# Impact of Noise on Performance of Machine Learning Algorithms

Ankit Samantaray
Trinity College Dublin
Dublin, Leinster
samantaa@tcd.ie

Raksha Kodandaramu
Trinity College Dublin
Dublin, Leinster
kodandar@tcd.ie

Ravi Yadav
Trinity College Dublin
Dublin, Leinster
yadavra@tcd.ie

## 1 INTRODUCTION

The cause of noise is often related to problems in the entry of data, error in computer and various other sources. Noise present in dataset leads to variability in the data which in turn leads to data swings. The amount of swing in data can be used to find the amount of noise. We have gone through the works of different researchers and their works based on the impact of Noise([2, 7]). This research paper measures the difference in performance of Naive Bayes and Random Forest on the Wine Quality data-set before and after introduction of Noise in the dataset. In Section 2, the work related to noise and its impact has been mentioned. The methodology adopted to achieve the results includes Data-set Analysis, Introduction of noise in Data-set, Feature engineering etc. are discussed in section 3. The section 4, consists of experimental results observed from the performance of different machine learning algorithms before and after introduction of noise in the Wine dataset, our inferred conclusions. In Section 5, we have mentioned 'Limitation and Future Scope' of our works.

**KEYWORDS:**Feature Engineering, Noisy Data-set, Data Cleaning, Naive Bayes, Random Forest, Wine Data-Set, Attribute Noise, Introduction of noise, Impact of noise and irrelevant features.

## 2 RELATED WORK

The wine quality dataset [1] we have used in this assignment had no noise. We went through the literature, to study the type of noise [4] namely 'Class Noise' and 'Attribute Noise' we also read the analysis of class and label noise[4].

Further, we studied, the development of techniques to deal with label noise[3]. The learning process uses feature selection[6]. Zhu and Wu [8] analyzed the impact of noise on attribute noise and its impact on learning models.

## 3 METHODOLOGY

To answer the research question, we analyzed the machine learning algorithms (Naive Bayes and Random Forest) with variation in amount of noise on the white wine dataset. We expect that noise should reduce the accuracy as it contains false values. The actual performance of machine learning model not only depends on choice of algorithm, but also on quality of dataset. Therefore, we analyzed the impact of noise by introducing noise in the dataset by two different approaches i.e. by adding noise in attributes and by adding irrelevant features in the dataset.

### 3.1 Dataset

We have used a famous multiclass Wine Quality Data-Set from UCI Machine Learning library which consisted of the red and white variants of the Portuguese "Vinho Verde" wine[1]. In this project, we have used only white wine dataset with 4898 instances. The data contains no missing values and consists of only numeric data. The data-set has 12 attributes and 'Quality' being the target variable which indicates the wine quality on scale of 0-10. The Figure 1 shows the count of the white wine samples with respect to the Quality (frequency distribution) with highest no. of samples belonging to scale 6.

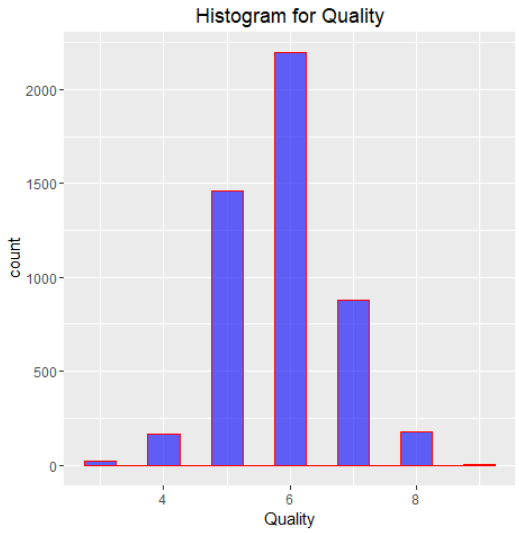


**Figure 1: Frequency Distribution of Quality feature**

### 3.2 Pre-processing

Initially, the 'Wine Quality' dataset had 10 classes with numeric values from 0 to 10. But, due to the imbalance frequency in class, we categorized into different grades of quality with 0- 5 being very bad (Poor) and 6-7 being considerable (Average) and above 8 being superior (Good) quality of wine. We had no missing data in this dataset, so there has been no data-imputation techniques involved. There was difference in scale for factors like fixed acidity, volatile acidity, therefore we used standardization on out dataset to get the values at same scale. Further, we found feature correlation as given in Figure 2.

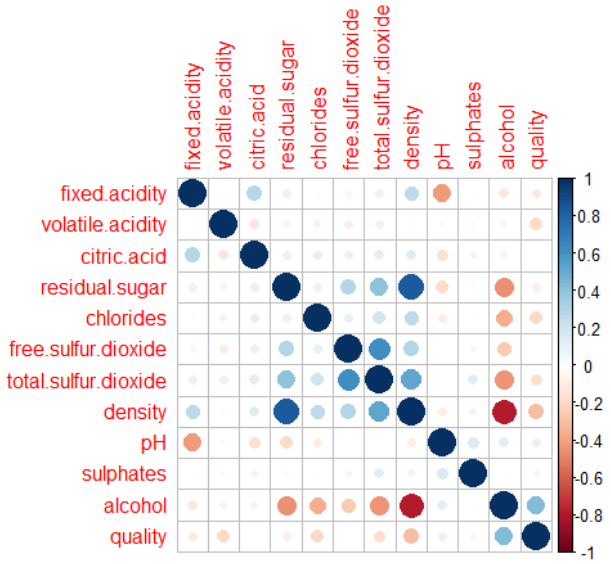


**Figure 2: Correlation Matrix**

## 3.3 Feature Selection

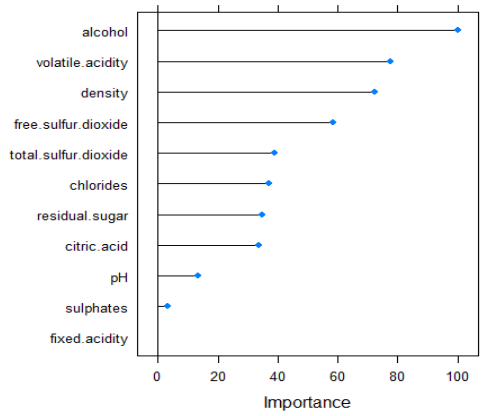


**Figure 3: Feature Importance**

The presence of irrelevant feature is considered to be noise. Figure 3 shows the feature importance. Therefore, after Pre-processing, we performed feature selection by recursive backward method with 10-cross validation and found decrease in accuracy. Hence, we selected all 12 feature for training the model for both machine learning algorithms.

## 3.4 Noise Introduction: Attribute Noise

In dataset, we can introduce noise by 2 methods i.e. *attribute noise* and *class noise*. In this project, we have focused on introduction of attribute noise via two methods.

*3.4.1 Add Noise to Feature Space with irrelevant new feature.* We have added extra irrelevant features with random values(generated from Rand Function) as seen in Figure 4 . This data is not a proper information. A machine learning model can either ignore or accept this data which leads to increase in complexity.

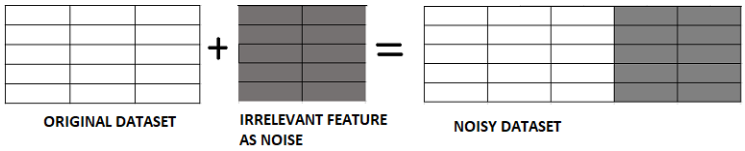


**Figure 4: Addition of Noise to Feature Space**

*3.4.2 Add Noise to Original Data (same Dimension).* We introduced the noise in the existing dataset using Gaussian Method. The attribute's value is replaced by random number following gaussian distribution. A diagrammatic representation is given in Figure 5. For the experiment, we have introduced different levels of noise( 10%,25%, and 50%)

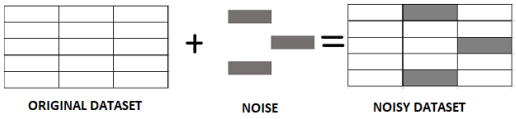


**Figure 5: Addition of Noise in Raw Data**

## 3.5 Algorithms

The machine learning algorithms: Naive Bayes(**NB**) and Random Forest (**RF**), analyzed the dataset with variation of (10%, 25% and 50%) noise and also evaluated by adding irrelevant new features (considered as noise). For each learning model, we performed cross validation(10-fold) for different hyper-parameters and selected the parameters which is contributing to high accuracy. The Figure 6 shows the selected hyper-parameters for each implementation performed in the project. We have used '*dplyr*' and '*caret*' packages of R language for the below implementations.

We have considered the original dataset as *W*, and split it in 70/30 with training set *X* and test *Y* respectively. We built 2 classifier, *C1* and *C2* based on Naive Bayes and Random Forest respectively. The accuracy of model was recorded in *X_C1* and *_C2* for the baseline comparison.

The first set of implementation was based on *addition of irrelevant feature* as noise in feature space as shown in Figure 5 . We introduced the noise by **adding 4 irrelevant features** (as they were randomly generated and not related to the dataset) in the original dataset and lets assume it to be *W2*(noisy dataset) as discussed in section 3.4. Just like the *W* dataset, we split it into training set *X'* and test set *Y'*. The classifier we again trained on X' set and recorded the results.

| ML Algorithm | Datasets | Optimized Hyperparameter Mtry |
|---|---|---|
| Random Forest | Original Dataset | 3 |
| | Dataset with addition of irrelevant features | 6 |
| | Dataset with 10% noise | 3 |
| | Dataset with 25% noise | 3 |
| | Dataset with 50% noise | 3 |

| ML Algorithm | Datasets | Optimized Hyperparameters | |
|---|---|---|---|
| | | Use kernel | Adjust |
| Naïve Bayes | Original Dataset | True | 5 |
| | Dataset with addition of irrelevant features | True | 6 |
| | Dataset with 10% noise | True | 5 |
| | Dataset with 25% noise | True | 5 |
| | Dataset with 50% noise | True | 5 |

**Figure 6: Optimized Parameter Value for different training model with various dataset**

The second and final set of implementation was executed by introducing noise randomly in original dataset with same dimension. In this approach, we experimented by **varying the percentage of noise** in $W$(Original dataset) and used as noisy dataset $W3$. We split the noisy dataset into a training set $X'$ and test set $Y'$. We learn classifier $C1'$ and $C2'$ based on NB and RF respectively for noise training dataset X'. Finally, we repeat the above steps for different level of noise i.e. 10%, 25%, and 50% noise and measure the accuracy's of $C1'$ and $C2'$.

## 4 RESULTS & DISCUSSION

In this section, we have discussed the impact of noise on developed machine learning models. Naive Bayes and Random Forest are the two machine learning algorithms which were implemented to study noise impact on multi-variant dataset. Due to the class imbalance, which leads to low accuracy, we transformed the target from 10 class variants into 3 class variants i.e. wine quality as poor, average, or good. The Figure 7 & 8 summarizes the evaluation of classification algorithm.

In the first method, we analyzed the impact of attribute noise by adding irrelevant features in dataset. We introduced almost 45% of noise by adding 5 features as noise in the original dataset. There was a decrease of 9% and 17% in accuracy for naive Bayes and random forest respectively against the original dataset as shown in Figures 7. This implies that the addition of irrelevant features made the algorithm too complex which shows that too many input features(irrelevant) do not properly regularize. This may end up in memorizing the noisy features which we have introduced.

In the second method, we have evaluated machine learning algorithm considering original dataset as benchmark with noise 'n'= 10%, 25% and 50% noise to the dataset. We calculated classification

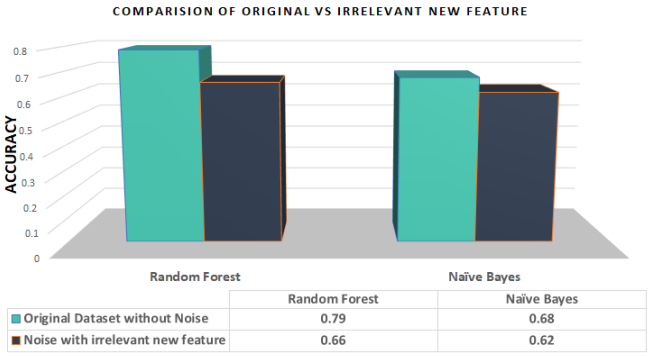


**Figure 7: Accuracy of ML Alogrithm with irrelevant new features as noise.**

accuracy of machine learning algorithm for each level of noise and recorded 5% and 9.3% decrease in accuracy at 50% noise for NB and RF classifier with respect to original dataset as seen in Figure 8.

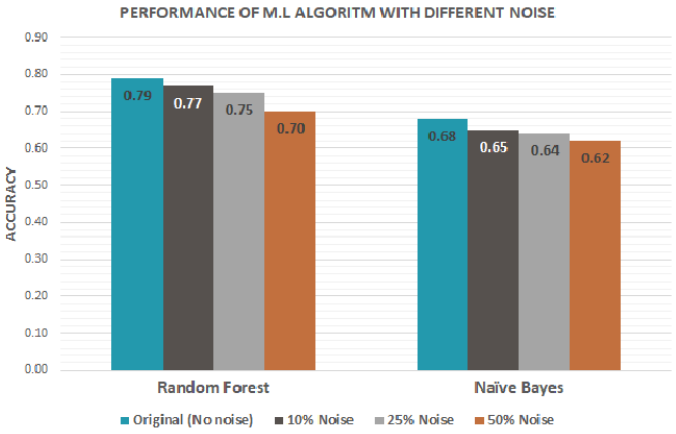


**Figure 8: Accuracy of ML Algorithm with various level of noise in Original Dataset**

In Summary, the result demonstrates that the

- highest accuracy was obtained from the learning model trained with original dataset. This implies the presence of noise affects accuracy of machine learning model.
- the lowest accuracy was present in model with highest noise. This means that in noisy dataset, the pre-processing techniques can increase accuracy in comparison to un-processed noisy data-set.
- there was an increase in the processing time as noise made the learning model complex and over-fit.

## 5 LIMITATION AND OUTLOOK

As the classes were not evenly distributed for the quality of white wine, it is possible that algorithm might leave some classes while training the model. Moreover, there were no missing data and correlated features. Therefore, we had to introduce noise to analyze the

impact. In future, we can evaluate the machine learning algorithm for

- noise in target feature i.e. 'Class Noise'.
- attribute noise only in training set with original test set.
- attribute noise only in test set with original training set.

## 6 ACKNOWLEDGEMENT

This analysis was conducted as part of the 2018/19 Machine Learning module CS7CS4/CS4404 at Trinity College Dublin [5].

## REFERENCES

[1] [n. d.]. https://archive.ics.uci.edu/ml/datasets/wine+quality
[2] Abhinav Atla, Rahul Tada, Victor Sheng, and Naveen Singireddy. 2011. Sensitivity of different machine learning algorithms to noise. *Journal of Computing Sciences in Colleges* 26, 5 (2011), 96–103.
[3] Benoît Frénay, Ata Kabán, et al. 2014. A comprehensive introduction to label noise.. In *ESANN*.
[4] Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2014), 845–869.
[5] Douglas Leith Joeran Beel. [n. d.]. *Noise detection in classification problems.* Ph.D. Dissertation. Trinity College Dublin, School of Computer Science and Statistics.
[6] Ahmad Abu Shanab, Taghi M Khoshgoftaar, and Randall Wald. 2011. Impact of noise and data sampling on stability of feature selection. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, Vol. 1. IEEE, 172–177.
[7] Hui Xiong, Gaurav Pandey, Michael Steinbach, and Vipin Kumar. 2006. Enhancing data analysis with noise removal. *IEEE Transactions on Knowledge and Data Engineering* 18, 3 (2006), 304–319.
[8] Xingquan Zhu and Xindong Wu. 2004. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review* 22, 3 (2004), 177–210.