STATLOG (SHUTTLE)

Comparative Study of Machine Learning Algorithms on Real Datasets

By

Victoire Djimna Noyum

May 11, 2020

CONTENTS

Introduction

Machine learning (ML) is the study of computer algorithms that improve automatically through experience for prediction. However, nowadays we are faced with the question of selecting appropriate ML methods to solve specific problems. The objective of this study is to compare the ML algorithms Logistic regression, KNN and SVM using the Statlog (Shuttle) dataset. To do so, we will first prepare our data, perform a basic visualization and statistical description of the data to better understand it, then we will perform data preprocessing, design prediction methods and finally we will compare the models in terms of accuracy, training and testing time.

1 Data preparation

1. **Download data**: from the UCI ML repository, using this link Statlog (Shuttle), we download the dataset **Statlog (Shuttle)**.

2. **Data description**:

   - Dataset has 58 000 samples and nine (09) features.
   - The label have seven (07) classes which has been coded as follows :

1	Rad Flow
2	Fpv Close
3	Fpv Open
4	High
5	Bypass
6	Bpv Close
7	Bpv Open

   - Approximately 80% of the data belongs to class 1.
   - Dataset is divided by two: the training dataset (43 500 samples) and the testing dataset (14 500 samples).

3. **Merging data** : we merge the training and testing dataset to do the basic visualization and statistical summary of data.

2 Basic visualization

To understand the data, we did some plots. First, we did the scatter plot for visualization of each feature and the bar-chart for label.

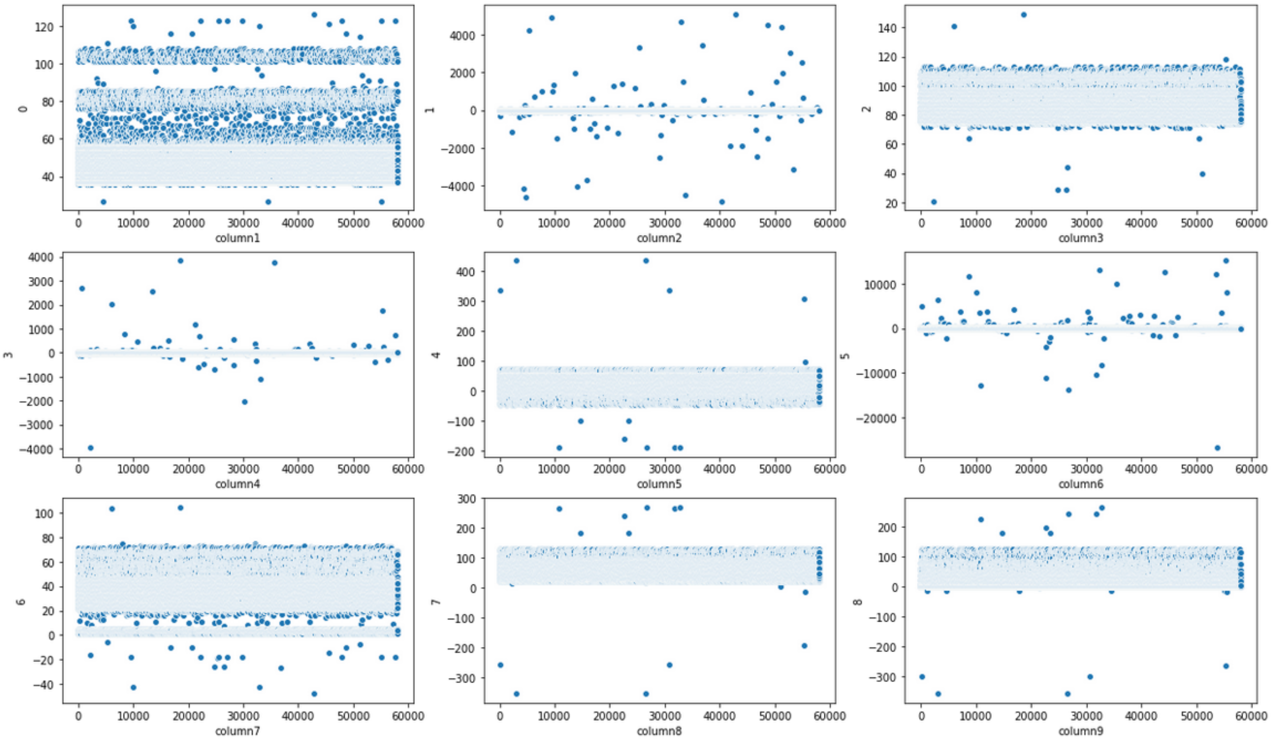


Figure 1: Scatterplot of features

According to Figure 1, the distribution of the components of each variable is normal because the data are uniformly distributed between the minimum and maximum value of each variable. So, we do not have the outliers.

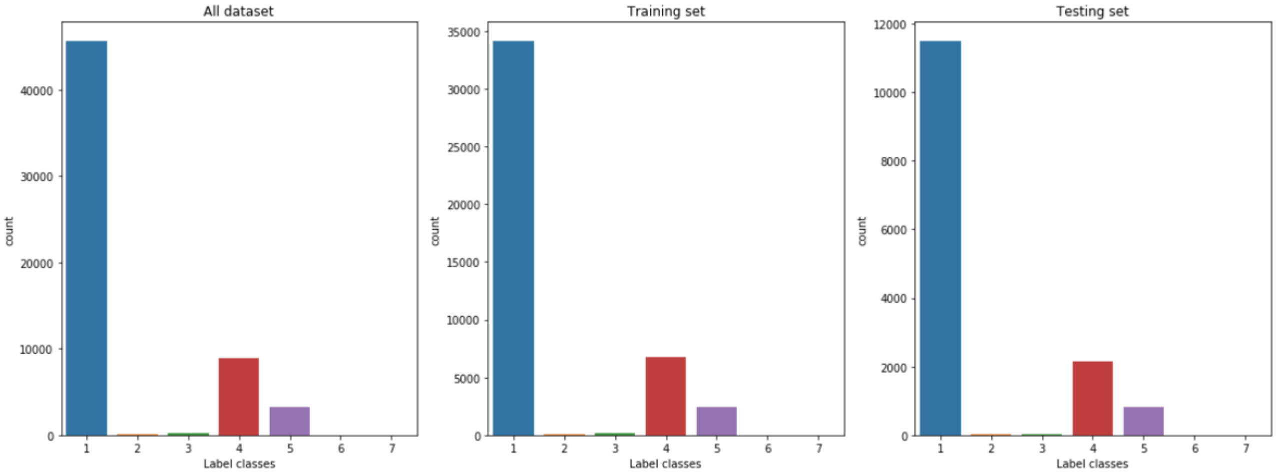
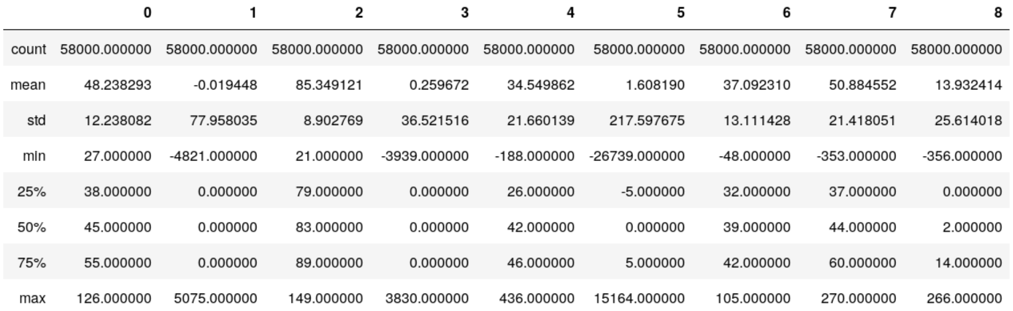


Figure 2: Bar-chart of each class of the label

In Figure 2, we have three graphs of the count of the number of each class of the label. The first one is for all the dataset, the second one, for the training set, and the last one, for the testing set. We can easily observe that there is approximately the same amount of each class in both the training and testing sets. Therefore, the data division was done in a balanced way.

3 Summary statistic

Table 1: Statistic description of the data



	0	1	2	3	4	5	6	7	8
count	58000.000000	58000.000000	58000.000000	58000.000000	58000.000000	58000.000000	58000.000000	58000.000000	58000.000000
mean	48.238293	-0.019448	85.349121	0.259672	34.549862	1.608190	37.092310	50.884552	13.932414
std	12.238082	77.958035	8.902769	36.521516	21.660139	217.597675	13.111428	21.418051	25.614018
min	27.000000	-4821.000000	21.000000	-3939.000000	-188.000000	-26739.000000	-48.000000	-353.000000	-356.000000
25%	38.000000	0.000000	79.000000	0.000000	26.000000	-5.000000	32.000000	37.000000	0.000000
50%	45.000000	0.000000	83.000000	0.000000	42.000000	0.000000	39.000000	44.000000	2.000000
75%	55.000000	0.000000	89.000000	0.000000	46.000000	5.000000	42.000000	60.000000	14.000000
max	126.000000	5075.000000	149.000000	3830.000000	436.000000	15164.000000	105.000000	270.000000	266.000000

The statistic description of data given by Table 1 show us that:

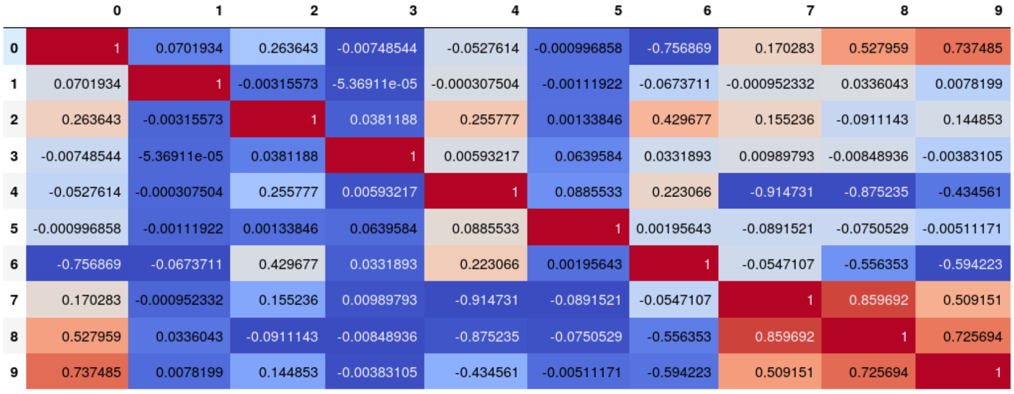
- The count of each attribute is 58 000. So, there are no missing values in the data.
- We have big differences between the standard deviations of the features.
- The value of the mean of attribute 8 is higher than that of its median. Therefore, the distribution of this feature is skewed to the right.

4 Training of Machine Learning Algorithms using Logistic Regression, KNN, SVM and Results

Here, we start with some pre-processing to design our prediction methods using Logistic Regression (LR), KNN, SVM, and then we do the box-plot visualization to illustrate a comparison among the algorithms in terms of accuracy, training and testing times.

4.1 Design of Ml Algorithms models

- **Correlation between the variables:** To optimize the training of the models, we build the correlation matrix to see if some features have the strong linear relationship.



	0	1	2	3	4	5	6	7	8	9
0	1	0.0701934	0.263643	-0.00748544	-0.0527614	-0.000996858	-0.756869	0.170283	0.527959	0.737485
1	0.0701934	1	-0.00315573	-5.36911e-05	-0.000307504	-0.00111922	-0.0673711	-0.000952332	0.0336043	0.0078199
2	0.263643	-0.00315573	1	0.0381188	0.255777	0.00133846	0.429677	0.155236	-0.0911143	0.144853
3	-0.00748544	-5.36911e-05	0.0381188	1	0.00593217	0.0639584	0.0331893	0.00989793	-0.00848936	-0.00383105
4	-0.0527614	-0.000307504	0.255777	0.00593217	1	0.0885533	0.223066	-0.914731	-0.875235	-0.434561
5	-0.000996858	-0.00111922	0.00133846	0.0639584	0.0885533	1	0.00195643	-0.0891521	-0.0750529	-0.00511171
6	-0.756869	-0.0673711	0.429677	0.0331893	0.223066	0.00195643	1	-0.0547107	-0.556353	-0.594223
7	0.170283	-0.000952332	0.155236	0.00989793	-0.914731	-0.0891521	-0.0547107	1	0.859692	0.509151
8	0.527959	0.0336043	-0.0911143	-0.00848936	-0.875235	-0.0750529	-0.556353	0.859692	1	0.725694
9	0.737485	0.0078199	0.144853	-0.00383105	-0.434561	-0.00511171	-0.594223	0.509151	0.725694	1

Figure 3: Correlation Matrix

The highest value is **0.85**. Using the fact that, we don't know the name of the features and its importance, this value is not enough to cancel one feature. So, we keep the nine (09) features.

- **Shuffle of data:** we saw above that the training and testing sets had divided in a balanced way. To have realistic results, we keep the merging data and we shuffle the data ten (10) times. Then, the general accuracy will be the mean of the ten (10) accuracies.
- **Normalization of data:** as we saw in the Table 1, according to the standard deviation values, the features have different spread. Then, the model can focus to this different spread to train which is not right. Therefore to avoid this, we normalize the data and then, all the feature have the spread in the same range.
- **Apply the Grid Search:** in order to tune the parameter of each model to come up with the best combination of parameters.

4.2 Models Comparison in term of accuracy

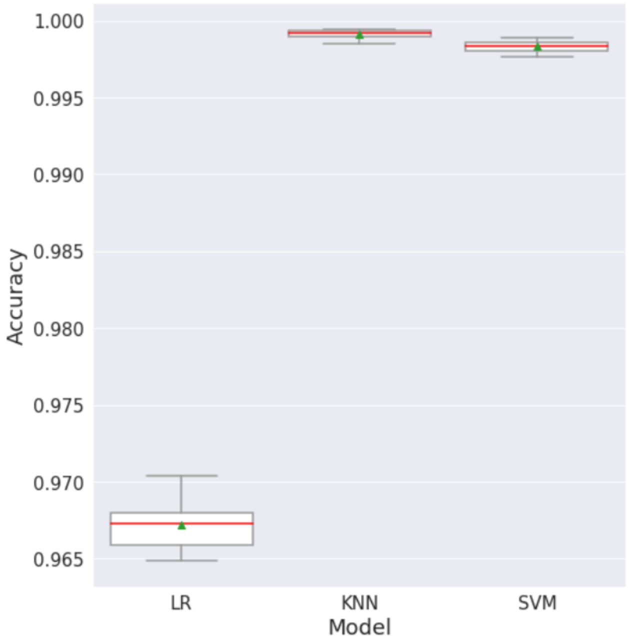


Figure 4: Box-plot of models accuracies

Table 2: Mean accuracies of each algorithm

	LR	KNN	SVM
Accuracy (%)	96.72	99.91	99.83

Figure 4 show us that KNN model is the most efficiently with the average accuracy of **99.91%** (Table 2). The second one is SVM model with 99.83% and the last one is LR model with 96.72%.

4.3 Models Comparison in term of training time

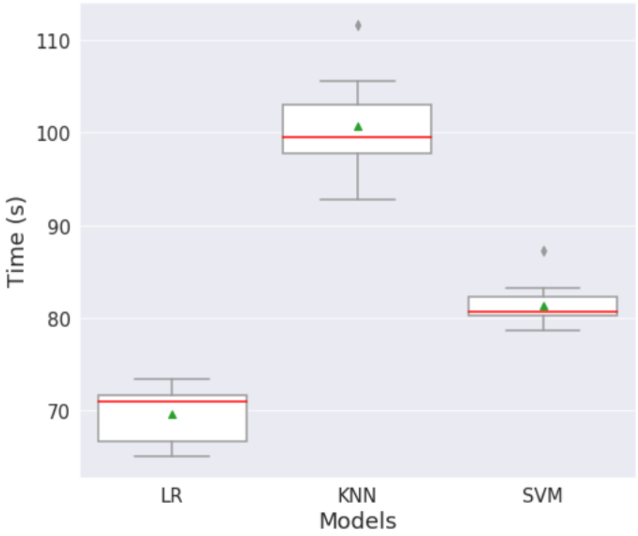


Figure 5: Box-plot of model training time

Table 3: Mean training time of each algorithm

	LR	KNN	SVM
Time (s)	69.66	100.70	81.35

According to Figure 5, KNN model take more time to train. It takes on average is 100.70s followed by the SVM model which takes 81.35s. The fastest model is LR with 69.66s.

4.4 Models Comparison in term of testing time

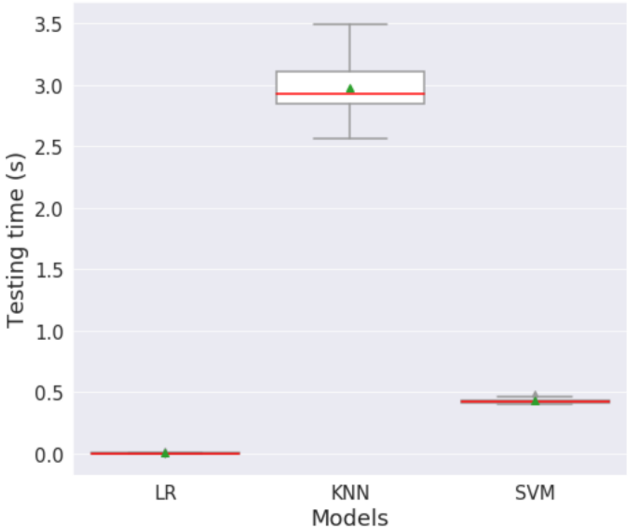


Figure 6: Box-plot of model testing time

Table 4: Mean testing time of each algorithm

	LR	KNN	SVM
Time (s)	0.001	2.97	0.42

Figure 6 show that, KNN model take more time to test. It takes on average is 2.97 s followed by the SVM model which takes 0.42 s. The fastest model is LR with 0.001 s.

In summary:

- KNN model gives the best average accuracy but takes more time to train and test.
- SVM model also gives an excellent accuracy average with a deviation of 0.08% from that given by the KNN model and takes on average less than 19.35 s than the KNN model to train and test.
- Logistic regression model gives the lowest average accuracy with a large deviation from the other two models. On the other hand this model takes less time to train and test.

Having these results, we can conclude that the best ML algorithm for this dataset is the **SVM model** in terms of performance and accuracy because it takes much less time to train than the KNN model and the difference in accuracy is just 0.08%.

Conclusion

The goal of this study was to compare the ML algorithm Logistic regression, KNN and SVM using the Statlog (Shuttle) dataset. We prepared the dataset, we did the visualization process to better understand the data, and we did the pre processing process, tuned the hyper parameters to design the prediction models. We can conclude that the SVM model is the best ML algorithm to perform the prediction using the Statlog (Shuttle) dataset.