# Diabetes Prediction using Machine Learning

John Zhou, Zichen Huang

# Introduction

01

**Problem Statement**:To predict the onset of diabetes using machine learning models.
**Motivation**: Diabetes is a global health issue. Early prediction can lead to timely interventions and better health outcomes.

02

**Approach**: Utilization of various regression and ensemble learning techniques to analyze and predict outcomes.Consideration of models ranging from simple linear regression to complex ensembles to capture underlying patterns.

03

**Improvement**: Previous studies often apply a one-size-fits-all model for all age groups. Our research expands on this by segmenting data based on age groups to enhance prediction accuracy.

04

**Limitation**:...

# Limitation

**Linear Regression (LR, Ridge, Lasso):**

- Assumes a linear relationship between predictors and outcome.
- Can be influenced by outliers.
- Multicollinearity can affect model estimates, though Ridge handles it better.
- Lasso may exclude relevant variables; Elastic Net might lack clarity on variable importance.
- Sensitive to feature scale, requiring standardization.

**PCR and PLS:**

- Original variable interpretability is compromised.
- Potential information loss if key components are discarded (PCR).
- PLS can be complex and harder to validate.

**Decision Trees:**

- Highly prone to overfitting without proper tuning.
- Sensitive to small data variations, leading to instability.
- Can be biased towards features with more categories.

**Random Forest:**

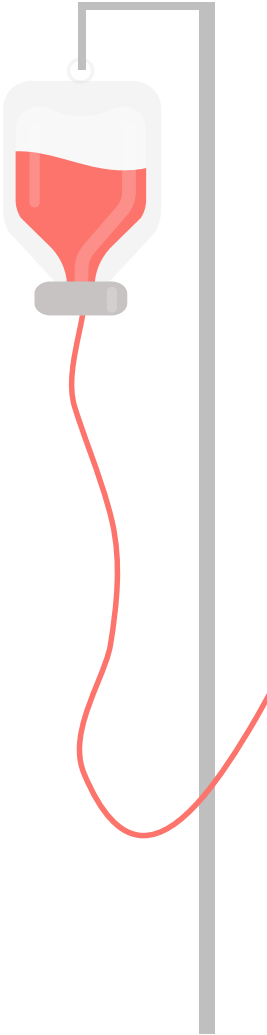
- Computationally intensive; large size can be memory-heavy.
- Although robust, can still overfit noisy data.
- More complex and less interpretable than simple models.

# Data Overview

This dataset comes from the National Institute of Diabetes and Digestive and Kidney Diseases. Its purpose is to predict the presence of diabetes in patients using specific diagnostic data included in the dataset. The dataset was curated with certain restrictions, including that all the subjects are female Pima Indians aged 21 or older.

| | max | min | average |
|---|---|---|---|
| Pregnancies | 17.00 | 0.000 | 3.845052 |
| Glucose | 199.00 | 44.000 | 121.686763 |
| BloodPressure | 122.00 | 24.000 | 72.405184 |
| SkinThickness | 99.00 | 7.000 | 29.108073 |
| Insulin | 846.00 | 14.000 | 155.548223 |
| BMI | 67.10 | 18.200 | 32.457464 |
| DiabetesPedigreeFunction | 2.42 | 0.078 | 0.471876 |
| Age | 81.00 | 21.000 | 33.240885 |
| Outcome | 1.00 | 0.000 | 0.348958 |

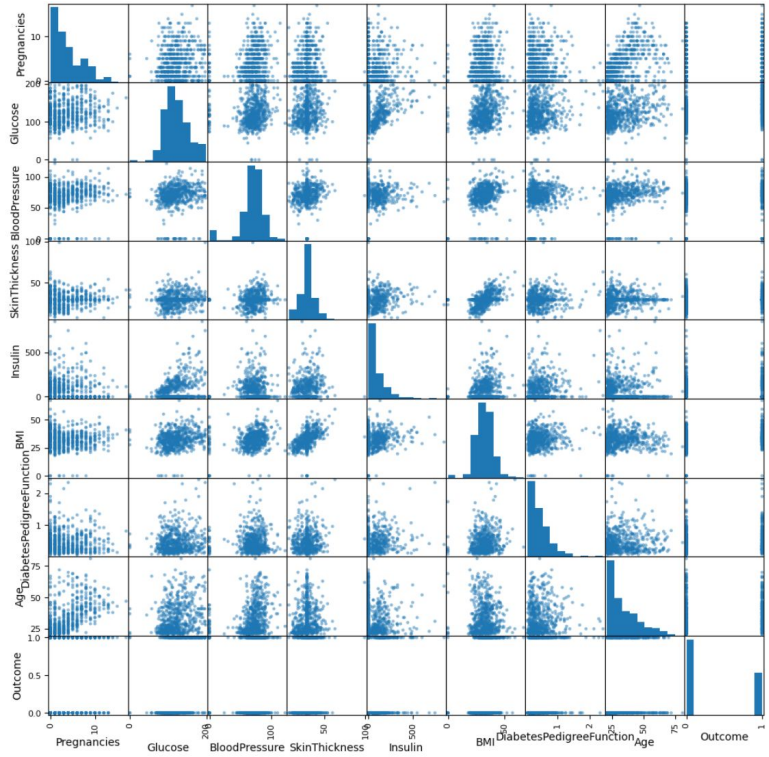
| | |
|---|---|
| Pregnancies | Number of times pregnant |
| Glucose | Plasma glucose concentration a 2 hours in an oral glucose tolerance test |
| Blood Pressure | Diastolic blood pressure (mm Hg) |
| Skin Thickness | Triceps skinfold thickness (mm) |
| Insulin | 2-Hour serum insulin (mu U/ml) |
| BMI | Body mass index (weight in kg/(height in m)^2). |
| Diabetes Pedigree Function | Diabetes pedigree function (a function which scores likelihood of diabetes based on family history). |
| Age | Age (years). |
| Outcome | Class variable (0 or 1) where 1 indicates diabetes and 0 indicates no diabetes. |



# Basic summary

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| **1** | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| **2** | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| **3** | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| **4** | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

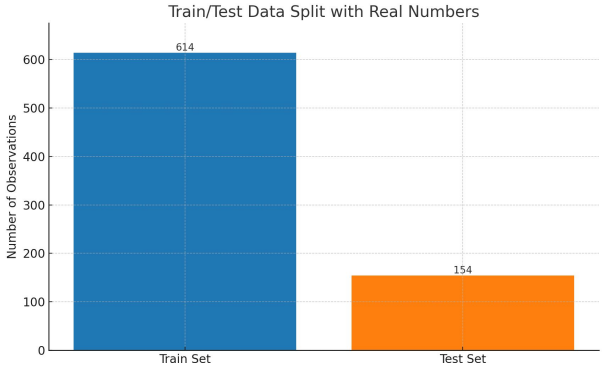
| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| Pregnancies | 1.000000 | 0.129406 | 0.189161 | 0.073995 | 0.063487 | -0.007550 | -0.047203 | 0.539582 | 0.193991 |
| Glucose | 0.129406 | 1.000000 | 0.223856 | 0.168225 | 0.424555 | 0.252352 | 0.122439 | 0.284982 | 0.490399 |
| BloodPressure | 0.189161 | 0.223856 | 1.000000 | 0.198708 | 0.063381 | 0.283517 | -0.008425 | 0.318606 | 0.146295 |
| SkinThickness | 0.073995 | 0.168225 | 0.198708 | 1.000000 | 0.128742 | 0.545312 | 0.103921 | 0.109514 | 0.188089 |
| Insulin | 0.063487 | 0.424555 | 0.063381 | 0.128742 | 1.000000 | 0.153159 | 0.107155 | 0.134318 | 0.199055 |
| BMI | -0.007550 | 0.252352 | 0.283517 | 0.545312 | 0.153159 | 1.000000 | 0.161797 | 0.005599 | 0.314197 |
| DiabetesPedigreeFunction | -0.047203 | 0.122439 | -0.008425 | 0.103921 | 0.107155 | 0.161797 | 1.000000 | 0.027692 | 0.163334 |
| Age | 0.539582 | 0.284982 | 0.318606 | 0.109514 | 0.134318 | 0.005599 | 0.027692 | 1.000000 | 0.238986 |
| Outcome | 0.193991 | 0.490399 | 0.146295 | 0.188089 | 0.199055 | 0.314197 | 0.163334 | 0.238986 | 1.000000 |



# Data Processing

The dataset contains 768 entries and 9 columns.

We split the data into 80% train data and 20% test data.




- **Model Selection**: Range of models to be executed: OLS, Ridge, Lasso, PCR, PLS, Decision Trees, Random Forest, and Boosting.
- **Parameter Tuning**: For regularization models: optimizing alpha and the l1_ratio for Lasso and Elastic Net. For tree-based models: tuning parameters such as max depth, min samples split, and min samples leaf.
- **Computing Environment:** Experiments will be conducted on Google Colab using stacks like scikit-learn, pandas, NumPy
- **Evaluation Metrics:** Mean Squared Error (MSE) for both training and test datasets.

- **Missing Data**: From an initial inspection, it appeared there might be some zero values in columns where it doesn't make sense (like Glucose, Blood Pressure, Skin Thickness, Insulin, BMI)
- **Solution:**We replace these zeros with the column's mean, excluding zeros from the calculation.

# Logistic regression

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Outcome | R-squared: | 0.312 |
| Model: | OLS | Adj. R-squared: | 0.303 |
| Method: | Least Squares | F-statistic: | 34.24 |
| Date: | Fri, 26 Apr 2024 | Prob (F-statistic): | 1.26e-44 |
| Time: | 21:29:43 | Log-Likelihood: | -305.89 |
| No. Observations: | 614 | AIC: | 629.8 |
| Df Residuals: | 605 | BIC: | 669.6 |
| Df Model: | 8 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.9655 | 0.118 | -8.177 | 0.000 | -1.197 | -0.734 |
| Pregnancies | 0.0164 | 0.006 | 2.932 | 0.003 | 0.005 | 0.027 |
| Glucose | 0.0066 | 0.001 | 10.427 | 0.000 | 0.005 | 0.008 |
| BloodPressure | -0.0020 | 0.002 | -1.349 | 0.178 | -0.005 | 0.001 |
| SkinThickness | -0.0008 | 0.002 | -0.379 | 0.704 | -0.005 | 0.003 |
| Insulin | -0.0002 | 0.000 | -0.875 | 0.382 | -0.001 | 0.000 |
| BMI | 0.0154 | 0.003 | 5.241 | 0.000 | 0.010 | 0.021 |
| DiabetesPedigreeFunction | 0.1218 | 0.050 | 2.445 | 0.015 | 0.024 | 0.220 |
| Age | 0.0030 | 0.002 | 1.763 | 0.078 | -0.000 | 0.006 |

| | | | |
|---|---|---|---|
| Omnibus: | 35.384 | Durbin-Watson: | 1.925 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 21.751 |
| Skew: | 0.322 | Prob(JB): | 1.89e-05 |
| Kurtosis: | 2.339 | Cond. No. | 1.68e+03 |

```
Training MSE: 0.1585820054514869
Test MSE: 0.13921144800508753
```

- About 31.2% of the variability in the outcome can be explained by the model
- Glucose has a positive coefficient (0.0066) with a very small p-value, suggesting that it is a significant predictor and that higher glucose levels are associated with an increased likelihood of the outcome occurring.
- BMI', and 'Pregnancies' as significant predictors for diabetes, which aligns with medical understanding.

# Best Subset Selection

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Outcome | R-squared (uncentered): | 0.498 |
| Model: | OLS | Adj. R-squared (uncentered): | 0.495 |
| Method: | Least Squares | F-statistic: | 201.9 |
| Date: | Fri, 26 Apr 2024 | Prob (F-statistic): | 5.56e-91 |
| Time: | 21:31:55 | Log-Likelihood: | -346.04 |
| No. Observations: | 614 | AIC: | 698.1 |
| Df Residuals: | 611 | BIC: | 711.3 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

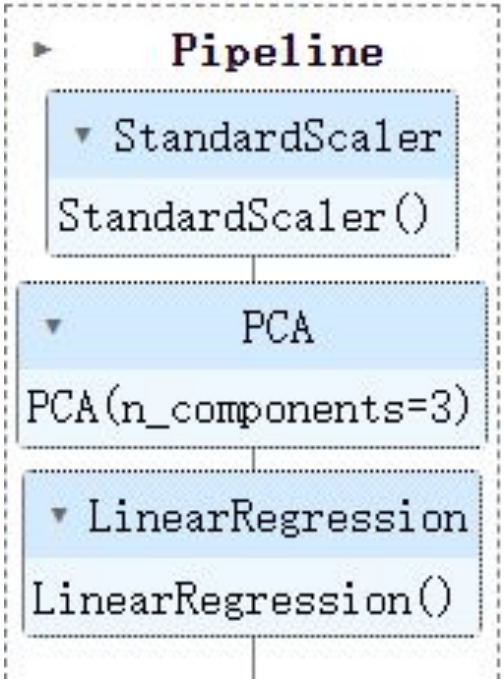
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Pregnancies | 0.0193 | 0.005 | 3.783 | 0.000 | 0.009 | 0.029 |
| Glucose | 0.0061 | 0.001 | 11.366 | 0.000 | 0.005 | 0.007 |
| BloodPressure | -0.0061 | 0.001 | -6.513 | 0.000 | -0.008 | -0.004 |

| | | | |
|---|---|---|---|
| Omnibus: | 174.970 | Durbin-Watson: | 1.992 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 53.069 |
| Skew: | 0.500 | Prob(JB): | 2.99e-12 |
| Kurtosis: | 1.964 | Cond. No. | 43.0 |

Training MSE: 0.18073553892411687
Test MSE: 0.1660702030271926

- Nearly half of the variability in the outcome can now be explained by the model, which is an improvement over the previous model.
- The significant predictors, according to this model, are 'Pregnancies', 'Glucose', and 'Blood Pressure'.
- The non-normal distribution of residuals remains a concern and could suggest that a linear model is not the best fit for this data.
- The MSEs are slightly higher than the previous model which had more predictors. This might happen when predictors that contribute to overfitting are removed.

# PCR

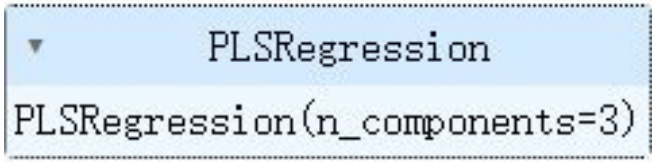



```
Coefficients: [0.14867637 0.01532704 0.06041417]
Training MSE: 0.1756678083054893
Test MSE: 0.14387061806946558
```

```
Accuracy: 0.7792207792207793
```

- PCA (n_components=3): Reducing the data to three principal components means the model is now working in a three-dimensional space.
- Using PCA reduces dimensionality, which might improve model performance. However, it loses the ability to interpret the model in terms of the original features directly.

# PLS



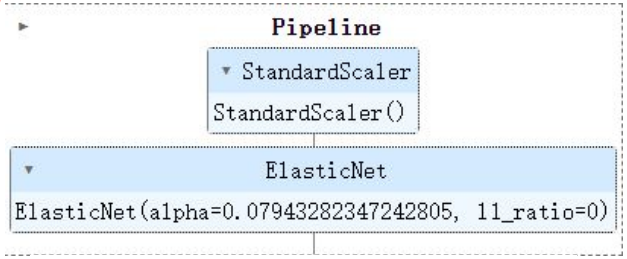

```
Top features for component 1: Index(['Glucose', 'BMI', 'Age', 'SkinThickness', 'Insulin', 'BloodPressure',
       'Pregnancies', 'DiabetesPedigreeFunction'],
      dtype='object')
Top features for component 2: Index(['BloodPressure', 'Glucose', 'SkinThickness', 'Age',
       'DiabetesPedigreeFunction', 'Pregnancies', 'Insulin', 'BMI'],
      dtype='object')
Top features for component 3: Index(['Insulin', 'DiabetesPedigreeFunction', 'Pregnancies', 'BloodPressure',
       'Age', 'BMI', 'Glucose', 'SkinThickness'],
      dtype='object')
```

```
Training MSE: 0.1586104519579192
Test MSE: 0.1392385345231498
```

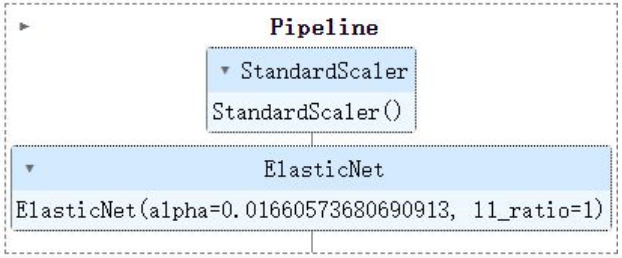
```
Accuracy: 0.8051948051948052
```

- The MSE values for PLS are in the same range as those obtained from previous model models.
- PLS can be particularly useful when there is a large number of predictors, many of which are correlated to avoid overfitting.
- The components in PLS can be more meaningful than those in PCA since they are designed to explain the dependent variable. The top features for each component give an insight into which combinations of features are most predictive for the outcome.

# Ridge & Lasso




Accuracy: 0.8051948051948052




```
['Pregnancies' 'Glucose' 'BloodPressure' 'SkinThickness' 'Insulin' 'BMI'
 'DiabetesPedigreeFunction' 'Age']
[ 0.04317758  0.18252337 -0.          0.         -0.          0.08392991
  0.02730522  0.0215632 ]
```

Accuracy: 0.7987012987012987

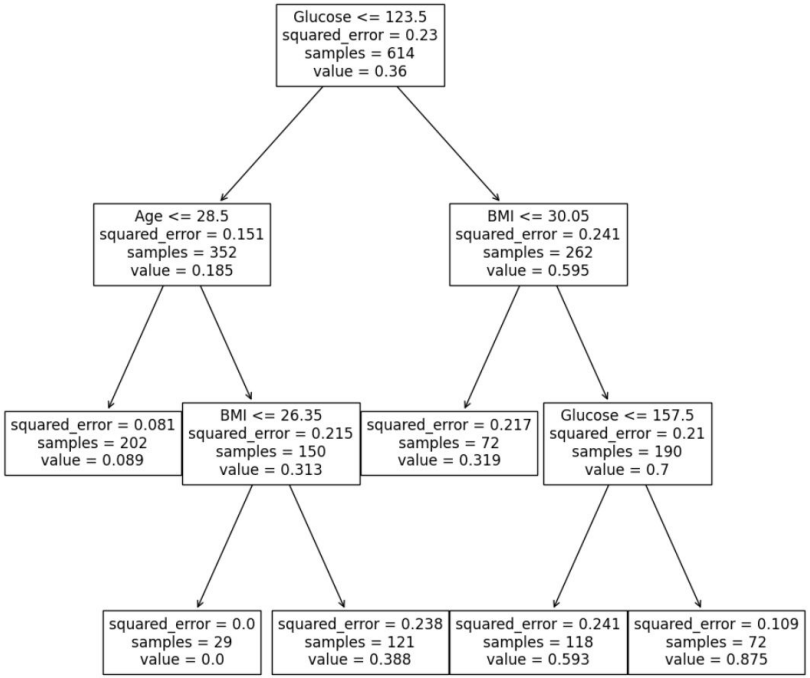
```
Training MSE: 0.15894900451394114
Test MSE: 0.13873909339962215
```

- Ridge regression, tends to shrink the coefficients for less important variables but does not set them to zero, which maintains all the features but reduces the effect of those that are not crucial.

```
Training MSE: 0.1601746976859794
Test MSE: 0.1392160448313773
```

- Lasso regression, with non-zero coefficients, suggests that features like 'Pregnancies', 'Glucose', and 'BMI' are important predictors for the outcome. 'BloodPressure' and 'Insulin' have been set to zero, implying that, according to Lasso, they might not be significant predictors given the presence of other variables.
- Lasso's feature elimination can be particularly useful as it is simpler and more interpretable while still being predictive.

# Decision Tree




Training MSE: 0.15821204447201495
Test MSE: 0.15299541631872235

Accuracy: 0.7207792207792207

- The decision tree has identified 'Glucose', 'BMI', and 'Age' as significant features for predicting the outcome, which aligns with medical understanding of diabetes risk factors.
- Prominence of ‘Glucose’ being the decision boundary of branches.
- Parameters determined through cross-validation.

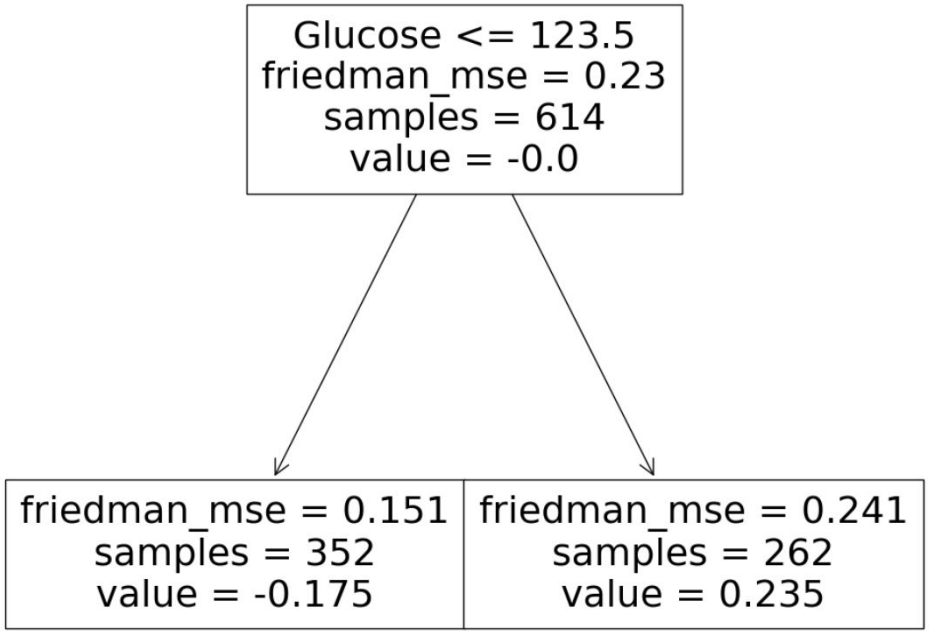
# Random Forest

```
For m is 3 Train MSE: 0.02355228013029316 Test MSE: 0.13693441558441558 Top three features are: ['Glucose', 'BMI', 'Age', 'DiabetesPedigreeFunction', 'BloodPressure', 'Insulin', 'Pregnancies', 'SkinThickness']
For m is 4 Train MSE: 0.023624755700325732 Test MSE: 0.13846233766233768 Top three features are: ['Glucose', 'BMI', 'Age', 'DiabetesPedigreeFunction', 'Insulin', 'BloodPressure', 'Pregnancies', 'SkinThickness']
For m is 5 Train MSE: 0.023418892508143323 Test MSE: 0.1424863636363636 Top three features are: ['Glucose', 'BMI', 'Age', 'DiabetesPedigreeFunction', 'BloodPressure', 'Insulin', 'SkinThickness', 'Pregnancies']
For m is 6 Train MSE: 0.02381986970684039 Test MSE: 0.13587597402597404 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'SkinThickness', 'Pregnancies']
For m is 7 Train MSE: 0.023841368078175895 Test MSE: 0.1362331168831169 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'SkinThickness', 'Pregnancies']
For m is 8 Train MSE: 0.023457166123778503 Test MSE: 0.1379831168831169 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'Pregnancies', 'SkinThickness']
For m is 9 Train MSE: 0.023457166123778503 Test MSE: 0.1379831168831169 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'Pregnancies', 'SkinThickness']
For m is 10 Train MSE: 0.023457166123778503 Test MSE: 0.1379831168831169 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'Pregnancies', 'SkinThickness']
For m is 11 Train MSE: 0.023457166123778503 Test MSE: 0.1379831168831169 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'Pregnancies', 'SkinThickness']
For m is 12 Train MSE: 0.023457166123778503 Test MSE: 0.1379831168831169 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'Pregnancies', 'SkinThickness']
For m is 13 Train MSE: 0.023457166123778503 Test MSE: 0.1379831168831169 Top three features are: ['Glucose', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'BloodPressure', 'Insulin', 'Pregnancies', 'SkinThickness']
```

| | M | Train Accuracy | Test Accuracy |
|---|---|---|---|
| 0 | 3 | 1.0000 | 0.8312 |
| 1 | 4 | 1.0000 | 0.8247 |
| 2 | 5 | 1.0000 | 0.7922 |
| 3 | 6 | 1.0000 | 0.8182 |
| 4 | 7 | 1.0000 | 0.8182 |
| 5 | 8 | 1.0000 | 0.8182 |
| 6 | 9 | 1.0000 | 0.8182 |
| 7 | 10 | 1.0000 | 0.8182 |
| 8 | 11 | 1.0000 | 0.8182 |
| 9 | 12 | 1.0000 | 0.8182 |
| 10 | 13 | 1.0000 | 0.8182 |

- The training and test MSE values generally decrease as the number of trees increases.
- However, after 7 trees, the improvement in MSE plateaus
- Prominence of ‘Glucose’ being the top 1 (most important) feature among 8, ‘BMI’ being the top 2 feature among 8.
- Training and test MSE stays relatively stable, suggesting an overall good performance of random forest.

# Boosting




```
test mse: 0.13592530460933203
train mse: 0.13556162889462128
```

```
{'learning_rate': 0.09099999999999998, 'max_depth': 2, 'n_estimators': 51} 0.1403051099604759
```

- Selected through cross-validation, the optimal learning rate is approximately 0.9, the max depth is 2, and number of estimators be 51.
- The test MSE for the optimal set of parameters is 0.140, close to the performance of random forest.

```
Accuracy: 0.7922077922077922
```

# Comparison : MSE

| | Model Name | Train MSE | Test MSE |
|---|---|---|---|
| 0 | OLS | 0.1586 | 0.1392 |
| 1 | Best Subset Selection | 0.1807 | 0.1661 |
| 2 | Forward Selection | 0.1807 | 0.1661 |
| 3 | PCR | 0.1757 | 0.1439 |
| 4 | PLS | 0.1586 | 0.1392 |
| 5 | Ridge | 0.1589 | 0.1387 |
| 6 | Lasso | 0.1602 | 0.1392 |
| 7 | Decision Tree | 0.1582 | 0.1530 |
| 8 | Random Forest | 0.0233 | 0.1411 |
| 9 | Boosting | 0.1356 | 0.1359 |

- Best subset selection and forward selection shows same training and testing mse, and select the same features as ‘pregnancies’, ‘glucose’, ‘blood pressure’.
- Random forest has the best performance in training set and maintains a low MSE on testing set.
- This represents the capacity to deal with overfitting and multiple features of random forest.
- Boosting appears to provide a great balance between training and test performance with the lowest test MSE.

# Comparison : Accuracy Score

| | Model Name | Accuracy |
|---|---|---|
| 0 | Logistic Regression | 0.7987 |
| 1 | PCR | 0.7792 |
| 2 | PLS | 0.8052 |
| 3 | Ridge | 0.8052 |
| 4 | Lasso | 0.7987 |
| 5 | Decision Tree | 0.7208 |
| 6 | Random Forest | 0.8312 |
| 7 | Boosting | 0.7922 |

- Decision Tree has the lowest accuracy score
- Random Forest has the highest accuracy score
- Comparing the results from MSE and accuracy score, we see some equivalences between MSE and classification error.
- Given the classification problem nature of this problem, we will choose accuracy score as a guideline for selecting optimal models.

# Further Explanation: Age Classification

We have selected Random Forest as our optimal predictive model. Our aim is now to refine its performance further. Given that individuals of varying ages may have distinct risk factors for diabetes, our objective is to identify which attributes significantly influence the likelihood of diabetes within different age groups.

For the purpose of our analysis, we are adopting the World Health Organization age classification: individuals aged 0-24 will be considered 'young adults', those between 25-44 will be 'adults', ages 45-60 will fall under 'middle-aged', and anyone above 60 will be categorized as 'senior'.

| | Age Group | Top 1 feature | Top 2 | Top 3 |
|---|---|---|---|---|
| 0 | Youth (0-24) | Glucose | BMI | DiabetesPedigreeFunction |
| 1 | Young (25-44) | Glucose | BMI | DiabetesPedigreeFunction |
| 2 | Middle (45-60) | Glucose | DiabetesPedigreeFunction | BMI |
| 3 | Elder (61-100) | BMI | Glucose | Pregnancies |

# Reference

World Health Organization (WHO)

# Q&A