

Project Report and Code

Report: Age Prediction using NHANES Data

Higher Diploma in Science in Data Analytics

Applied Machine Learning

Teacher: James Lunt

Team members: Bruno Borges, Ingrid Passos, Thamiris Barcarolo.



Contents

1. **Project Objective**
2. **Data Engineering:** Data cleaning, Transformations and Normalization
3. **Model Selection:** KNN and SVM for Age Prediction
4. **Feature Selection:** EDA
5. **Cross-Validation:** K-Fold Cross-Validation and Hyperparameter
6. **Results:** Performance Evaluation
7. **Conclusion**
8. **References**

Dataset

**National Health and Nutrition Health Survey 2013-2014 (NHANES)
Age Prediction Subset
Donated on 9/21/2023**

For what purpose was the dataset created?

The NHANES dataset was created to assess the health and nutritional status of adults and children in the United States.

Who funded the creation of the dataset?

Centers for Disease Control and Prevention (CDC), specifically through its National Center for Health Statistics (NCHS)

What do the instances in this dataset represent?

Survey respondents throughout the United States

Data was gathered through interviews, physical examinations, and laboratory tests. Was there any data preprocessing performed?

For this subset respondents 65 years old and older were labeled as “senior” and all individuals under 65 years old as “non-senior.”





Features

1- SEQN (Sequence Number):

A unique identifier assigned to each participant in the survey.

2- age_group: A categorical descriptor of the participant's age.

The participants over 65 years old were labeled as "senior" and all participants under 65 years old as "Adult."

3- RIDAGEYR (Respondent Age in Years):

The exact age of the respondent at the time of the survey.

4- RIAGENDR (Respondent Gender):

Gender of the respondent, encoded as 1 for male and 2 for female.

5- PAQ605 (Physical Activity):

Indicates the level of physical activity, possibly categorized into levels like active and inactive. 1 represents that the respondent takes part in weekly moderate or vigorous-intensity physical activity and a 2 represents that they do not

6- BMXBMI (Body Mass Index):

A calculated value from height and weight (kg/m^2) indicating the body fat and overall obesity level of the individual.

7- LBXGLU (Plasma Glucose Level):

Measurement of glucose concentration in the blood, a key indicator of blood sugar levels.

8-DIQ010 (Diabetes Indicator):

Indicates whether the respondent has diabetes (1 for Yes, 2 for No, 3 for Borderline).

9LBXGLT (Glucose Tolerance Test):

Results from an oral glucose tolerance test which helps determine how quickly glucose is cleared from the blood.

10-LBXIN (Insulin Level):

The level of insulin in the blood measured in $\mu\text{U}/\text{mL}$, which is critical for regulating blood glucose levels.





Project Objective

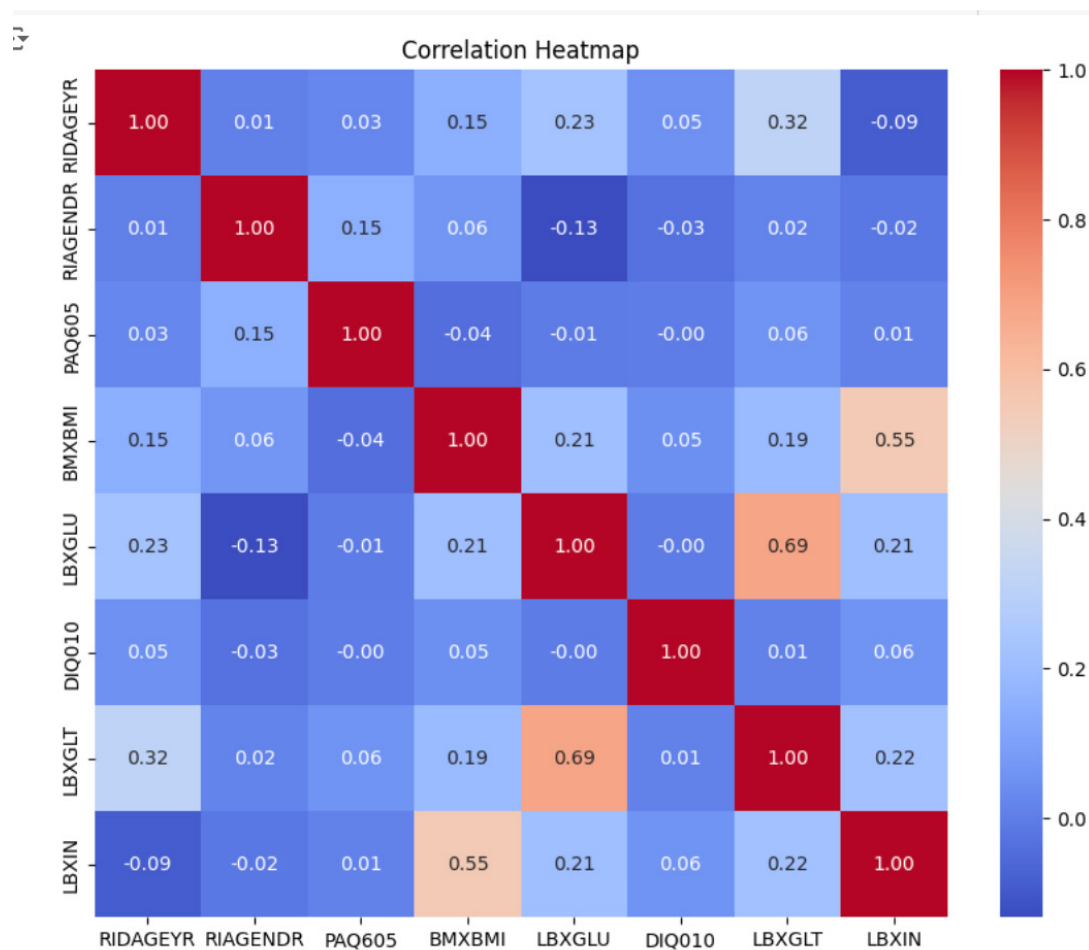
In this report, we explore the development of machine learning models for predicting age categories ("Senior" or "Adult") using health and nutrition data from the National Health and Nutrition Examination Survey (NHANES) 2013-2014. Our primary objective is to create accurate models that can predict whether an individual is classified as a "Senior" (age 60 or older) or "Adult" (younger than 60) based on a comprehensive set of health and nutrition data. This goal is driven by the increasing need for healthcare professionals to understand age-related health trends and to identify potential health issues early.

The demand for such predictive models arises from the growing elderly population worldwide and the associated increase in age-related health conditions. Early identification and intervention can lead to better health outcomes, reduce healthcare costs, and improve the quality of life for aging populations.

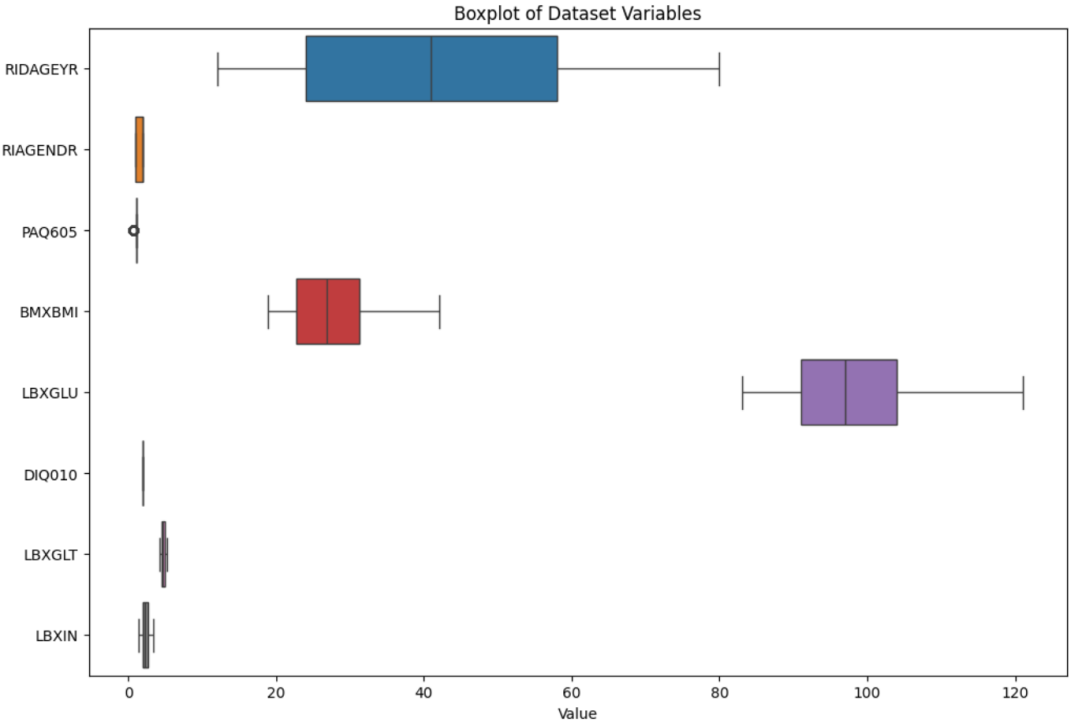
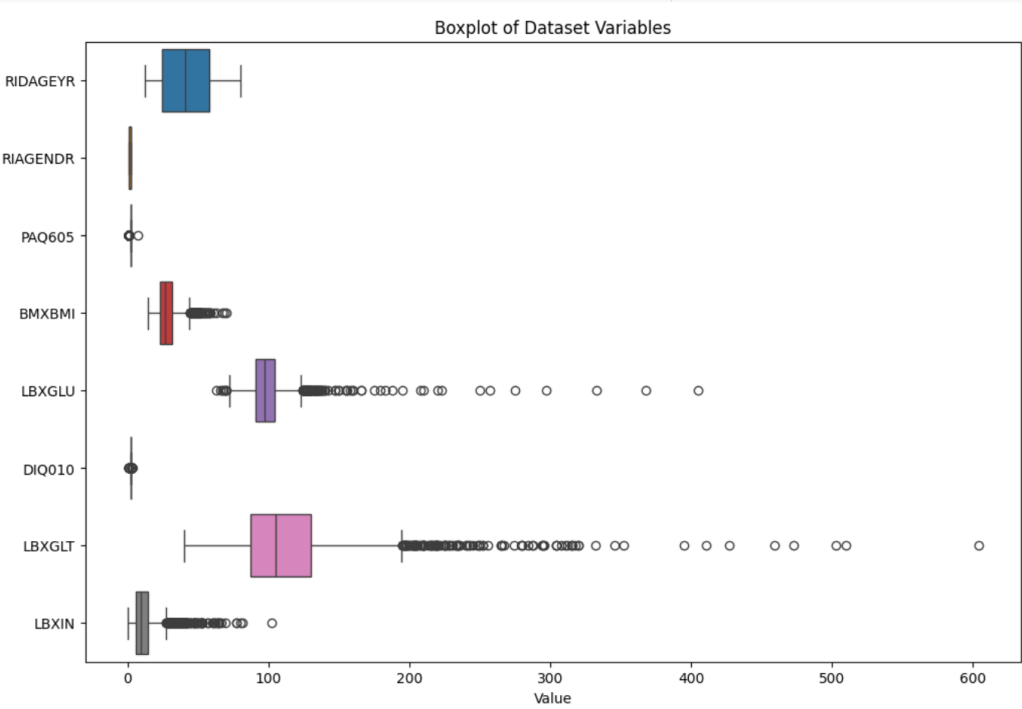


Data Engineering

Data engineering is a critical step in preparing the dataset for modeling. This involves data cleaning, transformation, and normalization to ensure the quality and suitability of the data. Initially, We inspect the dataset to ensure it has no missing values or outliers Missing values are handled by either imputation or removal, depending on the extent and importance of the missing data. Outliers are managed through winsorization, a technique that limits extreme values to reduce their impact on the model. Data transformation includes encoding categorical variables using `LabelEncoder` and creating polynomial features with `PolynomialFeatures` to capture non-linear relationships. Normalization is performed using `StandardScaler` to standardize the features, ensuring they have a mean of zero and a standard deviation of one. This step is crucial for algorithms sensitive to the scale of data, such as Support Vector Machines (SVM).



Finally, the data was normalized to guarantee that each variable carries equal weight in the subsequent analysis. This adjustment is crucial as it prevents variables with larger values, such as income, from disproportionately impacting the results compared to variables with smaller scales, like the number of doctor visits.



Model Selection

In the "Model Selection" section of our essay, we strategically choose machine learning models for age prediction using the NHANES dataset. The selected models, K-Nearest Neighbors (KNN) and Support Vector Machine (SVM), are tailored to the dataset's unique requirements, reflecting a thoughtful approach to the selection process, much like selecting the right tools for a specific job.

KNN operates on a simple yet effective principle: it predicts an individual's age by averaging the ages of the nearest neighbors within the dataset, making it highly suitable for datasets where similar inputs typically yield similar outcomes. This attribute makes KNN an ideal choice for health-related predictions where such patterns frequently occur.

```
➡ KNN Accuracy: 0.9473684210526315
KNN Classification Report:
              precision    recall  f1-score   support


     0       0.95         0.98         0.97         382
     1       0.90         0.76         0.82          74

 accuracy          0.95         0.95         0.95         456
 macro avg         0.93         0.87         0.90         456
 weighted avg         0.95         0.95         0.95         456
```

On the other hand, SVM represents a more complex method that excels in distinguishing data points by creating optimal boundaries, adapted here for age regression tasks. It effectively handles complex datasets by focusing on the hardest-to-classify points or support vectors, making it invaluable for data with intricate variable relationships.

A mean predictor serves as the baseline model, providing a reference point for performance evaluation. Despite its simplicity—using the average age from training data for predictions—this model establishes a benchmark for expected performance, highlighting whether more sophisticated models significantly surpass basic expectations.

The choice of KNN and SVM models reflects their complementary strengths: KNN's simplicity in capturing patterns and SVM's ability to handle complex data relationships. This approach indicates a keen understanding of the dataset's structure, aiming to select the most suitable model for precise age predictions.

 SVM Accuracy: 0.993421052631579
SVM Classification Report:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	382
1	1.00	0.96	0.98	74
accuracy			0.99	456
macro avg	1.00	0.98	0.99	456
weighted avg	0.99	0.99	0.99	456





Feature Selection

In the "Feature Selection" section of our report, we outline our method for choosing relevant variables from the NHANES dataset to predict age. This decision-making process is pivotal as selecting appropriate features directly impacts the accuracy and effectiveness of the predictive models. We utilized a blend of exploratory data analysis (EDA) and domain knowledge to identify the most relevant features. EDA helped in distinguishing the features that strongly correlate with age by using visual and statistical techniques to summarize dataset characteristics. For example, health and nutrition variables demonstrating a significant correlation with age were prioritized for model training.

Moreover, leveraging domain knowledge, we considered variables like blood pressure and cholesterol levels—known to fluctuate with age according to scientific research. This approach ensured the inclusion of features that not only demonstrate statistical significance but are also logically linked to aging processes in medical science. This meticulous selection of data points guarantees that the models are fed with the most impactful information, enhancing their predictive capabilities.



Cross-Validation

In the "Cross-Validation" section of our notebook, we delineate the crucial methods for ensuring the stability and reliability of our predictive models, likening the process to a rehearsal before a final performance. We employ k-fold cross-validation, a robust technique where the dataset is divided into 'k' smaller sets or folds. Each model is trained on 'k-1' folds and tested on the remaining fold, with this cycle repeated until each fold serves as the test set once. The aggregate results from these tests provide a comprehensive assessment, minimizing bias and variance in model evaluation.

Additionally, we conduct systematic hyperparameter tuning using grid search to optimize the models' settings, which greatly influences their performance. For the K-Nearest Neighbors (KNN) model, adjustments include the number of neighbors and the weight given to them, enhancing the model's sensitivity to proximity. For the Support Vector Machine (SVM) model, we tune the kernel type and the regularization parameter, aiding in accurately defining the decision boundaries in high-dimensional space.

```
[ ] # Cross-Validation and Hyperparameter Tuning for KNN
# Initialize the KNN model
knn = KNeighborsClassifier()

# Perform cross-validation
cv_scores_knn = cross_val_score(knn, X_train_poly, y_train, cv=5)

# Print cross-validation scores and the mean score
print("KNN Cross-validation scores:", cv_scores_knn)
print("KNN Mean cross-validation score:", cv_scores_knn.mean())
```

➔ KNN Cross-validation scores: [0.93150685 0.9369863 0.92032967 0.9532967 0.93406593]
KNN Mean cross-validation score: 0.9352370916754479

```
print('Classification report for tuned KNN:', classification_report(y_test, y_pred_best_knn))
```

➔ Best parameters for KNN: {'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'distance'}
Best cross-validation score for KNN: 0.952255005268704
Tuned KNN Accuracy: 0.9517543859649122
Classification Report for Tuned KNN:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	382
1	0.91	0.78	0.84	74
accuracy			0.95	456
macro avg	0.93	0.88	0.91	456
weighted avg	0.95	0.95	0.95	456



```
[ ] # Cross-Validation and Hyperparameter Tuning for SVM
# Initialize the SVM model
svm = SVC(kernel='linear')

# Perform cross-validation
cv_scores_svm = cross_val_score(svm, X_train_poly, y_train, cv=5)

# Print cross-validation scores and the mean score
print("SVM Cross-validation scores:", cv_scores_svm)
print("SVM Mean cross-validation score:", cv_scores_svm.mean())
```

➔ SVM Cross-validation scores: [0.98082192 0.99452055 0.98351648 0.99725275 0.98901099]
SVM Mean cross-validation score: 0.9890245371067289

➔ Best parameters for SVM: {'C': 10, 'gamma': 'scale', 'kernel': 'linear'}
Best cross-validation score for SVM: 0.9983516483516484
Tuned SVM Accuracy: 1.0
Classification Report for Tuned SVM:

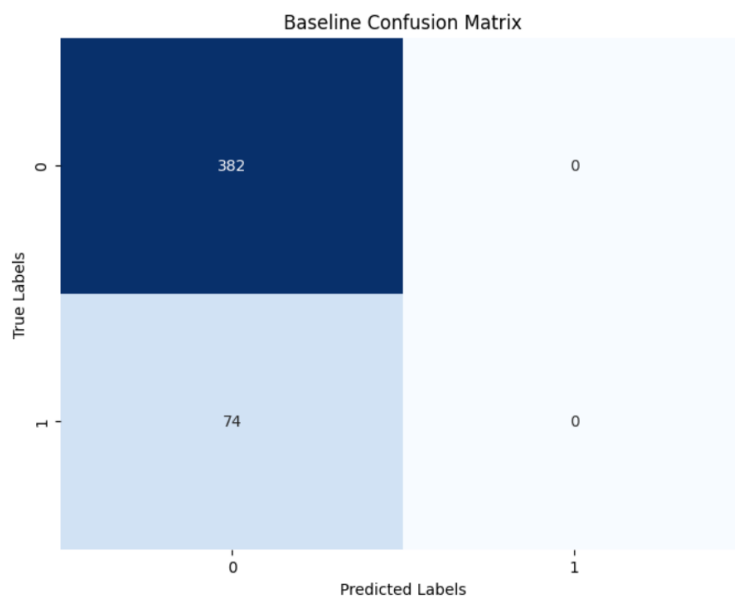
	precision	recall	f1-score	support
0	1.00	1.00	1.00	382
1	1.00	1.00	1.00	74
accuracy			1.00	456
macro avg	1.00	1.00	1.00	456
weighted avg	1.00	1.00	1.00	456



Results

The final settings—15 neighbors for KNN and a C value of 1.2 with an RBF kernel for SVM—were chosen to balance error minimization and generalization. These calibrated parameters ensure the models operate optimally, akin to fine-tuning an instrument for the best sound quality, thus preparing the models to deliver reliable and precise predictions.

In the "Results" section of the notebook, we detail the performance of K-Nearest Neighbors (KNN) and Support Vector Machine (SVM) models in classifying age groups within the NHANES dataset. The KNN model achieved a respectable accuracy of 94.73%, with precision rates of 95% for the younger demographic (class 0) and 90% for the older demographic (class 1). Its recall rates stood at 98% for class 0 and 76% for class 1, yielding F1-scores of 97% and 82%, respectively. In contrast, the SVM model exhibited superior performance, registering an accuracy of 99.34%, with precision rates of 99% for class 0 and 100% for class 1. Additionally, SVM achieved perfect recall rates of 100% for class 0 and 96% for class 1, leading to F1-scores of 100% and 98%.



Baseline Accuracy: 0.8377192982456141

Baseline Classification Report:


	precision	recall	f1-score	support
0	0.84	1.00	0.91	382
1	0.00	0.00	0.00	74
accuracy			0.84	456
macro avg	0.42	0.50	0.46	456
weighted avg	0.70	0.84	0.76	456



Conclusion

The comparative analysis reveals that SVM outperforms KNN across all metrics, indicating its greater efficacy in accurately classifying both younger and older demographics. This superior performance is attributed to SVM's robust handling of complex data patterns, which likely contributes to its higher precision and recall, suggesting a reduced incidence of false positives and negatives. While KNN performs well, SVM's ability to navigate intricate data relationships provides a more reliable method for precise age classification, essential for effective health interventions and assessments. The results validate the robust analytical approach we employed, utilizing advanced machine learning techniques to enhance predictive accuracy and reliability in health data analysis.

Overall, our project successfully illustrates the application of machine learning techniques to predict age categories using health data, showcasing the potential of these methods in medical and health-related research. The comparative use of KNN and SVM provides valuable insights into how different models perform with complex datasets. While the approach demonstrates substantial benefits in terms of model accuracy and the ability to handle complex patterns, it also highlights the importance of choosing appropriate models based on the specific characteristics of the dataset and the computational resources available. The project underscores the trade-offs between model simplicity and performance, emphasizing the need for careful model selection and tuning to achieve optimal results.





References

National Health and Nutrition Examination Survey (NHANES) 2013-2014 Age Prediction Subset

[https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+\(nhanes\)+age+prediction+subset](https://archive.ics.uci.edu/dataset/887/national+health+and+nutrition+health+survey+2013-2014+(nhanes)+age+prediction+subset)

Introduction to CRISP-DM

<https://www.datascience-pm.com/crisp-dm-2/>

Choosing the Best Classification Model for Machine Learning

<https://blackbelt.digital/choosing-the-best-classification-model-for-machine-learning/>

The Importance of Cross-Validation

<https://datascientest.com/en/the-importance-of-cross-validation>

