Age Detection

Francesco Mastrosimone, Salvatore Nocita
Politecnico di Torino
Student ids: s348159, s346378
emails: s348159@studenti.polito.it, s346378@studenti.polito.it

***Abstract*—This work addresses the problem of age estimation through vocal features extracted from audio recordings, as part of the winter project for the course "Data Science Lab: Process and Methods." Using two provided datasets, one for training and one for evaluation, we analyzed feature distributions, selected the most significant ones, and carried out a meticulous preprocessing phase. The pipeline included logarithmic transformations, normalization, and subsequent selection of relevant vocal features, such as MFCC coefficients and mel-spectrograms. After comparing several regression models, including Random Forest, SVR, and Ridge, the latter proved to be the most effective. Through the tuning of α and the inclusion of second-degree polynomial features, the model achieved optimal performance in terms of RMSE and r^2, demonstrating an excellent balance between complexity and generalization.**

I. Problem Overview

The competition proposed for the winter project of the "Data Science Lab: Process and Methods" course involves tackling an **age estimation problem**. To address this task, two datasets were provided: the first, named *development.csv*, contains various vocal features previously extracted from numerous audio files and the target age of the speakers. This dataset served as the starting point to build and train regression models. The second, called *evaluation.csv*, lacks labeled ages and is the dataset on which the trained models were applied for age recognition.

Both datasets contain detailed information about the recorded voice signals. The recordings were sampled at a frequency of 22.050 kHz, ensuring good temporal resolution for vocal feature analysis. Each observation includes a variety of features extracted from the audio signals, such as:

- **Voice-specific features**, including *mean, maximum, and minimum pitch*, which represent the fundamental frequency of the signal, and measures such as *jitter* and *shimmer*, describing temporal and amplitude variations of the signal.
- **Speech-related metrics**, like the *number of words spoken*, *number of characters*, and *silence duration*, which provide insights into speech rhythm and content.

These data form the basis for training machine learning models aimed at predicting the exact age (represented as a continuous float value) of the speakers.

The applications of this type of problem are numerous, including:

- Voice assistance systems and artificial intelligence, where the user's estimated age can enhance personalized interaction.
- Demographic control in market research, enabling more detailed audience analysis.
- Security systems, to validate user identities or restrict access to certain features.

Based on the *age distribution* in the development dataset (Figure 1), we observe that the data are not balanced. Most speakers fall into the age range of 20 to 30 years, with progressively fewer observations for higher ages. This imbalance could affect the ability of machine learning models to generalize to underrepresented age groups. Additionally, if the features are strongly correlated with age, the lack of uniform distribution across classes could hinder the model's ability to recognize significant differences among the less-represented age groups.

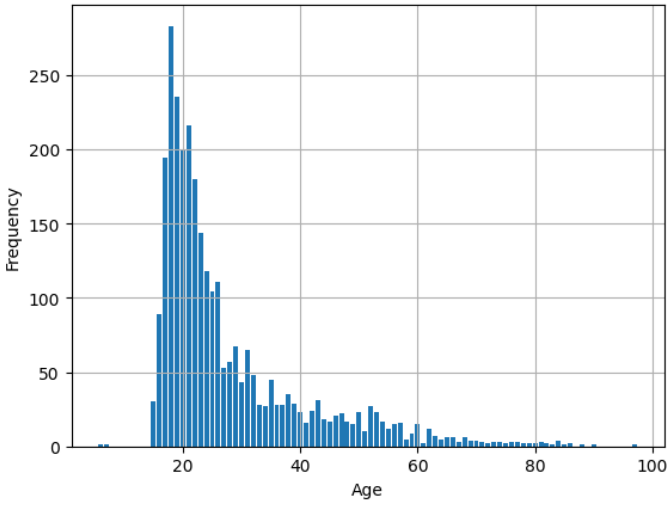

Fig. 1. Age distribution

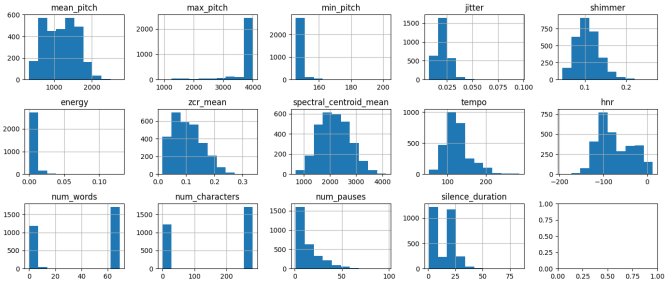

Fig. 2. Feature distribution

The distributions of numerical features (Figure 2) reveal several interesting trends:

Voice-Specific Features:

- *Mean pitch* exhibits a relatively uniform distribution, highlighting a wide variability among speakers. In contrast, *max pitch* and *min pitch* are more concentrated within specific intervals, with extreme values less frequent.
- *Jitter* and *shimmer*, metrics describing frequency and amplitude variations, show asymmetric distributions with most values close to zero, indicating vocal signal stability for most speakers.

Speech-Related Metrics:

- *Number of words* and *number of characters* display a bimodal distribution, suggesting two distinct groups of recordings: some with very brief content and others more verbose.
- *Silence duration* Exhibits significant variability, largely due to the varying durations of the audio messages, highlighting the importance of normalizing this attribute.

Spectral Metrics: The distribution of *spectral centroid mean*, representing signal "brightness," is slightly asymmetric, with a well-defined peak around a mean value and a tail toward higher frequencies. *HNR* (Harmonic-to-Noise Ratio) is predominantly distributed toward negative values, indicating a prevalence of natural vocal signals over noise.

These distributions highlight the diversity of available data and provide useful indications for preprocessing, such as feature normalization and transformation.

II. Proposed Approach

Our approach was structured into seven phases:

- Analysis of selected feature distributions
- Elimination of irrelevant features
- Further feature selection
- Extraction of vocal features from audio
- Creation of the final dataset
- Testing of various regression models
- Tuning the selected model and final predictions

Data Preprocessing

After ensuring that the dataset contained no null values and that the attribute values all belonged to the same domain, we began by filtering the features provided in the *development dataset* to eliminate those irrelevant to the task.

The first feature analyzed was *ethnicity*. After analyzing the distribution of ethnicities in the dataset, we observed many underrepresented ethnic groups and a highly unbalanced age distribution within them, which caused the model to unfairly exploit the dataset's structure. This hypothesis was further confirmed by calculating *cosine similarity* of centroids between pairs of dinstinct ethnicities. This calculation always yielded to values close to 1, indicating a strong similarity between categories.

Additionally, the overall analysis revealed that, despite the presence of approximately 160 different ethnicities in the dataset, this feature proved to be among the least relevant in building a preliminary model based on the *RandomForestRegressor*. For this reason, we decided to exclude the *ethnicity attribute* from the predictive model, as the speaker's ethnicity did not appear to be a determining factor in differentiating their vocal characteristics.

Upon examining the distributions of the *number of words* and *number of characters* attributes, it became evident that the two distributions are highly similar. Listening to a sample of audio files, we discovered that groups of speakers were reciting the same text, leading to identical values for these attributes. Furthermore, many audio files reported both *number of words* and *number of characters* as zero, even when speech was present, suggesting issues with the transcription. Based on these observations, we decided to eliminate these features from the dataset.

Moreover we decided to eliminate the features *minimum pitch* and *maximum pitch* for two main reasons. First, both variables exhibit low variability within the dataset, making them uninformative for the model. Additionally, the presence of *mean pitch* renders these two features redundant, as the mean pitch effectively captures the central information related to the fundamental frequency of speech, making the minimum and maximum values unnecessary for improving model performance. This decision helped reduce informational noise, resulting in a more streamlined and effective dataset.

After this phase, we transformed and normalized several features:

- The categorical attribute gender was encoded into two binary variables: gender male and gender female, enabling a direct numerical representation useful for machine learning models.
- The *silence duration* feature was normalized relative to the total audio duration and the *duration* consequently has been added to the dataset, due to its relevance in the RandomForestRegressor model.
- Finally, the attributes *energy, mean of zcr, tempo, number of pauses, jitter and shimmer* were transformed to a **logarithmic scale**. While their linear scale showed a wide range of values skewed to the right, the logarithmic scale highlighted a more meaningful range, making the data more interpretable. This approach aligns with the fact that human perception of sound intensity is logarithmic by nature.

Feature Correlation: The Feature Correlation Matrix is presented below, offering interesting insights. The analysis of the feature correlation matrix revealed that some variables exhibit strong mutual correlations. To evaluate the impact of these correlations on model performance, several tests were conducted using a RandomForestRegressor model, where, each time, one of the two highly correlated features were progressively removed. The results showed that eliminating such variables did not improve the model's performance. On the contrary, the best performance was achieved by retaining

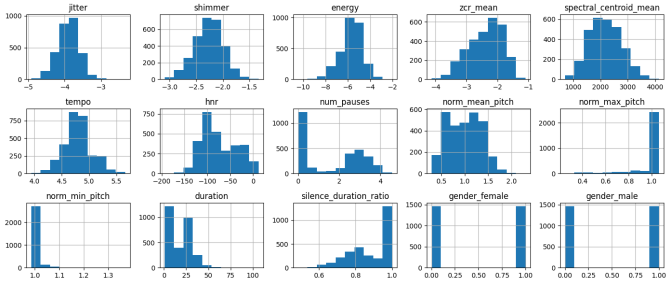

Fig. 3. Feature distribution after transformation

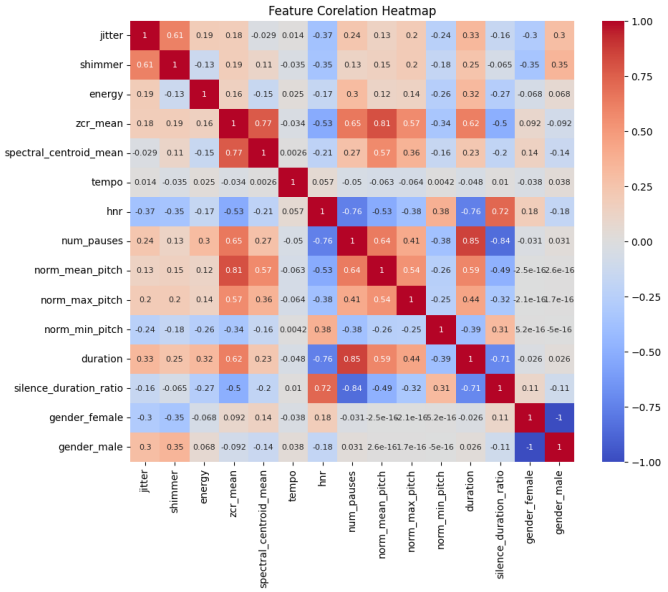

Fig. 4. Feature Correlation Matrix

all the features in the dataset. Therefore, it was decided to keep the entire set of attributes for the subsequent stages.

Feature Extraction: The extraction of the best parametric representation of acoustic signals is an important task to produce a better recognition performance. The efficiency of this phase is important for the next phase since it affects its behavior [1]. The following features were selected thanks to the support of the Python library **Librosa** [2]:

1) ***MFCC coefficients*** (Mel-Frequency Cepstral Coefficients), which capture fundamental information about timbre and sound quality, often influenced by physical characteristics associated with age. 60 coefficients were extracted for this feature, both for mean and standard deviation.
2) The ***onset per second*** metric was included for its utility in representing the density of acoustic events, which may vary based on the rhythm and speaking style often correlated with age.
3) For ***Mel spectrogram*** extraction, normalization was applied to 7 seconds for each audio file (randomly chosen within the audio file) to ensure comparable samples of equal length [4]. The spectrogram was divided into 64 bands and the mean was computed to highlight the energy distribution across frequencies—an aspect that can reflect age-related vocal changes.
4) The ***variance of the spectral centroid*** was chosen to describe the variability of spectral balance over time, which can indicate instability or changes in vocal quality.
5) Finally, metrics related to ***fundamental frequency*** (median, standard deviation, and 95th and 5th percentiles) were included to characterize pitch and signal stability, elements that significantly vary with age.

This targeted selection ensures that the extracted features align closely with the analysis objectives, providing an effective representation for age detection.

Supported by the metrics provided by the RandomForestRegressor, we evaluated various combinations for different random state values, combining MFCC coefficients and Mel coefficients (means and variances)thanks to the use of *ParameterGrid*. The best configuration consisted of 13 mean MFCC coefficients and the first 20 out of 64 coefficients from the Mel spectrogram.

The first 20 coefficients of the Mel spectrogram, at a sampling rate of $22,050$ Hz, capture most of the relevant information for human voice, making it possible to ignore the remaining 44 coefficients without losing significant information for many vocal applications. This approach improves efficiency and performance by eliminating superfluous features and focusing on the most significant components of the signal, ensuring optimal informational contribution for the age estimation task.

Model Selection

At the model selection stage, we decided to conduct the initial tests using the **Random Forest Regressor**, an ensemble learning model particularly suitable for datasets with numerous features and potential nonlinear relationships between variables. The choice of this model was motivated by its ability to handle complex and non-normalized datasets, as well as its robustness to irrelevant variables, thanks to its intrinsic feature selection mechanism. Additionally, Random Forest proved useful during the data processing phase for identifying the relative importance of variables, providing valuable insights into the most significant features. The results obtained with this model were evaluated in terms of $RMSE$ and r^2 metrics to assess the level of accuracy achieved.

Subsequently, we considered the **Support Vector Regressor** (*SVR*). This model is particularly suitable for datasets with a relatively small number of features, due to its ability to capture complex relationships between variables through the use of nonlinear kernels. Given that our dataset includes vocal characteristics that may exhibit nonlinear relationships with age, *SVR* emerged as a valid option. However, *SVR* is sensitive to parameter selection and requires proper hyperparameter tuning (e.g., kernel, C, and ϵ), which has not yet been addressed in this preliminary phase.

Finally, we evaluated the **Ridge regression model**, a regularized form of linear regression. Ridge regression is particularly well-suited for datasets with many variables, especially when these exhibit multicollinearity, as highlighted by our correlation matrix analysis. The $L2$ regularization term helps penalize parameter coefficients, preventing overfitting issues and improving the model's generalization capabilities. Moreover, Ridge is especially efficient for medium-sized datasets, such as ours, with approximately 50 variables, and performs well when all features contribute in a balanced way to the final output.

We tested each model without hyperparameter tuning, repeating the training process over 50 different *random state values*, obtaining the following results:

- **RandomForestRegressor**: $RMSE$ of 10.594653 and r^2 of 0.3001754.
- **SVR**: $RMSE$ of 10.784653 and r^2 of 0.2973002.
- **Ridge**: $RMSE$ of 10.384653 and r^2 of 0.3490263.

The results led us to select the Ridge regression model for the following reasons:

1) *Handling multicollinearity*: Ridge regression is particularly effective in mitigating the negative effects of multicollinearity among variables, as observed in our correlation matrix.
2) *Balance between bias and variance*: Thanks to the $L2$ regularization term, Ridge effectively balances the trade-off between bias and variance, demonstrating better generalization capabilities compared to the other models considered [3].
3) *Superior performance*: Ridge achieved the best RMSE among the tested models, indicating higher prediction accuracy, and a significantly higher r^2 value, suggesting that the model better explains data variability compared to Random Forest and SVR.
4) *Computational efficiency*: Ridge is computationally less intensive than Random Forest and SVR, making it a more practical choice for rapid iterations and subsequent optimizations.

Hyperparameters Tuning

Tuning the Ridge regression model, the only parameter to optimize is the value of α. This parameter represents the $L2$ regularization coefficient in the cost function, balancing bias and variance.

The objective function minimized by Ridge is defined as:

$$\text{MIN } \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} w_j^2$$

A higher value of α forces the coefficients to be smaller, reducing the risk of overfitting. Conversely, very low values of α reduce regularization, increasing the risk of overfitting when the dataset contains noise or collinearity among variables.

We performed a search starting with initial values in a logarithmic scale: $0.01, 0.1, 1, 10, 100, 1000$. After analyzing the model's performance for each value of α in terms of $RMSE$ and r^2, we progressively narrowed down the range of values to refine the choice of the optimal parameter.

In the tuning pipeline, we integrated **PolynomialFeatures** to introduce nonlinear relationships into the data and **StandardScaler** to normalize the variables, ensuring uniform contributions from all features.

The tuning process identified $\alpha = 178$ as the value that enabled the Ridge model to achieve the best performance. Additionally, the inclusion of **second-degree polynomial features**, significantly contributed to improving the model's performance. These results ensure an adequate trade-off between bias and variance, reducing the risk of overfitting without excessively penalizing the model's ability to fit the training data.

III. Results

In the context of the competition, our Ridge regression model, optimized through α tuning and the inclusion of second-degree polynomial features, achieved a final score of RMSE = 9.463 in the overall leaderboard. This result highlights the effectiveness of the adopted pipeline, which successfully balanced model complexity and generalization ability, demonstrating excellent predictive performance compared to the competition standards.

IV. Discussion

However, we acknowledge that there is room for improvement in our work, particularly through the exploration of advanced techniques for dataset balancing or the implementation of more complex models capable of more accurately capturing the nonlinear relationships between vocal features and age.

V. Conclusion

In this project, we developed an approach for age estimation based on vocal data, integrating a complete pipeline for preprocessing, feature selection, and modeling. The choice of the Ridge model, combined with the tuning of α and the use of second-degree polynomial features, proved effective in balancing bias and variance while capturing nonlinear relationships within the data.

Despite the strong performance achieved, certain limitations remain evident, particularly the model's difficulty in generalizing to age groups that are underrepresented in the dataset. To overcome these limitations, future work could explore dataset balancing techniques or the use of more complex models, such as deep neural networks, capable of capturing even more intricate relationships within the data. Overall, the project demonstrates how a well-crafted combination of analytical and modeling techniques can effectively address complex problems in the field of age estimation based on vocal features.

References

[1] L. Muda, M. Begam, and C. Elamvazuthi, "Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques," *Journal of Computing*, vol. 2, no. 3, pp. 138–143, 2010.

[2] Librosa Development Team, "Librosa: Python library for audio and music analysis," 2023. [Online]. Available: https://librosa.org. [Accessed: Jan. 12, 2025].

[3] K. Hechmi, T. N. Trong, V. Hautamäki, and T. Kinnunen, "VoxCeleb enrichment for age and gender recognition," *arXiv preprint arXiv:2109.13510v2*, 2021. [Online]. Available: https://arxiv.org/abs/2109.13510. [Accessed: Jan. 17, 2025].

[4] V. S. Kone, P. Jadhav, A. Anagal, U. Kulkarni, and S. Anegundi, "Voice-based Gender and Age Recognition System," in *2023 Int. Conf. on Advancement in Computation and Computer Technologies (InCACCT)*, 2023, pp. 74–79. DOI: 10.1109/InCACCT57535.2023.10141801.

[5] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer program]," 2001. [Online]. Available: http://www.praat.org. [Accessed: Jan. 20, 2025].