INTRODUCTION TO ARTIFICAL INTELLIGENCE

STROKE PREDICTION ANALYSIS

STROKE PREDICTION DATASET

The Stroke Prediction dataset aims to predict the occurrence of strokes in individuals by providing various demographic, clinical, and lifestyle-related features.

PROCESS

MODEL D1

Makes use of the original dataset.

MODEL D2

Transformed and normalized version of the original dataset by utilization of the MinMax scaling technique.

MODEL D3

Transformed and normalized version of the original dataset by utilization of the Standard scaling technique.

DATA PREPARATION

```
# Load stroke dataset
df = pd.read_csv('stroke.csv')
x = df.drop(columns=['stroke', 'id'])
y = df['stroke']

# Identify categorical columns and apply label encoding
categorical_columns = ['gender', 'ever_married', 'work_type', 'Residence_type', 'smoking_status']
label_encoder = LabelEncoder()
for column in categorical_columns:
    x[column] = label_encoder.fit_transform(x[column])
```

DATA PREPARATION

```
# Fill null values in the original dataset
imputer = SimpleImputer(strategy='median')
x_imputed = imputer.fit_transform(x)

D1 = pd.DataFrame(x_imputed, index=x.index, columns=x.columns)

# Declare scaler
minmax_scaler = MinMaxScaler()
standard_scaler = StandardScaler()

# Scale imputed data
x_minmax_scaled = minmax_scaler.fit_transform(x_imputed)
x_standard_scaled = standard_scaler.fit_transform(x_imputed)
D2 = pd.DataFrame(x_minmax_scaled, index=x.index, columns=x.columns)
D3 = pd.DataFrame(x_standard_scaled, index = x.index, columns=x.column

# Train and test (original dataset)
x_train, x_test, y_train, y_test = train_test_split(D1, y, test_size
=0.2, random_state=0)

# Train and test (minmax scaled dataset)
x_train_minmax, x_test_minmax, y_train_minmax, y_test_minmax
= train_test_split(D2, y, test_size=0.2, random_state=0)

# Train and test (standard scaled dataset)
x_train_standard, x_test_standard, y_train_standard, y_test_standard
= train_test_split(D3, y, test_size=0.2, random_state=0)

feature_names = x.columns.tolist()
```

CLASSIFICATION MODELS

RESULTS (D1)

Although model D1 (Original dataset) shows good accuracy, there is still room for improvement as it is outperformed by other models.

K-NN MODEL

- **Accuracy:** 0.9422700587084148
- **Confusion Matrix:** [[962, 6], [53, 1]]

DECISION TREE MODEL

- **Accuracy:** 0.9187866927592955
- **Confusion Matrix:** [[933, 35], [48, 6]]

CLASSIFICATION MODELS

RESULTS (D2)

Despite its strong performance and high accuracy, model D2 (MinMax Scaling) falls short of other alternatives.

K-NN MODEL

- **Accuracy:** 0.9461839530332681
- **Confusion Matrix:** [[966, 2], [53, 1]]

DECISION TREE MODEL

- **Accuracy:** 0.9129158512720157
- **Confusion Matrix:** [[927, 41], [48, 6]]

CLASSIFICATION MODELS

RESULTS (D3)

Based on accuracy, consistency, low misclassifications, and improved generalizability, the **D3 model (Standard Scaling) stands as the recommended model.**

K-NN MODEL

- **Accuracy:** 0.9471624266144814
- **Confusion Matrix:** [[967, 1], [53, 1]]

DECISION TREE MODEL

- **Accuracy:** 0.9178082191780822
- **Confusion Matrix:** [[932, 36], [48, 6]]

CLASSIFICATION MODELS

RECOMMENDATIONS

High Accuracy: Both the k-NN and Decision Tree models have high levels of accuracy for D3, indicating that the model can procure more accurate predictions.

Consistency: D3 performs consistently well across the k-NN and Decision Tree models. Standard scaling has a positive impact on model performance.

Low Misclassification: D3 shows a lower number of misclassifications compared to the other two models.

Improved generalizability: D3 addresses variations in feature scales and reduces the impact of outliers the best.

CLUSTERING

DATA PREPARATION

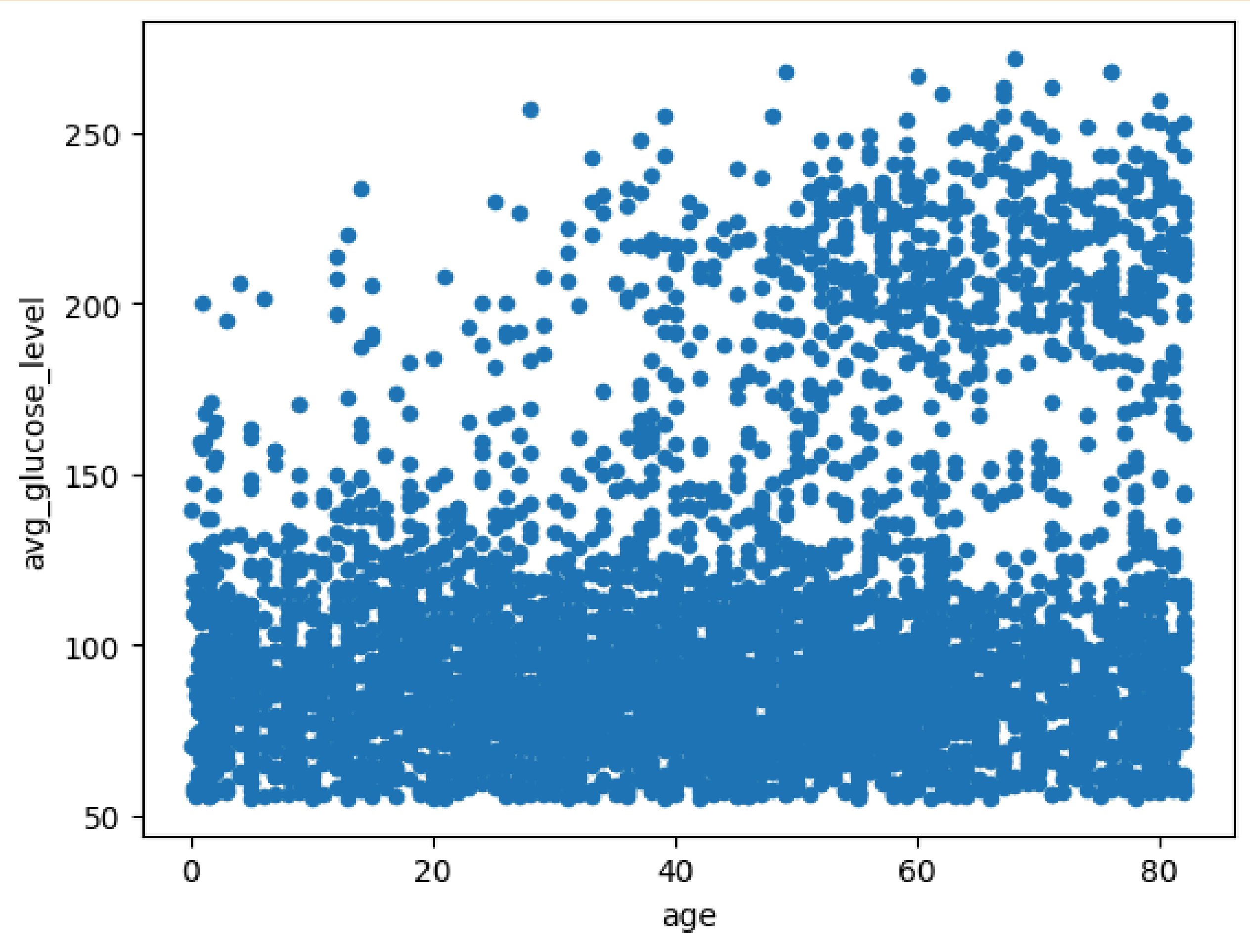
```
D1 = D1[['age', 'avg_glucose_level']]
D1.plot.scatter('age', 'avg_glucose_level')
```

D1

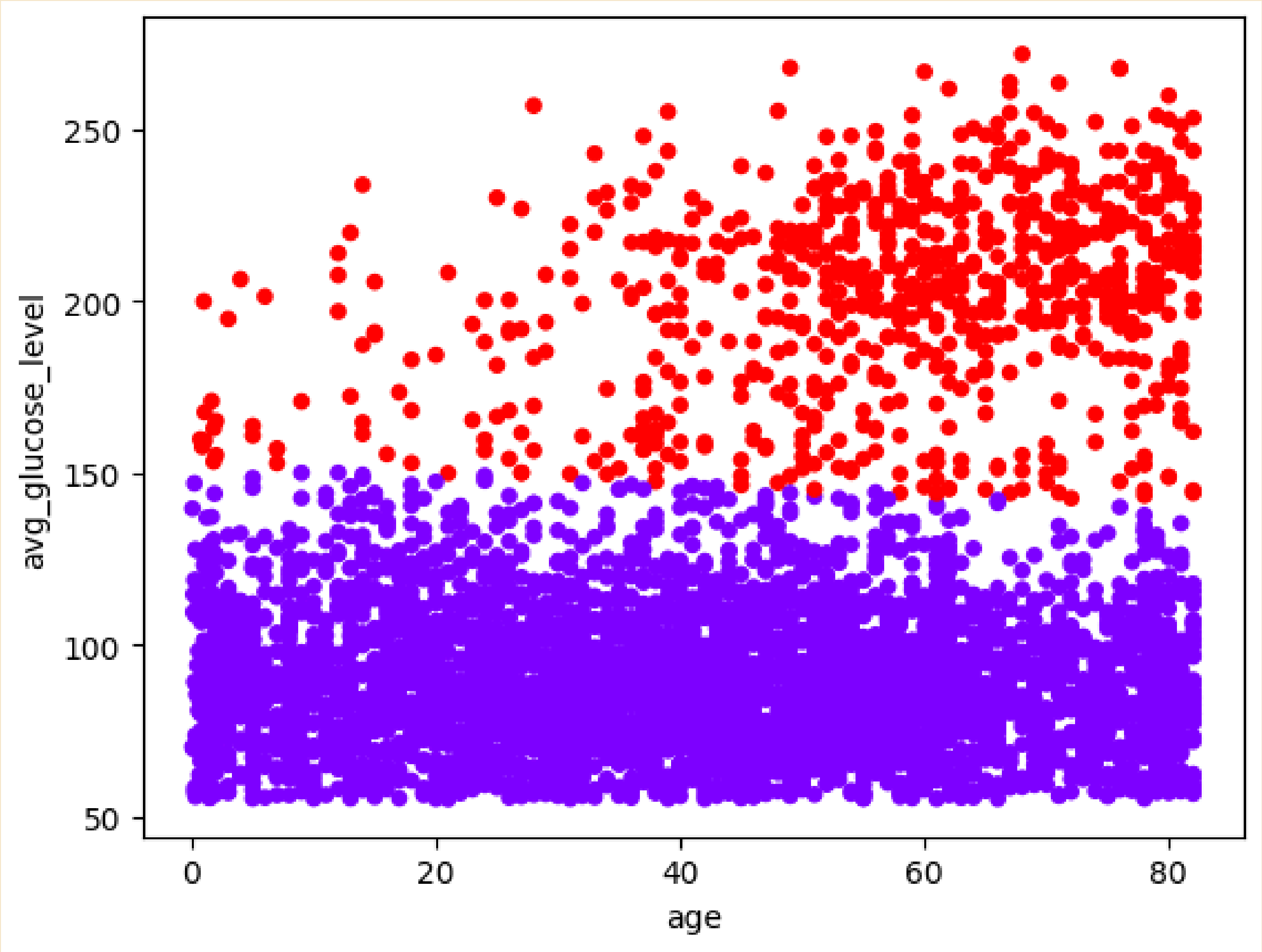
Silhouette scores

```
(2, 0.6558)
(3, 0.3992)
(4, 0.4099)
(5, 0.3933)
(6, 0.3695)
(7, 0.3393)
(8, 0.3477)
(9, 0.3392)
(10, 0.3375)
```

Unclustered



Clustered



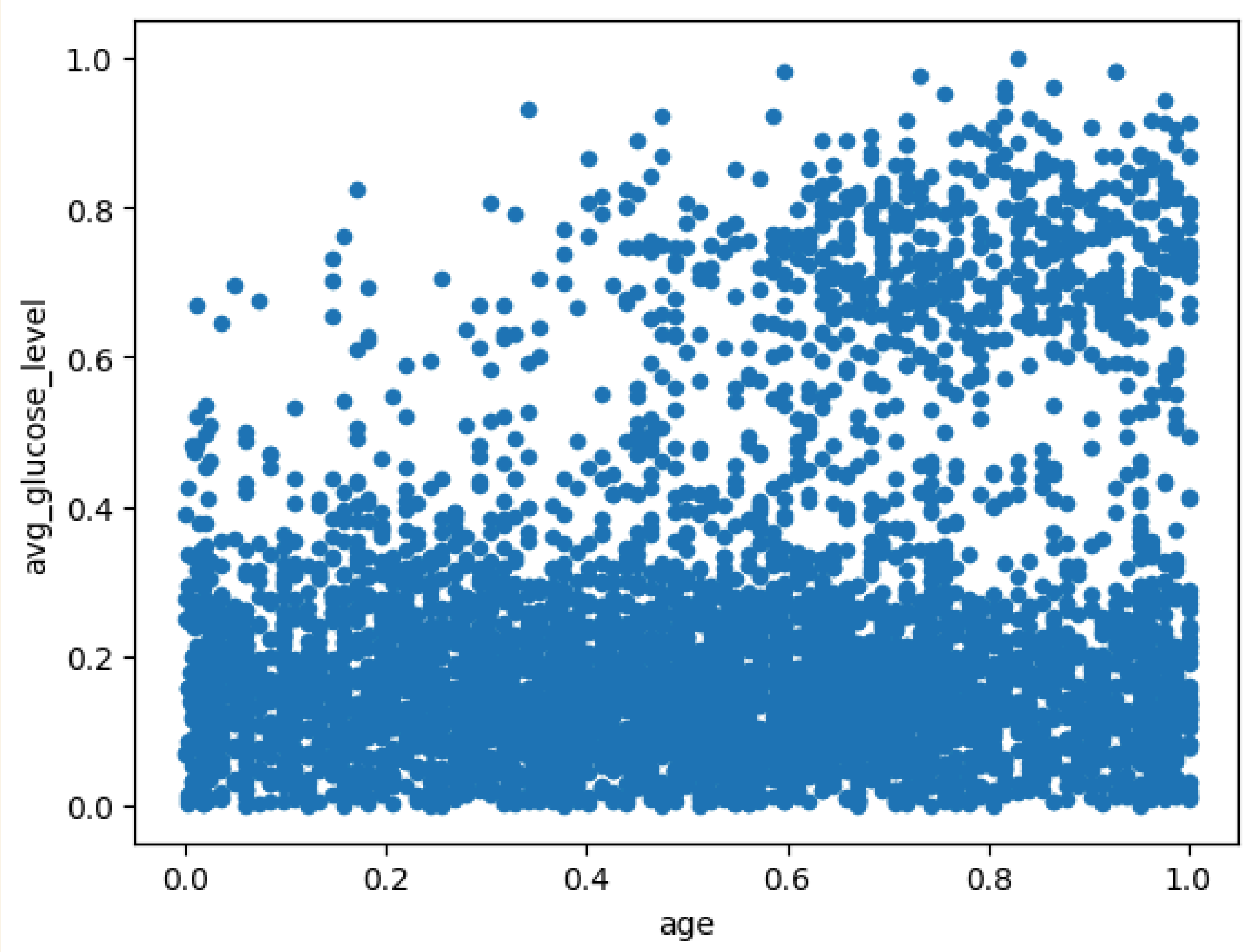
D2

MinMax Scaling

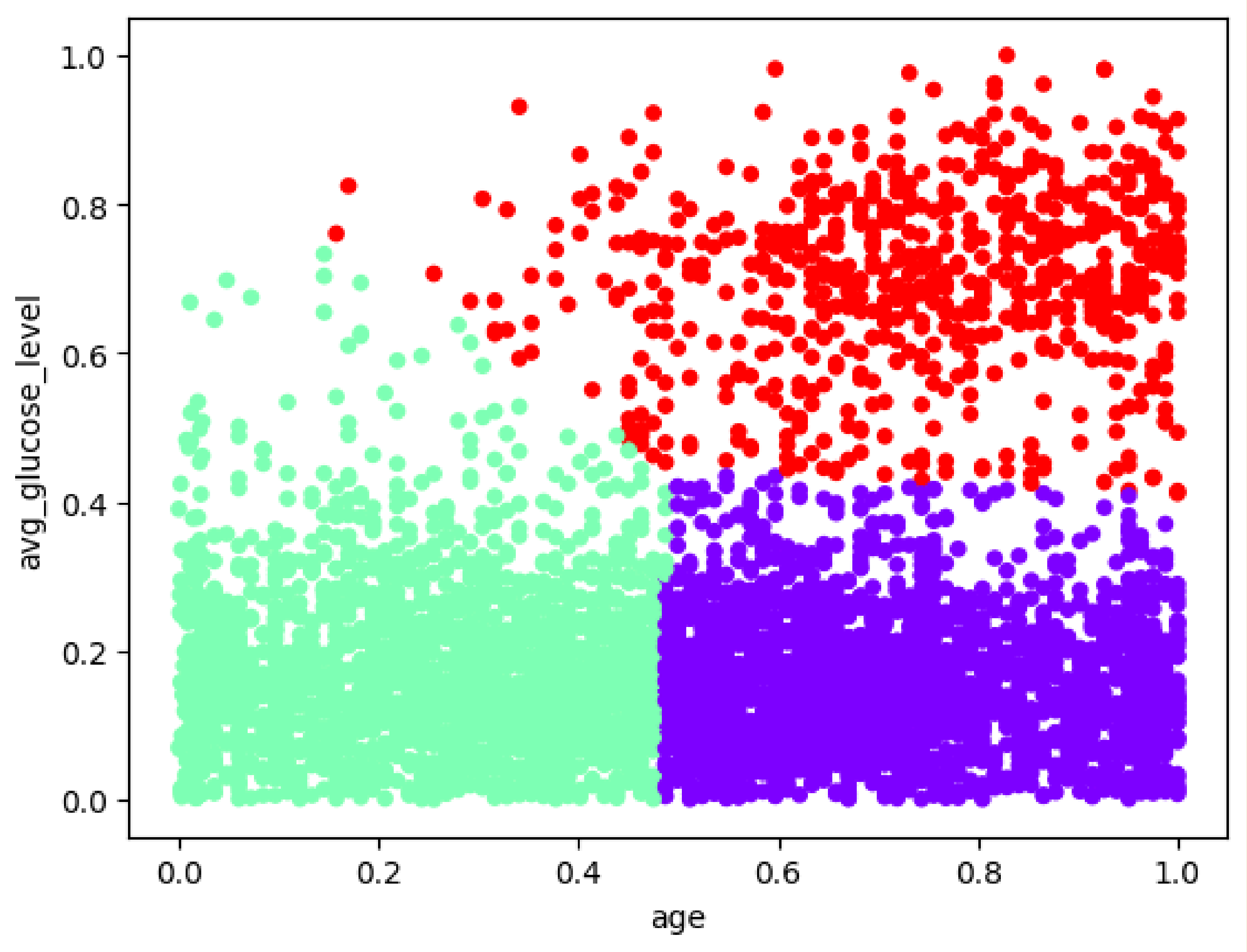
Silhouette scores

```
(2, 0.4297)
(3, 0.4993)
(4, 0.4366)
(5, 0.3895)
(6, 0.3921)
(7, 0.3982)
(8, 0.3532)
(9, 0.3666)
(10, 0.3637)
```

Unclustered



Clustered



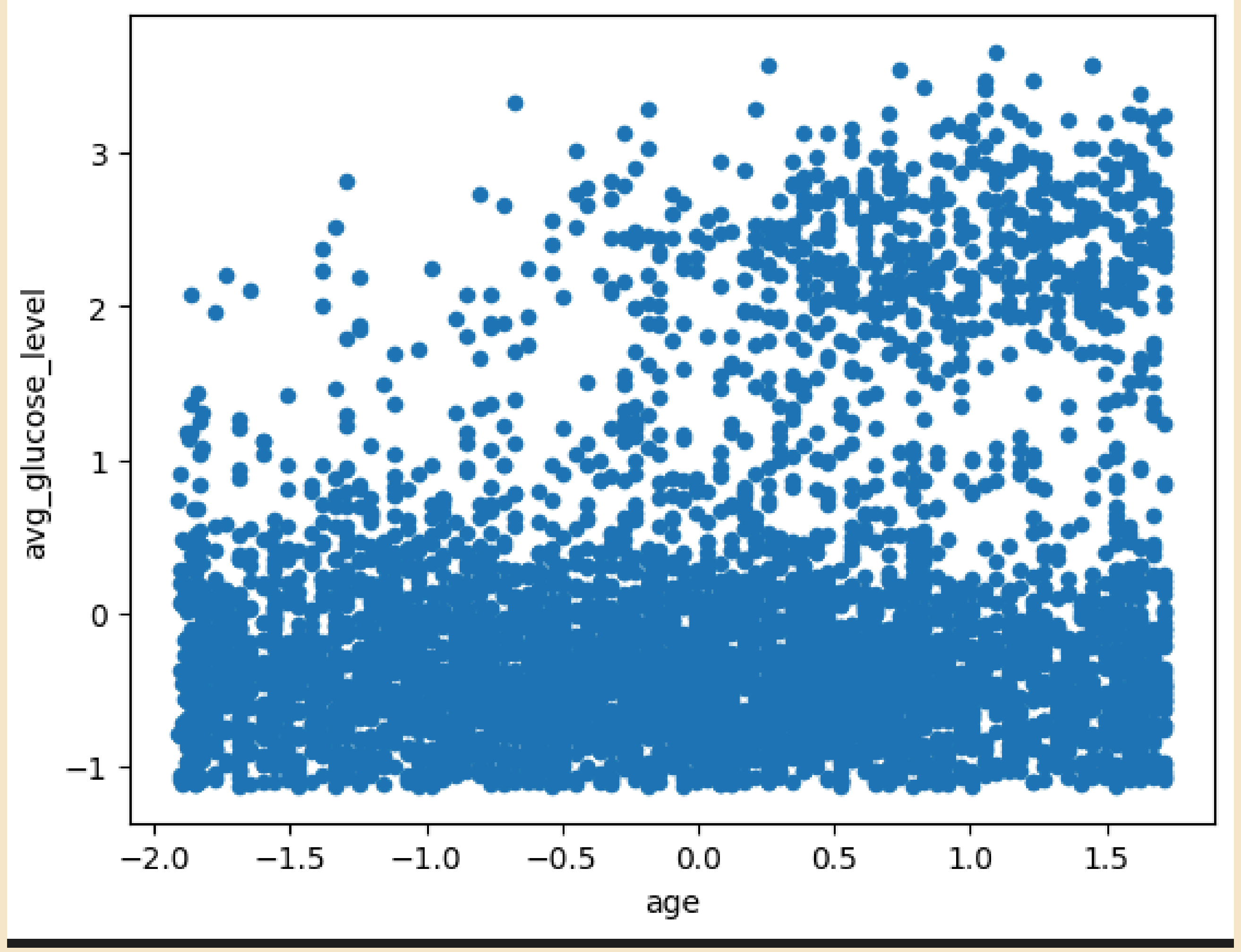
D3

Standard Scaling

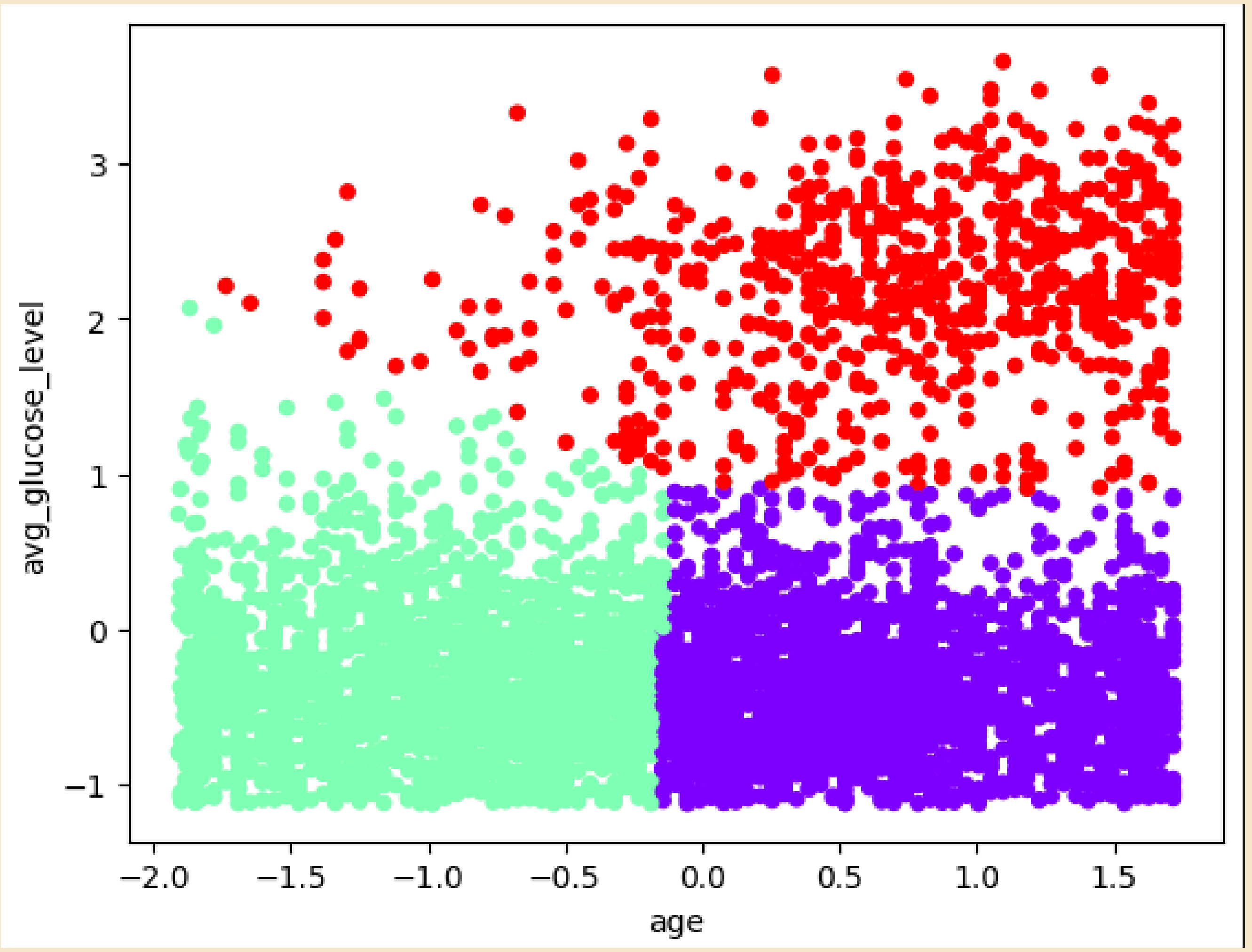
Silhouette scores

```
(2, 0.3858)
(3, 0.4847)
(4, 0.4061)
(5, 0.3582)
(6, 0.3638)
(7, 0.3817)
(8, 0.3638)
(9, 0.3532)
(10, 0.3424)
```

Unclustered



Clustered



CLUSTERING MODELS

RECOMMENDATIONS

Explore other types of clustering.

Figure out whether standardization is neccesary.

Try different kinds of standardization.

Experiment with which axes to use.

CONCLUSION

CLASSIFICATION

The classification models trained on the stroke dataset achieved **relatively high accuracy scores.** This indicates their **ability to predict the occurrence of strokes** with moderate to high accuracy whilst finding meaningful patterns that allow for **effective classification of stroke cases**.

CLUSTERING

The **unstandardized dataset** resulted into **more suitable clusters** compared to its standardized counterparts. The low silhouette scores of the standardized datasets suggests that the **K-Means process may not be the optimal method to use for them**.