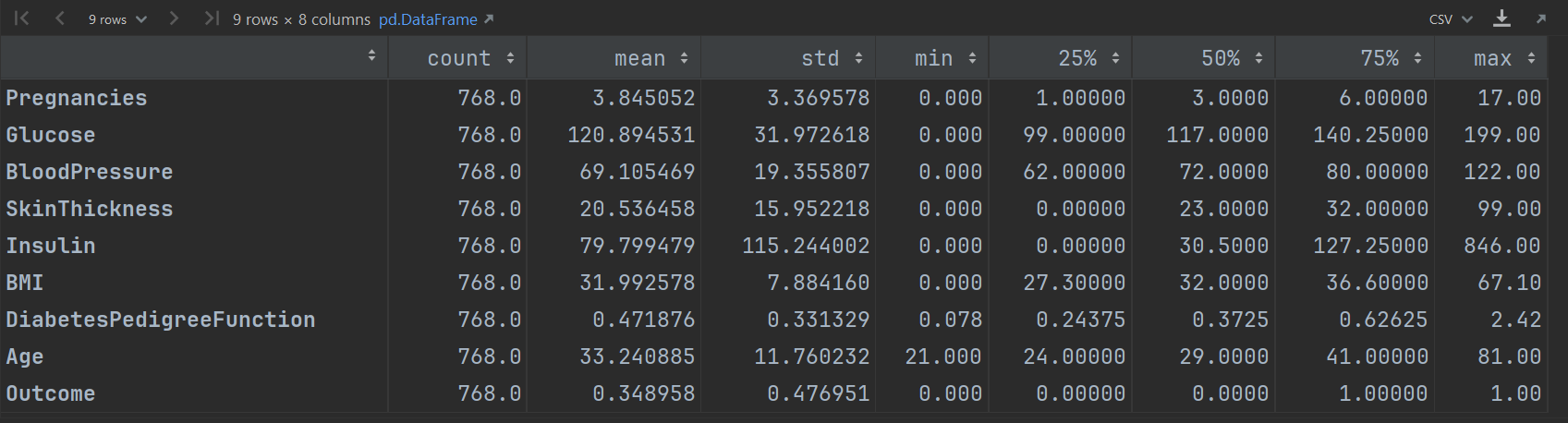
Diabetes dataset

----------------

Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

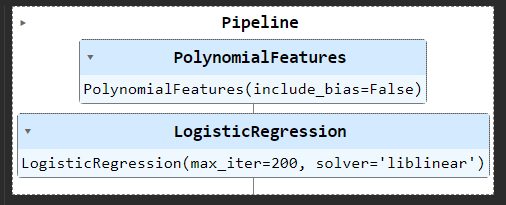


We will build and analyse the basics and implementation of several model validation techniques, mentioned below:

* Hold Out Validation
* K-fold Cross-Validation.
* Stratified K-fold Cross-Validation
* Leave One Out Cross-Validation.

1. Holdout Validation Approach - Train and Test Set Split

The holdout validation approach refers to creating the training and the holdout sets, also referred to as the 'test' or the 'validation' set. The training data is used to train the model while the unseen data is used to validate the model performance. We will use the 70:30 ratio split for the diabetes dataset.



**Accuracy scores for each fold: 0.7489177489177489**

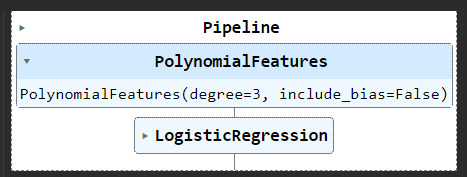
**Mean cross-validation score: 74.8918%**

**MSE: 0.2510822510822511**

1. K-fold Cross-Validation

In k-fold cross-validation, the data is divided into k folds. The model is trained on k-1 folds with one fold held back for testing. This process gets repeated to ensure each fold of the dataset gets the chance to be the held back set. Once the process is completed, we can summarize the evaluation metric using the mean or/and the standard deviation

We will use 10-fold cross-validation for our problem statement.



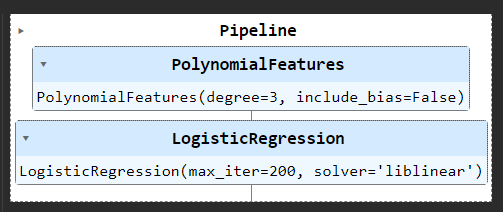
**Accuracy scores for each fold: [0.74025974 0.71428571 0.74025974 0.72727273 0.75324675 0.72727273 0.80519481 0.77922078 0.71052632 0.71052632]**

**Mean cross-validation score: 74.0807%**

**MSE: 0.18229166666666666**

1. Stratified K-fold Cross-Validation

Stratified K-Fold approach is a variation of k-fold cross-validation that returns stratified folds, i.e., each set containing approximately the same ratio of target labels as the complete data.



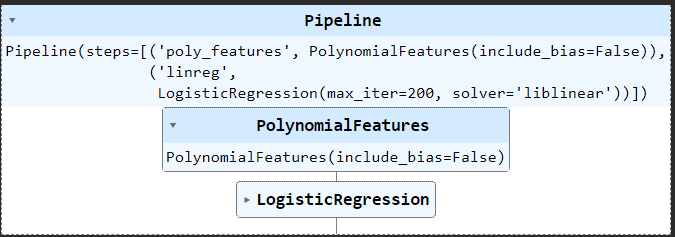
**Accuracy scores for each fold: [0.734375 0.73046875 0.75 ]**

**Mean cross-validation score: 73.8281%**

**MSE: 0.18229166666666666**

1. 5-fold cross validation (default)

In k-fold cross-validation k-5 is the default



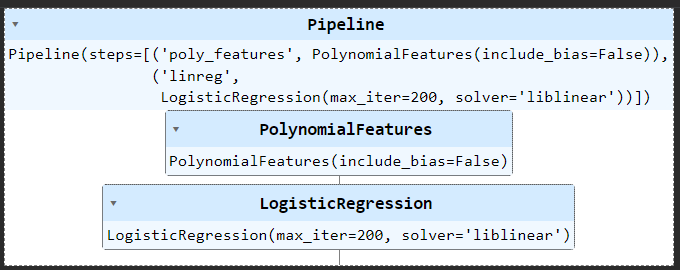
**Accuracy** scores **for each fold: [0.74675325 0.77922078 0.75974026 0.78431373 0.79084967]**

**Mean cross-validation score: 77.2176%**

**MSE: 0.19270833333333334**

1. Leave One Out Cross-Validation (LOOCV)

LOOCV is the cross-validation technique in which the size of the fold is “1” with “k” being set to the number of observations in the data. This variation is useful when the training data is of limited size and the number of parameters to be tested is not high.

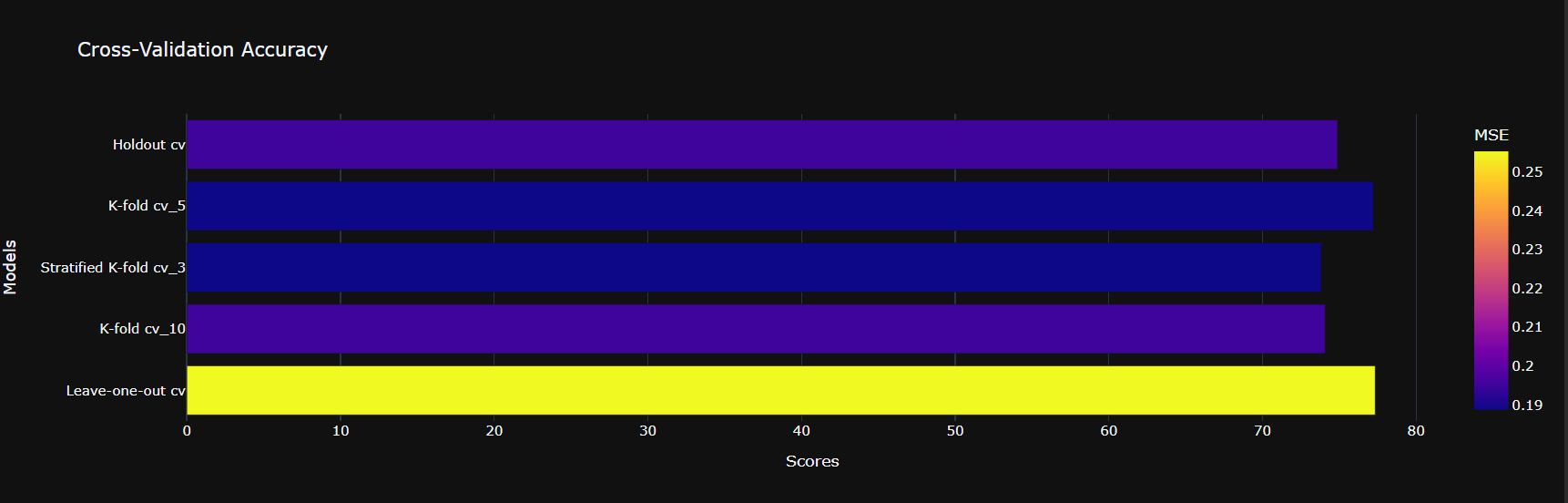


**Accuracy scores for each fold: [1. 1. 1. 1. 1. 1. 0. 0. 1. 0. 1. 1. 1. 1. 1. 0. 0. 0. 1. 0. 1. 1. 1. 0.** **1. 1. 1. 1. 1. 1. 1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.**

**0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 0. 0. 0. 1. 0. 0. 1. 1. 0. 1. 0. 1]**

**Mean cross-validation score: 77.3438%**

**MSE: 0.19270833333333334**

****

The best performance of the models for this dataset Leave One Out Cross-Validation with **Mean cross-validation score: 77.3438%** and **MSE: 0.255411.**

Even the MSE is higher than other models, the cross-validation result is more representative because it represents the performance of the system on the 80% of the data instead of just the 20% of the training set.

We can conclude that the cross-validation technique improves the performance of the model and is a better model validation strategy. The model can be further improved by doing exploratory data analysis, data pre-processing, feature engineering, or trying out other machine learning algorithms instead of the logistic regression algorithm we built in this guide.

Hi Jay,

Good work, yes the MSE is a good approach to compare models, and alternative is to use ‘cross\_val\_score’ or both, where you could come out for a different assumption.

Hi Radhika,

Good Work, I am trying to figure out which K model or yours has better CVscores K=10 or K=5, but you have missed the information when you have cut and paste the ‘5 – fold cross validation’

Hi John,

Good work!

I came to the same conclusion and our values are close, even our datasets are different.