Report 2

# **Summarize**:

|  |
| --- |
|  |

# **Feature Selection**

Feature Selection in python is the process where you automatically or manually select the features in the dataset that contribute most to your prediction variable or output in which you are interested.

As not all the features selected are helpful to the model and some may even damage its performance, so we chose the right features carefully for these reasons

1. Improve the accuracy
2. Reduce overfitting
3. It enables the machine learning algorithm to train faster.
4. Reduces the complexity and makes it easier to interpret

Featurewiz: can automatically detect if the problem is regression or classification and it’s used for automatically creating and selecting important features in your dataset that will create the best model with higher performance.

* SULOV: stands for searching for uncorrelated list of variables
* Recursive XGBoost: After selecting the features with less coloration and high mutual score, it finds the best features among the remaining ones.

Explanation: it was proved , because input numerical and output categorical but regression input numerical and output numerical.

# **Hyperparameter Tuning:**

In SVM Model and logistic regression model use three hyperparameter C, max\_iter and penalty.

We've got different accuracy and performance example in SVM:

* When 'C': 0.01, 'max\_iter': 1000, 'penalty': 'l2' 🡪accuracy model = 0.423
* When 'C': 10, 'max\_iter': 10000, 'penalty': 'l2' 🡪 accuracy model = 439
* This is mean accuracy better when ‘C’ is big but this is lead to overfitting

We've got best hyperparameter in logistic regression mode when C=1.0092528860766845, max\_iter=1000, 'penalty': 'l2'

In Decision Trees Model use three hyperparameter criterion, splitter, max\_depth

We've got different accuracy and performance example in Decision Trees:

* When criterion: 'gini', splitter: 'best', max\_depth: 50 🡪 accuracy model = 0.849
* When criterion: 'entropy', splitter: 'random', max\_depth: 70 🡪accuracy model= 0.831
* When criterion: 'gini', splitter: 'random', max\_depth: 60 🡪 accuracy model = 0.831
* When criterion: 'entropy', splitter: 'best', max\_depth: 80 🡪 accuracy model = 0.858
* When criterion: 'gini', splitter: 'best', max\_depth: 40 🡪 accuracy model = 0.853

# **Conclusion**

In this phase we use regression(the previous phase) pre Processing but don’t use feature selection because in this phase different the previous phase(input is numerical and output categorical)

Use many hyperparameter for train model and select best hyperparameter in SVM, logistic and decision tree it is proved by training many model and analysis result

In SVM model select one vs one model because this is more better from one vs all but it is more Complicated