Age Prediction Dataset

HTTPS://WWW.KAGGLE.COM/DATASETS/ABDULLAH0A/HUMAN-AGE-PREDICTION-SYNTHETIC-DATASET

Dataset Information

General Description

This dataset comprises synthetic data curated for the purpose of age prediction, using a comprehensive suite of health and lifestyle metrics. It features 3,000 records, each with 25 unique attributes, to offer insights into the relationship between various health and lifestyle indicators and age. This dataset is ideally suited for developing predictive models in healthcare and wellness contexts.

Features and Characteristics

Feature/Characteristic	Units	Description
Height	cm	The height of the individual
Weight	kg	The weight of the individual
Blood Pressure	s/d	Blood pressure (systolic/diastolic) in mmHg
Cholesterol Level	mg/dL	Cholesterol level
Body Mass Index (BMI)	None	Approx. BMI calculated using height and weight
Blood Glucose Level	mg/dL	Blood glucose level of the individual
Bone Density	g/cm^2	Bone density
Vision Sharpness	None	Scale from 0 (blurry) to 100 (perfect)
Hearing Ability	dB	Hearing ability in decibels
Physical Activity Level	None	Categorized as Low, Moderate, or High
Smoking Status	None	Categorized as Never, Former, and Current
Alcohol Consumption	None	Categorized as None, Occasional or Frequent
Diet	None	Categorized as Balanced, High Protein, Low Carb, etc.
Chronic Diseases	None	What, if any, chronic diseases
Medication Use	None	Categorized as None, Regular, Occasional
Family History	None	Presence of family history of age-related conditions
Cognitive Function	None	Self-reported scale from 0 (poor) to 100 (excellent)
Mental Health Status	None	Self-reported scale from 0 (poor) to 100 (excellent)
Sleep Patterns	Hours	Average amount of sleep per night
Stress Levels	None	Self-reported scale from 0 (poor) to 100 (high)
Pollution Exposure	None	Exposure to pollution measured in arbitrary units
Sun Exposure	Hours	Average sun exposure per week
Education Level	None	Highest level of education attained
Income Level	USD	Annual income
Gender	None	Male or female
Age	Years	Target variable representing age of the individual

Issues with these features come down to the subjectivity of some of the data collected. For example, the definition of a moderate physical activity level will likely be different between two individuals.

However, the results may still be worth exploring and better defined in a follow-up study if determined to be viable.

Potential Uses

Finding a correlation between certain factors and an individual's age could lead to finding what affects the aging process in humans. For example, if there is a correlation between sleep patterns and age, we could look further into what affects sleep patterns and perform a follow-up experiment to see what happens if we fix sleep patterns in individuals. By finding enough of these correlations, we may be able to understand how we can best increase life expectancy with current technology.

Kaggle Contributions

Contribution 1

https://www.kaggle.com/code/dalraejin29/age-prediction-using-ml-approach

Considered Learning Tasks

There are two learning tasks mentioned:

1. Predict the age or age group of a human using regression for age and classification for age group
2. Evaluate the most significant factors that contribute to predictions

Data Analysis and Preparation

To prepare the data, null values were replaced with "none" as was the intent of the dataset. Next, the blood pressure feature is in units systolic/diastolic, so they are separated at the "/" symbol and made into two separate columns and another age group feature was added for the classification model. An encoder to make it more compatible with machine learning models.

The data was also shown to resemble a bell curve or have relatively even distribution across most features and a correlation matrix was shown.

Learning Models

1. Multilinear Regression
    a. Features with high correlation and variance inflation factor values were also removed from the set. A multiple linear regression model was fitted and assessed the p-values of coefficients. High p-values were removed. After this process, only bone density and smoking status were significant enough to predict age.
    b. After all this, the model had an R-Squared score of 86.4%, which is generally considered strong.
2. Classification using SVM and Decision Tree
    a. Decision Tree resulted in 62% accuracy while SVM resulted in 73% accuracy
    b. The SVM was decent at predicting adults, senior, and elderly but was not good at predicting young and middle-aged age groups.

Contribution 2

https://www.kaggle.com/code/abdullah0a/precisionforest-advanced-model-tuning-and-optimiz

This dataset had many spelling errors, but it was one of the only other ones that described what they were doing, why, and had some results.

Considered Learning Tasks

Predicting the age of a human and determining the most important features.

Data Analysis and Preparation

Like Contribution 1, systolic and diastolic were separated, null values replaced with 0, and encoding was used to make the data easier for the models to work with.

Learning Models

To calculate feature importance, an XGBoost model was used and the key features influencing the model's results were identified. This was combined with RandomForest to improve accuracy. Overall, this resulted in a mean squared error of 50.205.

Performance of Each Model

Contribution 1 had a much more thorough data preparation process than Contribution 2 that resulted in removing many more features and had the best performing model, multilinear regression. Contribution 2 did not filter the data as much, but it did combine multiple models with hopes of increasing accuracy yet decided to include mean squared error instead of accuracy. Unfortunately, in my opinion, the mean squared error being used in this case does not provide a significant metric as it does not relate to any other models.

Anticipated Contributions

Proposed Changes

Something that seemed to work well for Contribution 1 was meticulously pruning the data for better predictions, however, they also bring up a good point about the ease of measuring some of the results. For example, bone density was a significant predictor of age but would be a much more invasive process to measure on a human as opposed to eyesight. To that end, I would like to further explore even highly correlated features with hopes of discovering something else from this dataset. One other change I would like to do is introduce k-fold validation. With the bell curve distribution, there is a risk of the test-train set having wild variations of each feature, and I believe using k-fold validation will decrease the chances of that distribution having a significant effect on the results of the model.

Expected Results

I expect that bone density and vision will be highly correlated with age since I believe both get worse with age after reaching a certain age. I also believe that there will be societal correlations, with smokers and drinkers being more prevalent among the older generations, it might be found in the data that there is a correlation between those habits and age. This will have to be something I consider, as society has a

strong effect on how we lead our lives, certain habits can be correlated with age that have no significant biological impacts.