing how different variations of the hyper-parameters `tol` and `c` affected the accuracy. `tol` is the tolerance for the stopping criteria and `c` is the regularisation parameter similar to SVMs (See 4). However the hyper-parameters failed to increase or decrease the accuracy of the model. The accuracy of the model for $(tol, c) \in \mathbb{R} \times \mathbb{R}$ remained fixed at 97.7971%. The option to add a bias/intercept to the linear decision hyperplane was also experimented on. The performance increase was minimal when enabled (See Table 7). This might be because the data was normalised to float around 0, and hence the linear decision surface would not need to be 'elevated' to 'reach the data'.

8. EVALUATION

As can be seen by Table 8, the best algorithm for sex prediction from this dataset is ANN with an accuracy of 99.1%. The neural network was run with 1 hidden layer of 6 nodes with ReLU as an activation function as can bee seen by Table 2. Neural Networks are very scalable to different types of problems, mainly because the architecture can be enlarged the more complex the problem is. However. it is important to mention that all algorithm performed extremely good on the dataset with an average accuracy around 97%. More-

Fit Intercept	MM	MF	FM	FF
True	316	5	9	304
False	316	5	13	300

Table 7: Confusion Matrix of a Logistic Regression Classifier with and without an intercept. Enabling an intercept allows the model to correctly classify 4 additional females than without.

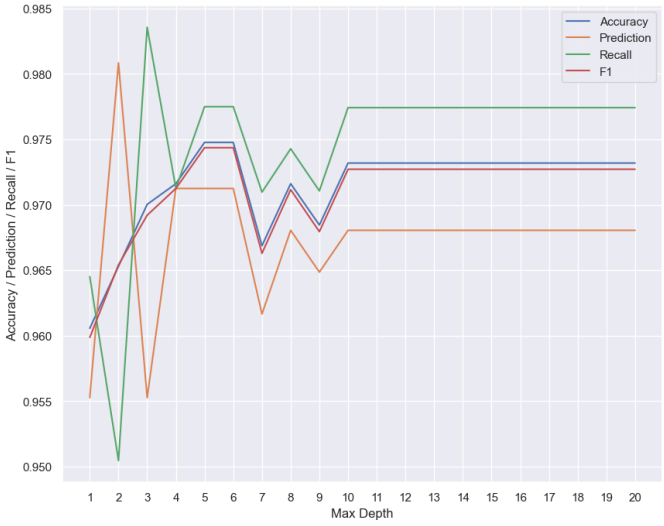

Figure 6: Graph showing the performance of the Decision Tree with varying values for the maximum depth of a tree.

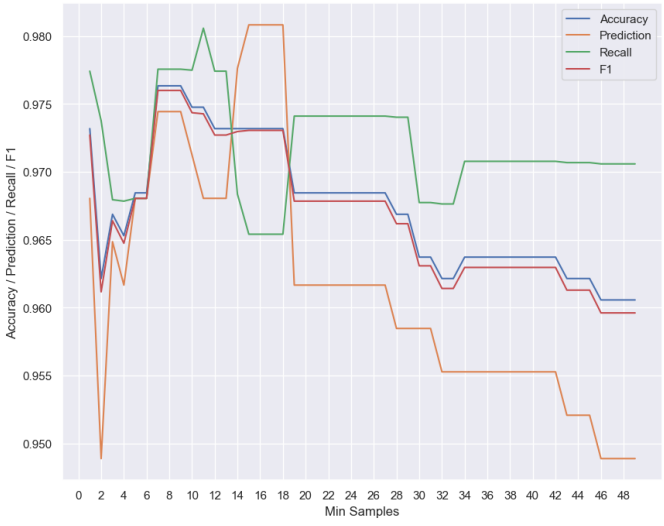

Figure 7: Graph showing the performance of the Decision Tree with varying values for the minimum number of samples needed in order to perform a split.

Method	Accuracy	Hyperparameters
ANN	**0.991**	1 Hidden (6 nodes), ReLU
SVM	0.979	`C=4`, RBF Kernel
KNN	0.984	Euclidean Distance, `k=4`
DT	0.976	Max Depth=5, MinSamplesLeaf=8
LR	0.978	Fit Intercept/Add Bias

Table 8: Highest accuracies of the different algorithms across different hyperparameter variations.

Item	Completed
Implemented artificial neural network	Yes
Implemented support vector machine	Yes
Implemented k-means clustering	Yes
Implemented decision tree learning	Yes
Implemented logistic regression	Yes
Evaluated artificial neural network	Yes
Evaluated support vector machine	Yes
Evaluated k-means clustering	Yes
Evaluated decision tree learning	Yes
Evaluated logistic regression	Yes
Overall comparison of methods and discussion	Yes

Table 9: Statement of Completion

over, although hyperparameter tweaking did manage to improve the performance of the algorithms, without doing so would still have resulted in a good classifier as almost all accuracies hovered around 95% regardless of the hyperparameters. In fact, most models were only able to be improved by at most 2% more. Due to the small size of the dataset, all methods trained on the dataset in negligible times. Scaling to real world problems, ANNs might require more computationally intensive workloads in order to achieve good performance.

9. CONCLUSIONS

In conclusion, the results of this project demonstrate that machine learning techniques, specifically ANNs, SVM, KNN, decision trees, and logistic regression, can be effectively applied to the problem of predicting the sex of a speaker based on a given dataset. The data was preprocessed and out of the 20 features only 8 were kept, and various hyperparameters for the machine learning algorithms were tested. An interesting observation to note is that there was a duplicate feature in the data which would have added an extra layer of complexity to the machine learning models if it had not been detected and removed. The best results were obtained using an ANN with 1 hidden layer of 6 neurons, achieving an accuracy of 99.1%. Another interesting observation is that an ANN with one hidden layer with one neuron performed well on the dataset, but adding an addition hidden layer with one neuron made the model flop. Overall, all of the algorithms performed extremely well. This project showed that sex prediction from such a dataset is not only possible but accurate and reliable. Moreover, it is possible to scale this project into real-world applications as this proof-of-concept demonstrated impeccable performance.

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own, of work of another person, whether or not such work has been published, and as may be further elaborated in Faculty or University guidelines" (University Assessment Regulations, 2009, Regulation 39 (b)(i), University of Malta).

We, the undersigned, declare that the assignment submitted is our work, except where acknowledged and referenced.

We understand that the penalties for committing a breach of the regulations include loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

(N. B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

Gabriel Hili	
Student Name	Signature
Student Name	Signature
Student Name	Signature

ICS3206	**ICS3206 - Course Project**
Course Code	Title of work submitted

01/06/23

Date