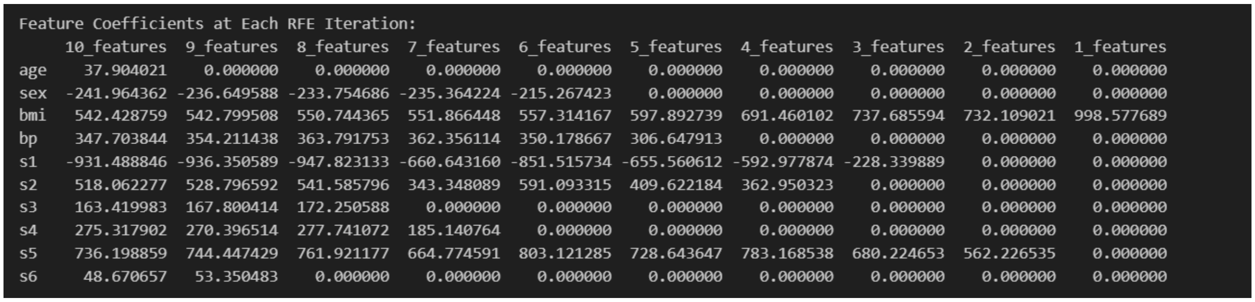
Introduction :

This report summarizes the findings from Recursive Feature Elimination (RFE) applied to the **Diabetes dataset**.
RFE was used to identify the most relevant features affecting **diabetes progression**.
The analysis includes feature ranking, comparison with other selection methods, **and key dataset insights**.

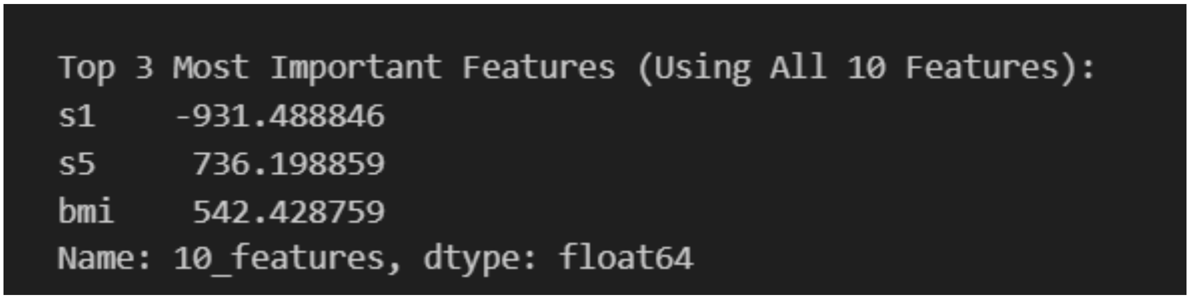
Feature Coefficients at Each RFE Iteration:

The table below shows how feature importance changed as features were eliminated step by step



```
Feature Coefficients at Each RFE Iteration:
     10_features  9_features  8_features  7_features  6_features  5_features  4_features  3_features  2_features  1_features
age    37.904021    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
sex  -241.964362 -236.649588 -233.754686 -235.364224 -215.267423    0.000000    0.000000    0.000000    0.000000    0.000000
bmi   542.428759  542.799508  550.744365  551.866448  557.314167  597.892739  691.460102  737.685594  732.109021  998.577689
bp    347.703844  354.211438  363.791753  362.356114  350.178667  306.647913    0.000000    0.000000    0.000000    0.000000
s1   -931.488846 -936.350589 -947.823133 -660.643160 -851.515734 -655.560612 -592.977874 -228.339889    0.000000    0.000000
s2    518.062277  528.796592  541.585796  343.348089  591.093315  409.622184  362.950323    0.000000    0.000000    0.000000
s3    163.419983  167.800414  172.250588    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
s4    275.317902  270.396514  277.741072  185.140764    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
s5    736.198859  744.447429  761.921177  664.774591  803.121285  728.643647  783.168538  680.224653  562.226535    0.000000
s6     48.670657   53.350483    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000    0.000000
```

Top 3 Most Important Features:

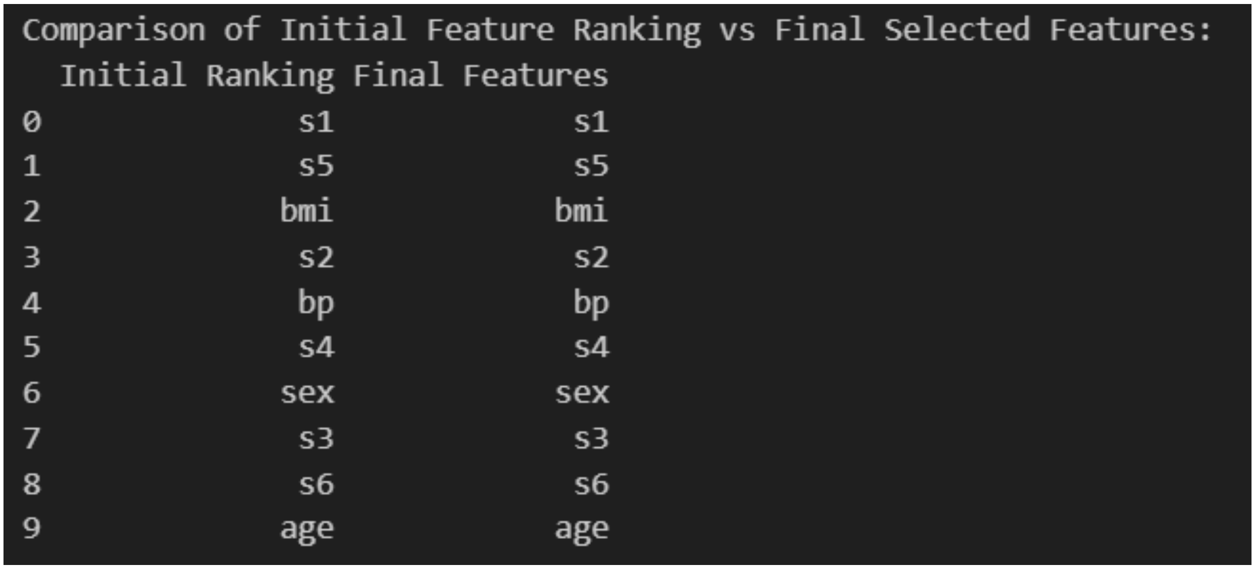


```
Top 3 Most Important Features (Using All 10 Features):
s1    -931.488846
s5     736.198859
bmi    542.428759
Name: 10_features, dtype: float64
```

What we can infere is

- s1 has the strongest negative impact , Higher values slow down diabetes progression.
- s5 is highly positive , Higher values increase diabetes progression significantly.
- bmi (Body Mass Index) is a major contributor , Obesity is a strong factor in diabetes risk.

Comparison of Initial vs. Final Feature Selection :



```
Comparison of Initial Feature Ranking vs Final Selected Features:
  Initial Ranking Final Features
0              s1             s1
1              s5             s5
2             bmi            bmi
3              s2             s2
4              bp             bp
5              s4             s4
6             sex            sex
7              s3             s3
8              s6             s6
9             age            age
```

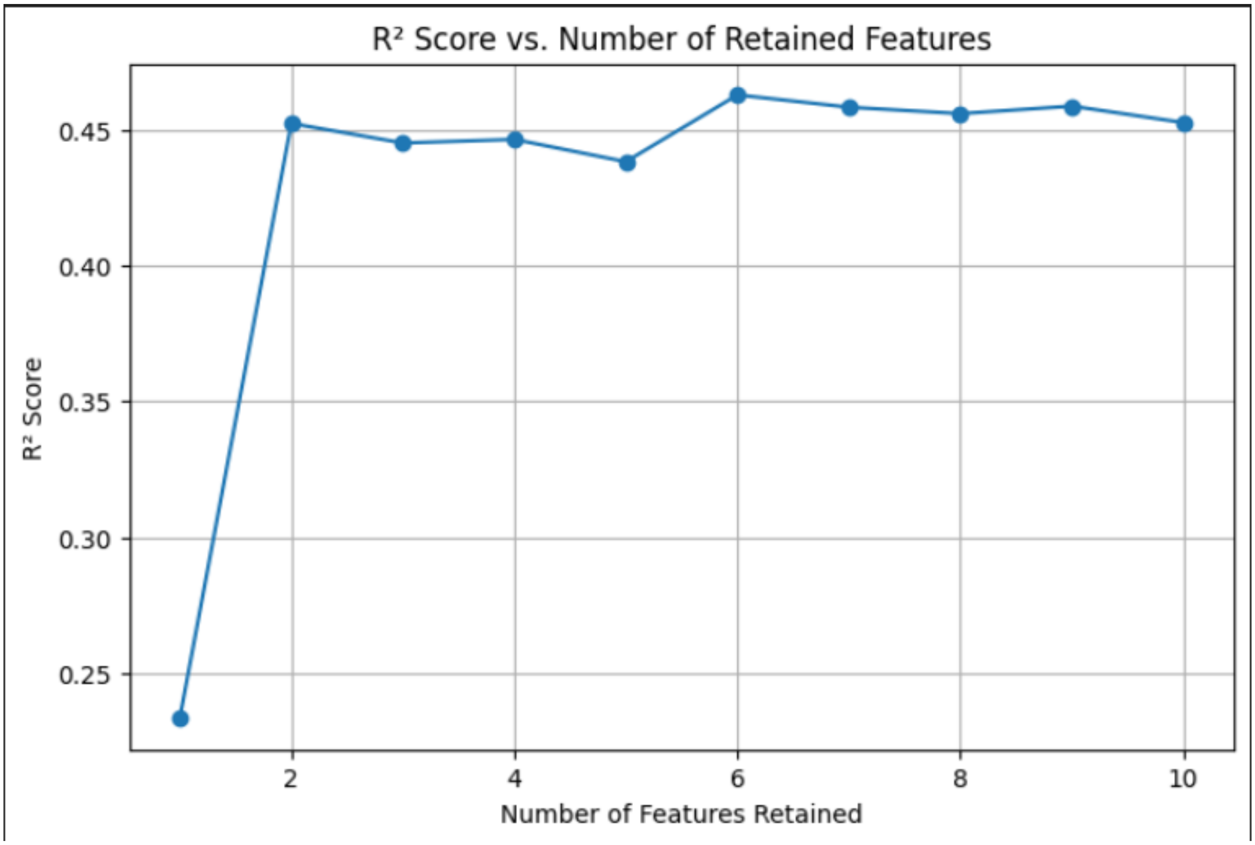
- RFE retained all 10 features because removing any feature led to an R² drop greater than 0.01.
- The strongest predictors (s1, s5, and bmi) remained highly ranked even after RFE.
- Age and sex had the lowest impact but were still retained since they contributed meaningfully.

Conclusion :

RFE retained all 10 features because removing any one caused an R² drop greater than 0.01, meaning each feature contributed enough to keep. BMI, s5, and s1 were the strongest predictors, showing that body mass and blood serum levels are key factors in diabetes progression. BP had moderate importance, while age and sex had the least impact but were still retained as they added some value.

Unlike LASSO, which might have forced some features to zero, RFE ranked features without removing any, ensuring all useful predictors stayed. Diabetes progression is influenced by multiple factors working together, not just one, making it important to keep all relevant features for better predictions.

This graph shows the changes in r2 when number features are retained



Key Findings of mine :

The R^2 score, which measures how well the model explains the variability in diabetes progression, was 0.4526 when using all 10 features, indicating that the model explains 45.26% of the variance. However, using the 0.01 R^2 improvement threshold, RFE determined that only 6 features were necessary, as removing any more led to a significant drop in performance. This suggests that while all features contribute to some extent, selecting the most relevant ones improves model efficiency without sacrificing accuracy. By reducing the number of features from 10 to 6, we maintain a strong predictive ability while ensuring a more interpretable and optimized model.