Human Age Prediction Based on Health and Lifestyle Factors

CIS 9660- Data Mining for Business Analytics - Prof. Chaoqun Deng
Zicklin School Of Business, Baruch College, CUNY

Team 1: Darshi Shah, Dhruv Sharma, Krishi Shah, Nikita Gautam, Vrinda Arora

Introduction:
Age is more than a number - it's a reflection of health, lifestyle, and biological processes. In this project, we use machine learning to predict a person's age based on various health and lifestyle factors. Accurate age prediction has potential applications in personalized healthcare, early risk detection, and wellness planning. Our analysis is based on a synthetic Kaggle dataset containing 3,000 records and 26 features including numerical, categorical, and multi-valued attributes.

Motivation for the research:

This project aims to analyze how various physiological and lifestyle factors contribute to aging such as mental health (stress, cognitive function) and physical health (activity levels, vision, hearing). It explores and quantifies the relationship between observable features (e.g., blood pressure, hearing ability) and a person's actual age. Predicting age using measurable health and behaviour indicators enables earlier detection of health risks and provides a more objective view of aging. Biological age often differs from chronological age and our approach offers deeper insight into this gap using data-driven techniques.

Additionally, many health metrics that change with age can be easily recorded, making this analysis scalable and practical in real-world applications. In relation, as personalized healthcare and wellness platforms grow, age prediction models can enhance their effectiveness by supporting tailored health interventions. Other stakeholders include insurers and public health systems which can use age predictions to improve risk assessments, pricing fairness, and demographic health planning. This ultimately helps consumers to manage their costs as well in the rising rate environment.With rising healthcare costs and aging populations, accurate biological age estimation has significant value for clinical and policy decision-making. Healthcare law is one area which needs better advocacy especially in countries with no overarching public system. Lastly, this study also examines the predictive power of biological markers (e.g., blood pressure, bone density) versus lifestyle attributes (e.g., smoking, activity level), helping inform smarter strategies in healthcare and product design for medicine as producers can know their consumers better.

Data Description & Variable Introduction:
The dataset was sourced from Kaggle and includes **3,000 observations** across **26 variables**, combining both numerical and categorical attributes:

Numerical Variables (14): Height, Weight, BMI, Blood Pressure (split into Systolic and Diastolic), Cholesterol, Blood Glucose, Bone Density, Vision Sharpness, Hearing Ability, Cognitive Function, Stress Levels, Pollution Exposure, Sun Exposure, and Age (target).

Categorical Variables (12): Gender, Physical Activity Level, Smoking Status, Alcohol Consumption, Diet, Chronic Diseases, Medication Use, Family History, Mental Health, Sleep Patterns, Education Level, Income Level.

Target Variable: Age ranges from 18 to 89.

	min	max
Height (cm)	141.13	198.11
Weight (kg)	32.54	123.60
Blood Pressure (s/d)	NaN	NaN
Cholesterol Level (mg/dL)	148.81	331.30
BMI	12.05	43.33
Blood Glucose Level (mg/dL)	69.87	185.74
Bone Density (g/cm²)	-0.22	2.00
Vision Sharpness	0.20	1.06
Hearing Ability (dB)	0.00	94.00
Cognitive Function	30.38	106.48
Stress Levels	1.00	10.00
Pollution Exposure	0.01	10.00
Sun Exposure	0.00	11.99
Age (years)	18.00	89.00

	Unique Values
Gender	2
Physical Activity Level	3
Smoking Status	3
Alcohol Consumption	2
Diet	4
Chronic Diseases	3
Medication Use	2
Family History	3
Mental Health Status	4
Sleep Patterns	3
Education Level	3
Income Level	3

Data Cleaning:

Missing values were detected due to the usage of the word ‘Never’ as a unique value in several categorical columns, including Alcohol Consumption, Chronic Diseases, Medication Use, Family History, and Education Level. These were replaced with "Don't Have" to preserve categorical structure. No missing values were found in the numerical features. The Blood Pressure (s/d) column was split into two numeric fields - Systolic_BP and Diastolic_BP - and the original column was removed. Categorical features were transformed using One-Hot Encoding, converting them into a binary format suitable for machine learning. This expanded the dataset from 26 to 56 columns.

```
Gender                          0
Height (cm)                     0
Weight (kg)                     0
Blood Pressure (s/d)            0
Cholesterol Level (mg/dL)       0
BMI                             0
Blood Glucose Level (mg/dL)     0
Bone Density (g/cm²)            0
Vision Sharpness                0
Hearing Ability (dB)            0
Physical Activity Level         0
Smoking Status                  0
Alcohol Consumption             0
Diet                            0
Chronic Diseases                0
Medication Use                  0
Family History                  0
Cognitive Function              0
Mental Health Status            0
Sleep Patterns                  0
Stress Levels                   0
Pollution Exposure              0
Sun Exposure                    0
Education Level                 0
Income Level                    0
Age (years)                     0
Systolic_BP                     0
Diastolic_BP                    0
dtype: int64
```

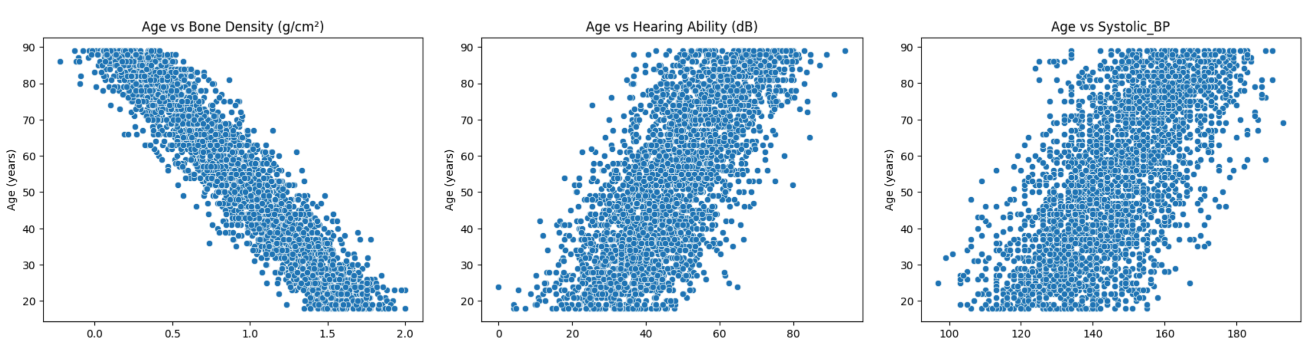
```
   Height (cm)  Weight (kg) Blood Pressure (s/d)  Cholesterol Level (mg/dL)  \
0   171.148359    86.185197              151/109                 259.465814
1   172.946206    79.641937              134/112                 263.630292
2   155.945488    49.167058              160/101                 207.846206
3   169.078298    56.017921               133/94                 253.283779
4   163.758355    73.966304              170/106                 236.119899

         BMI  Blood Glucose Level (mg/dL)  Bone Density (g/cm²)  \
0  29.423017                   157.652848              0.132868
1  26.626847                   118.507805              0.629534
2  20.217553                   143.587550              0.473487
3  19.595270                   137.448581              1.184315
4  27.582078                   145.328695              0.434562

   Vision Sharpness  Hearing Ability (dB)  Cognitive Function  ...  \
0          0.200000             58.786198           44.059172  ...
1          0.267312             54.635270           45.312298  ...
2          0.248667             54.564632           56.246991  ...
3          0.513818             79.722963           55.196092  ...
4          0.306864             52.479469           53.023379  ...

   Sleep Patterns_Excessive  Sleep Patterns_Insomnia  Sleep Patterns_Normal  \
0                       0.0                      1.0                    0.0
1                       0.0                      0.0                    1.0
2                       0.0                      1.0                    0.0
3                       0.0                      1.0                    0.0
4                       0.0                      0.0                    1.0

   Education Level_Don't Have  Education Level_High School  \
0                         1.0                          0.0
1                         0.0                          0.0
2                         1.0                          0.0
3                         1.0                          0.0
4                         0.0                          0.0

   Education Level_Postgraduate  Education Level_Undergraduate  \
0                           0.0                            0.0
1                           0.0                            1.0
2                           0.0                            0.0
3                           0.0                            0.0
4                           0.0                            1.0

   Income Level_High  Income Level_Low  Income Level_Medium
0                0.0               0.0                  1.0
1                0.0               0.0                  1.0
2                0.0               0.0                  1.0
3                0.0               1.0                  0.0
4                1.0               0.0                  0.0

[5 rows x 56 columns]
```

Exploratory Data Analysis (EDA)

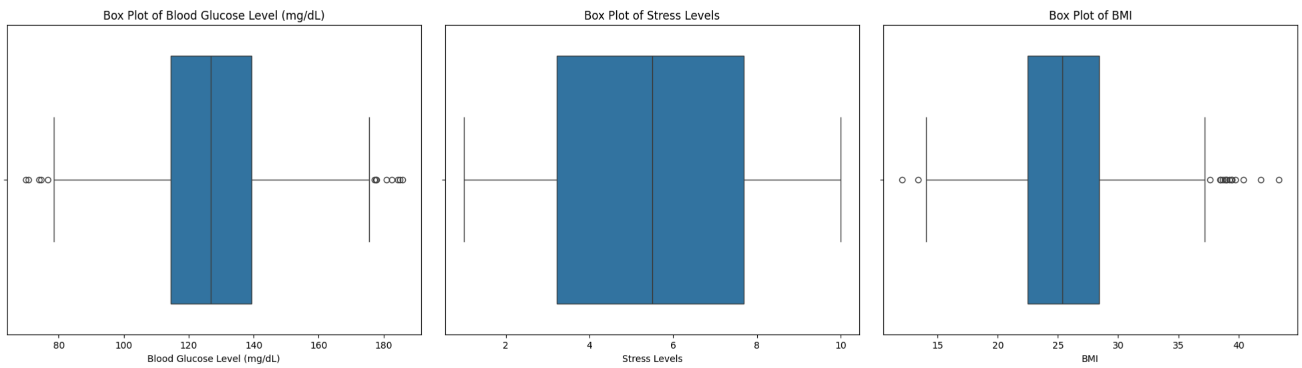
A summary table of the numerical features showed key statistics such as mean, minimum, maximum, and standard deviation.

index	Height	Weight	Cholesterol	BMI	Blood Glucose	Bone Density	Vision	Hearing	Cognitive	Stress	Pollution	Sun	Age (years)	Systolic	Diastolic
count	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000
mean	169	73	234	26	127	1	0	47	64	5	5	6	53	146	96
std	9	13	25	4	18	0	0	14	12	3	3	3	21	16	10
min	141	33	149	12	70	0	0	0	30	1	0	0	18	97	60
25%	162	63	217	22	114	1	0	37	56	3	3	3	36	135	89
50%	168	71	234	25	127	1	0	47	64	5	5	6	53	146	95
75%	176	82	251	28	139	1	1	57	72	8	7	9	72	157	103
max	198	124	331	43	186	2	1	94	106	10	10	12	89	193	133

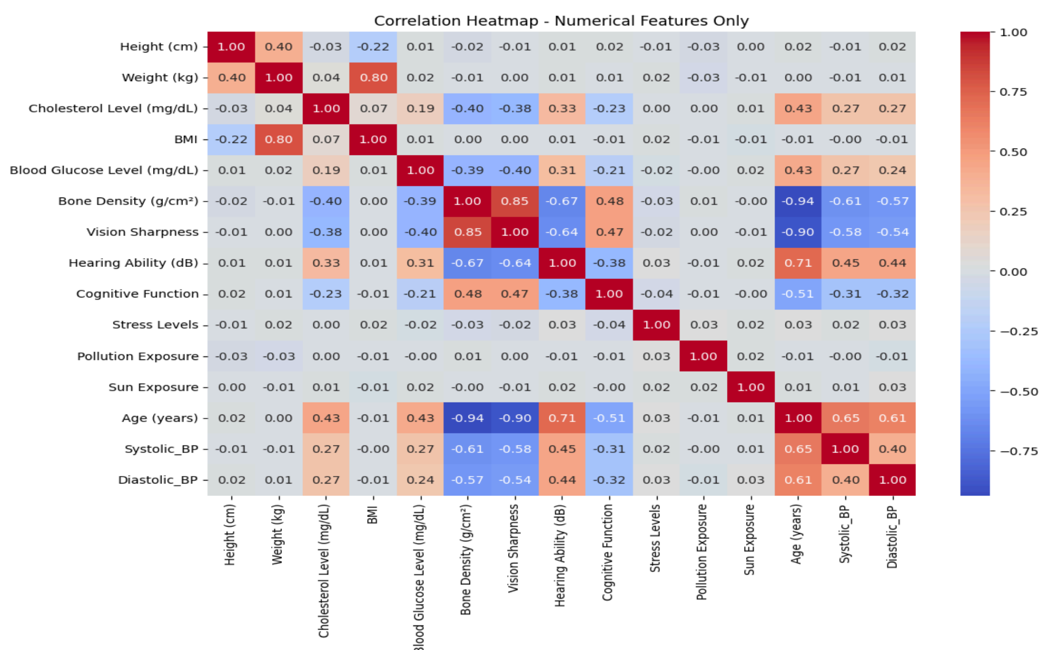
Scatterplots were used to examine the relationships between Age and other numerical features. These visualizations helped confirm that variables like Hearing Ability, Systolic_BP, and Bone Density exhibit strong correlations with age.



Boxplots revealed outliers in several numerical features, including high values in BMI, Blood Glucose, and Stress Levels, as well as extremes in Bone Density, Vision Sharpness, and Hearing Ability. These outliers were retained, as they reflect realistic variability in health profiles and enhance model robustness.



A correlation heatmap was created to understand relationships among numerical features. The strongest positive correlations with Age were found in Hearing Ability (r = 0.71), Systolic_BP (r = 0.65), and Diastolic_BP (r = 0.61). Negative correlations were observed with Bone Density (r = -0.94), and Vision Sharpness (r = -0.90)



Modeling and Evaluation:

Multiple regression-based algorithms were evaluated to determine the most effective model for age prediction. Linear Regression served as the baseline model, achieving a strong R^2 score of 0.93, despite its simplicity and linear assumptions. The Random Forest Regressor was effective in capturing non-linear relationships, yielding an R^2 of 0.92 and RMSE of 5.66. However, the XGBoost Regressor outperformed the others and was selected as the final model due to its robustness, scalability, and ability to control overfitting. After hyperparameter tuning using GridSearchCV, XGBoost achieved the best performance. Feature importance analysis from XGBoost revealed that Bone Density, Vision Sharpness, and Hearing Ability were the most influential predictors of age, while lifestyle-related features such as smoking status had comparatively lower predictive impact. Performance was evaluated using MAE, MSE, RMSE, and R^2.

	Model	MAE	MSE	RMSE	R^2 Score
0	Linear Regression	4.25	28.48	5.34	0.93
1	Random Forest	4.49	32.01	5.66	0.92
2	XGBoost (Default)	4.67	34.46	5.87	0.92
3	XGBoost (Tuned)	4.36	29.91	5.47	0.93

Feature Insights:

While correlation analysis showed that features like Hearing Ability (r = 0.71), Systolic_BP (r = 0.65), and Diastolic_BP (r = 0.61) had the strongest linear relationships with age, the model's feature importance told a deeper story. XGBoost identified Bone Density, Vision Sharpness, and Hearing Ability as the most influential predictors in actually estimating age. Interestingly, some highly correlated features like Diastolic_BP were ranked lower in predictive importance, while less correlated variables like Smoking Status contributed more significantly when combined with others. This highlights the model's ability to capture complex interactions beyond basic correlation.

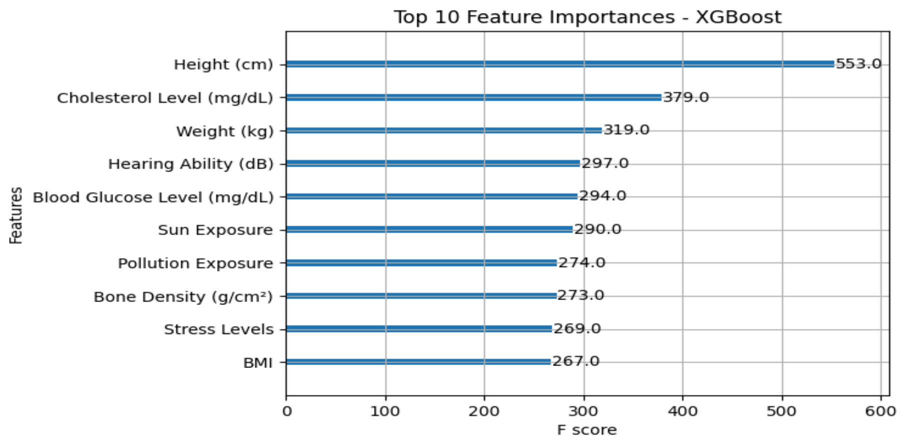
```
Top 10 most correlated features with Age
Hearing Ability (dB): 0.7124
Systolic_BP: 0.6461
Diastolic_BP: 0.6111
Cholesterol Level (mg/dL): 0.4324
Blood Glucose Level (mg/dL): 0.4286
Stress Levels: 0.0291
Height (cm): 0.0203
Sun Exposure: 0.0092
Weight (kg): 0.0025
```

Top 10 Predictive Features:

	Feature	Importance
5	Bone Density (g/cm^2)	0.736961
6	Vision Sharpness	0.142290
7	Hearing Ability (dB)	0.016714
12	Systolic_BP	0.013393
21	Smoking Status_Never	0.010138
13	Diastolic_BP	0.009534
8	Cognitive Function	0.005440
2	Cholesterol Level (mg/dL)	0.004908
4	Blood Glucose Level (mg/dL)	0.004511
20	Smoking Status_Former	0.003365

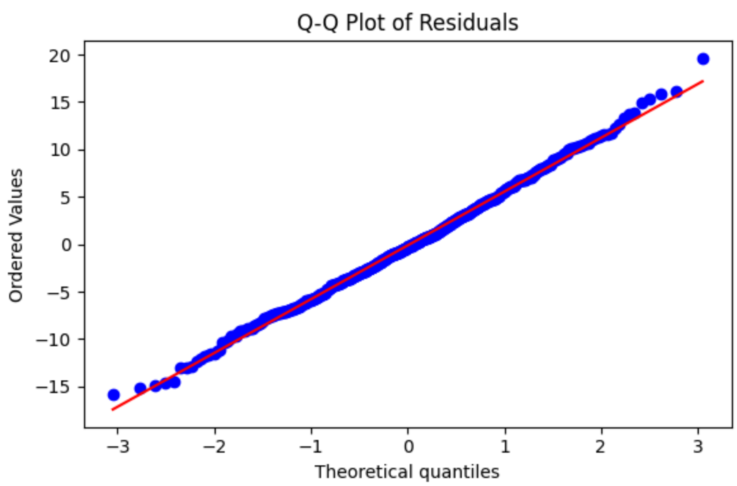
Feature Importances from XGBoost:

The F-score plot from XGBoost revealed that features such as **Height**, **Cholesterol**, and **Weight** were most frequently used in the model's decision trees. While this doesn't always align with correlation values, it highlights the model's reliance on a broader set of health indicators to capture subtle and nonlinear patterns in age prediction.



Q–Q Plot of Residuals

Residuals follow a near-normal distribution, suggesting a well-calibrated model.



Variable Significance:

According to the OLS Regression Summary the following variables have p-values > 0.05 and therefore are statistically insignificant: Height, Weight, BMI, Pollution Exposure and Sun Exposure. All other variables are statistically significant with p-values < 0.05. Vision Sharpness and Bone Density having higher coefficients (-23.4 and -28.5 respectively) which shows a highly negative relationship with Age variable. So, as Age goes up, Vision/Bone Density goes down which is scientifically proven as well with aging patterns. Those two items have a negative influence on aging.

```
                              OLS Regression Results
==============================================================================
Dep. Variable:            Age (years)   R-squared:                       0.938
Model:                            OLS   Adj. R-squared:                  0.937
Method:                 Least Squares   F-statistic:                     850.4
Date:                Sun, 11 May 2025   Prob (F-statistic):               0.00
Time:                        15:16:10   Log-Likelihood:                -7332.7
No. Observations:                2400   AIC:                         1.475e+04
Df Residuals:                    2357   BIC:                         1.500e+04
Df Model:                          42
Covariance Type:            nonrobust
==================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                              14.4792      2.621      5.523      0.000       9.339      19.620
Height (cm)                        -0.1190      0.073     -1.642      0.101      -0.261       0.023
Weight (kg)                         0.1274      0.081      1.566      0.117      -0.032       0.287
Cholesterol Level (mg/dL)           0.0314      0.005      6.587      0.000       0.022       0.041
BMI                                -0.4079      0.232     -1.762      0.078      -0.862       0.046
Blood Glucose Level (mg/dL)         0.0330      0.006      5.177      0.000       0.020       0.046
Bone Density (g/cm²)              -23.4002      0.510    -45.852      0.000     -24.401     -22.399
Vision Sharpness                  -28.5135      1.003    -28.424      0.000     -30.481     -26.546
Hearing Ability (dB)                0.1355      0.010     13.324      0.000       0.116       0.155
Cognitive Function                 -0.0573      0.011     -5.440      0.000      -0.078      -0.037
Stress Levels                       0.0052      0.041      0.127      0.899      -0.076       0.086
Pollution Exposure                 -0.0080      0.037     -0.216      0.829      -0.081       0.065
Sun Exposure                       -0.0242      0.031     -0.785      0.433      -0.085       0.036
Systolic_BP                         0.0974      0.009     11.425      0.000       0.081       0.114
Diastolic_BP                        0.1356      0.013     10.223      0.000       0.110       0.162
Gender_Female                       7.0411      1.284      5.483      0.000       4.523       9.559
Gender_Male                         7.4381      1.358      5.479      0.000       4.776      10.100
```

Insights & Interpretation:

There are a few takeaways from this research that we can getaher. First, we can determine that cardiovascular and metabolic factors (especially hearing ability, blood pressure, and glucose) are strong indicators of age. Additionally, Linear Regression provided the most accurate and interpretable results. XGBoost's feature analysis also revealed moderate contributions from sun exposure, stress, and pollution, highlighting the role of the environment in aging. Lastly, residual plots and Q-Q analysis confirmed that errors were symmetrically distributed and unbiased which led us to confirm that the model was accurate.

Practical Implications:

The implications of this project go beyond academic interest due to a variety of factors. First, it will aid in healthcare monitoring. Our model could be used in routine checkups to flag individuals whose biological age appears inconsistent with their chronological age. Second, it will allow for better preventative techniques such as enhancing health tools. Health apps can integrate such models to give users insights into how lifestyle changes affect their biological aging. Another stakeholder that benefits is insurance providers. They might find predictive age assessments helpful in creating better health risk models. Lastly, this will help researchers in their medical research to promote better care for all. Studies in gerontology and wellness could use similar frameworks to track aging trends in populations.

Limitations:

This project used synthetic data, which may not reflect real-world complexity. Placeholder values for missing data may oversimplify health histories, and one-hot encoding increases dimensionality. Additionally, the dataset lacks time-series information.

Future work:

To improve real-world impact, future work should apply these models to clinical datasets and incorporate customer data. Expanding the target to biological or perceived age could also enhance healthcare relevance.

Conclusion:

Based on our evaluation results, the tuned XGBoost model demonstrated strong predictive capability, achieving an R^2 score of 0.93 and low RMSE. This confirms that physiological and lifestyle features like bone density, hearing ability, and blood pressure are reliable indicators of aging. The findings were consistent across different evaluation methods and visual analyses, such as residual plots and feature importance charts. These results not only validate our choice of data mining approach but also suggest that non-invasive health attributes can be effectively used for estimating biological age. Overall, our conclusions logically follow from the statistical evidence and model performance, supporting the real-world viability of such predictive tools in healthcare and wellness applications.

Works Cited:

Abdullah, M. “Human Age Prediction Synthetic Dataset.” *Kaggle*, 4 Sept. 2024, www.kaggle.com/datasets/abdullah0a/human-age-prediction-synthetic-dataset

Lind, Tara. “Predicting Human Life Exepectancy.” *RPubs*, RStudio, 2024, rpubs.com/taralind/lifeexpectancyproject

Oliveira, Willian, et al. “Estimation of Human Age Using Machine Learning on Panoramic Radiographs for Brazilian Patients.” *Nature News*, Nature Publishing Group, 24 Aug. 2024 www.nature.com/articles/s41598-024-70621-1