Effect of Using Numerical Data Scaling on Supervised Machine Learning Performance

Mona Ali Mohammed
Department of Computer Science, Faculty of Sciences, University of Omar Almukhtar, Albaida, Libya

Effect of Using Numerical Data Scaling on Supervised Machine Learning Performance

Abstract

Before building machine learning models, the dataset should be prepared to be a high quality dataset, we should give the model the best possible representation of the data. Different attributes may have different scales which possibly will increase the difficulty of the problem that is modeled. A model with varying scale values may suffers from poor performance during learning. Our study explores the usage of Numerical Data Scaling as a data pre-processing step with the purpose of how effectively these methods can be used to improve the accuracy of learning algorithms. In particular, three numerical data Scaling methods with four machine learning classifiers to predict disease severity were compared. The experiments were built on Coronavirus 2 (SARS-CoV-2) datasets which included 1206 patients who were admitted during the period between June 2020 and April 2021. The diagnosis of all cases was confirmed with RT-PCR. Basic demographic data and medical characteristics of all participants was collected. The reported results indicate that all techniques are performing well with Numerical Data Scaling and there are significant improvement in the models for unseen data. lastly, we can conclude that there are increase in the classifier performance while using scaling techniques. However, these methods help the algorithms to better understand learn the patterns in the dataset which help making accurate models.

Keywords

Feature Scaling, supervised machine learning, Support Vector Machines classifier, Naïve Bayes classifier, Decision Trees classifier, k-nearest neighbors classifier.

تأثير استخدام تحجيم مقياس البيانات العددية على تعلم الاله بإشراف

منى علي محمد

الملخص

قبل إنشاء نماذج التعلم الآلي ، يجب إعداد مجموعة البيانات لتكون مجموعة بيانات عالية الجودة ، وبأفضل تمثيل ممكن للبيانات. قد يكون للخواص المختلفة مقاييس مختلفة مما قد يزيد من صعوبة صياغة المشكلة. قد يعاني النموذج من ضعف الأداء أثناء التعلم مع استخدام مقاييس مختلفة للقيم. تعرض دراستنا استخدام تحجيم مقياس البيانات العددية كخطوة للمعالجة المسبقة

للبيانات بهدف بيان مدى فعالية هذه الأساليب في تحسين دقة خوارزميات التعلم. على وجه الخصوص ، تمت مقارنة ثلاث طرق لتحجيم مقياس البيانات العددية مع أربعة خوارزميات تصنيف من خوارزميات للتعلم الآلي للتنبؤ بخطورة الامراض. تم بناء التجارب على مجموعة بيانات Coronavirus 2 (SARS-CoV-2) والتي شملت 1206 مريض خلال الفترة بين يونيو 2020 وأبريل 2021. تم تأكيد تشخيص جميع الحالات باستخدام RT-PCR, جمعت البيانات الأساسية والخصائص الطبية لجميع المرضى. تشير النتائج إلى أن جميع الخوارزميات المستخدمة تعمل بشكل جيد مع تحجيم مقياس البيانات العددية وهناك تحسن كبير في ادائها في بيانات الاختبار. أخيرًا ، يمكننا أن نستنتج أن هناك تحسن في أداء خوارزميات التصنيف أثناء استخدام تحجيم مقياس البيانات العددية. الخلاصة تساعد هذه الأساليب الخوارزميات على فهم أفضل للتعلم الأنماط مما يساعد في صنع نماذج دقيقة.

الكلمات المفتاحية : تحجيم الخواص، التعلم الآلي بإشراف ، مصنّف آلات المتجهات الداعمة ، مصنف Naïve Bayes ، مصنف أشجار القرار ، مصنف K لأقرب الجيران.

1. Introduction

All machine learning(ML) algorithms take input data to generate the output. The input data remains in a tabular form consisting of rows (instances or observations) and columns (variables or attributes), and these attributes are often known as features. The feature is the basic building block of real world datasets, it represents a measurable piece of data. the feature is generally a numeric representation of an aspect of real world phenomena or data [1][2].

Generally, ML systems fit mathematical notations to the existing data to derive some insights, it improves automatically using experiments and existing data. Supervised, unsupervised, and semi-supervised ML are all types of ML. Typically, Supervised ML deals with data sets that contain both inputs and the corresponding desired class labels. Learning classifiers to address classification problems is a fundamental issue in data mining. The classification algorithms category are used within supervised learning to predict the class of an unseen instance[1][3].

The models take features as input. The quality of data can vary significantly and has an immense effect on model performance. So, we should give the model the best possible representation of the data, otherwise, it may not give a good accuracy[2][4]

The quality of a dataset's features can be improved in the pre-processing stage. Preprocessing data is necessary to reduce the impact of data distortion or outliers and increase the predictive performance of the model for unseen data[5][4]. Data Scaling is a recommended pre-processing step while working with a different range of independent variables or features of data. Feature Scaling is the name of the technique that transforms and normalizes numerical input variables [2].

ML approaches are an effective way to increase the adoption of information technology in hospitals, it improves automatically using experiments and existing data, which makes it suitable to predict individualized disease risk and clinical decision making[1][6][7][8].

This study explores the usage of Numerical Data Scaling as a data pre-processing step with the purpose of how effectively these methods can be used to improve the accuracy of learning algorithms. In particular, three numerical data Scaling methods with four ML classifiers to predict Coronavirus 2 (SARS-CoV-2) severity will be compared. The classifiers used to run the experiments and evaluate the results using common classification algorithms (Naive

Bayes, K-Nearest Neighbors(KNN), Decision Tree(DT), and Support Vector Machines(SVM) [1][9] [10][11][12].

This paper is structured as follows, Section 2 describes related work and a literature review. The material and methods that have been used in this work proposed in Section 3. The results of experiments will be introduced in section4, Finally, we will present our conclusion in Section 5

2. Literature review

2.1 Related work

Since Most of the published literature represented their study output in terms of the accuracy of the machine learning (ML) algorithm very limited work investigated data scaling methods [13].

One of previous research,[14] stated that min-max scaling has good performance in terms of speed, accuracy, and quantity of support vectors in Support Vector Machines(SVM). On the other hand, [15]provide on many ML algorithms standardization scaling before training. When applied normalization between 0 to 1 K-Nearest Neighbors (KNN) achieved accuracies 87.3% on the dataset based on shape features and leaf color histograms to identify plants[16]. In addition, the accuracy obtained is 98.8% on Naïve Bayesian while applying normalization between 0 to 1 regarding plant identification using leaf features dataset [17].

Another study identified plants based on the type of leaf venation using SVM. The author was applied min-max normalization and the resulting accuracy was 77.57% [18].

Different normalization methods showed that the performance of ML algorithms and the selection of normalization methods are interconnected. Also, their study shows that SVM has the maximum accuracy and Naïve Bayes has the best performance in terms of accuracy and lowest fitting times[19].

Data normalization has used on the dataset extracted from the leaf venation feature, four ML algorithms include KNN, Naïve Bayesian, ANN, and SVM with Radial Basic Function (RBF) kernels, and linear kernels applied for the normalized dataset. The results show that the min-max normalization technique with SVM that uses the RBF kernel can provide the best performance results .addition to that, the KNN algorithm is quite stable compared to SVM and Artificial Neural Network(ANN) while Naïve Bayesian has the most stable performance against the use of min-max normalization techniques as well as standardization [20].

More robust ML algorithms such as XGBoosting(XGB), Linear Regression(LR), DT, and Random Forest (RF) with scaling methods such as Standard Scaler, MinMax Scaler, Max Abs Scaler, Robust scaler, and Quantile Transformer use on heart failure patient datasets[21] . Their study demonstrated that Random Forest (RF) has a higher performance with Standard Scaler and Robust Scaler. Additionally, the performance of Decision Tree (DT) remained unchanged with scaling[22].

Recently, six data scaling methods with eleven ML algorithms evaluated to detect patients with heart diseases dataset, The result shows that classification and regression Trees, along with Robust Scaler or Quantile Transformer, outperform all other ML algorithms

Finally, Many studies bolstered the effect of data scaling techniques on different ML algorithms [13][20][19]. furthermore, one of the main challenges associated with ML is choosing the appropriate scaling method.

2.2 The Machine Learning Techniques that are used in this study

Supervised machine learning(ML) deals with datasets that contain both inputs and the corresponding desired outputs. The classification algorithms category is used within supervised learning when the outputs are discrete to a limited set of values.

In this study, we have implemented four supervised ML algorithms (Naive Bayes,
K-Nearest Neighbors(KNN), Decision Tree(DT), and Support Vector Machines(SVM)).

2.2.1 Support Vector Machines(SVM)

It is a powerful technique used by a supervised learning approach for classification. It plots the data items as a space split into categories based on statistical learning frameworks. Then, it finds the hyperplane with the maximum distance between the target data points[10][12][23][24] [25].

Computing the (soft-margin) SVM classifier amounts of minimizing an expression of the form

$$\frac{1}{n}[\sum_{i=1}^{n} max(0.1 - yi(W^{t}x_{i} - b))] + ʎ||W||^{2}$$

2.2.2 Naïve Bayes

It is known as one of the best classification algorithms and creates fast ML. it is a simple technique for constructing classifiers models based on the Bayes Theorem that assigns class labels as vectors of some n features, where the class labels are drawn from some finite set [26][27][28][29].

Naive Bayesian classifier works on assuming the effect of an attribute value on a given class is independent of the values of the other attributes as follows:

$$X = (x1,x2,x3,\dots\dots\dots.,xn)$$

$$P(y\backslash x1,x2,\dots..xn) = \frac{p(x1\backslash y)p(x2\backslash y)\dots\dots p(xn\backslash y)p(y)}{p(x1)p(x2)\dots\dots p(xn)}$$

$$\mathrm{P(y\backslash x1,x2,\dots..xn)} \propto \mathrm{p(y)}\prod_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{p(xi\backslash y)}$$

$$\mathrm{y = argmaxy\ p(y)}\prod_{\mathrm{i=1}}^{\mathrm{n}}\mathrm{p(xi\backslash y)}$$

2.2.3 Decision Trees(DT)

It is one of the easiest tools to decision systems and easy to understand, it is powerful and popular tools to build classification and regression models[24][30][31][32]. It uses a tree structure which that provides sequential nonlinear analysis in algorithmic relationship and their possible consequences, including outcomes. The tree consists of nodes that symbolize a dataset's features, branches symbolize the decision rules, and leaves symbolize the class, It built top-down from a root node and involves partitioning the data into subsets that contain instances with similar values which break down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The split of a node attempts to minimize the impurity of the node. If a split is unable to achieve any improvement in terms of reducing impurity, the node is not split and is declared as a leaf node. If a split is can reduce impurity, then the split providing the maximum reduction in impurity is selected and two branches are formed, forming two new nodes [24][31][33] .

2.2.4 K-Nearest Neighbors(KNN)

One of the simplest and most common classifiers, As the name implies, KNN finds the closest K (number of neighbors) nearest neighbor points to the target point. Then, it predicts the output of the target point [34][35][36]. By default, KNN function employs Euclidean Distance which can be calculated with the following equation:

$$D(p,q) = \sqrt[2]{(p1-q1)^2 + (p2-q2)^2 + \cdots \dots \dots \dots \dots .(pn-qn)^2}$$

3. Materials and Methods

3.1 The Used Dataset: our dataset includes 1206 patients who were admitted during the period between June 2020 and April 2021. The diagnosis of all cases was confirmed with RT-PCR. Basic demographic data and medical characteristics of all participants was collected. Laboratory investigations include hematological parameters, coagulation parameters, liver function tests (LFT), and renal function tests (RFT).

3.2 Data pre-processing: In ML, the performance of the model depends on data preparation and data handling. In contrast, before building ML models the used dataset should be a high quality dataset to run the experiments and evaluate the results of different techniques. we have preprocessed our dataset in three steps. In these subsections, we will describe these steps.

3.2.1 Data Cleaning

In the data cleaning step, missing data will be handled by applying Missing Value Imputation for Classification Tasks (MVICT) method which has been proposed in[37]. MVICT has focused on imputation using mean, middle, and mode of the available values which belong to the same class of the missing value. In our dataset, the missing values will be replaced by the mean of the available values which belong to the same class of the missing value.

After handling the missing values, it is often necessary to deal with categorical features. Many ML algorithms require numerical data in nature. This means that it must be one of the elements of data pre-processing is categorical variables encoding which convert categorical data to a numerical form. there are many techniques designed for this purpose. the most common technique is One Hot Encoding. In One Hot Encoding , each category apply to a vector that contains 1 and 0 denoting the presence or absence of the feature. The number of vectors depends on the number of categories for features. It is the most common correct way to deal with categorical data where no relationship exists between categories[38][39][40].

In our study, the used dataset has an attributes stored as categorical values, and to convert it to a numerical form we will use One Hot Encoding method. Out of 18 columns, only 2 had a string value and the rest of them had a numerical input. Therefore, to convert the string input into a numerical input One Hot Encoding has been applied.

3.2.2 Feature Extraction

In fact, not all of the features contribute to the definition or determination of class labels. In theory, too many features that may adversely affect the model performance are redundant or even irrelevant. it is important to select the most useful features to improve the quality of the feature set in many ML tasks. In practice, however, increasing the size of the feature may slow down the running time of model training, and require a large amount of system memory which reduces the performance of an algorithm[9][41][42][6]. According to the physicians, we selected the most significant features, in addition, the 9 selected variables with (p value < 0.05).

Besides the Feature Selection, we used the generating features to create new features. it allows us to use less complex models which are faster to run and easier to understand and maintain.

3.2.3 Feature Scaling

the scale and distribution of the dataset is different for each variable, so it is often necessary to transform the different numerical features to fall within a similar range. In our dataset we have 10 columns including the last column which will be the class that we are trying to predict with models algorithms. However, C reactive protein(CRP), Lymphocytes(Lymph), Lactate dehydrogenase (LDH), D Dimer, Urea, and Red blood cells(RBC), these were numerical features.

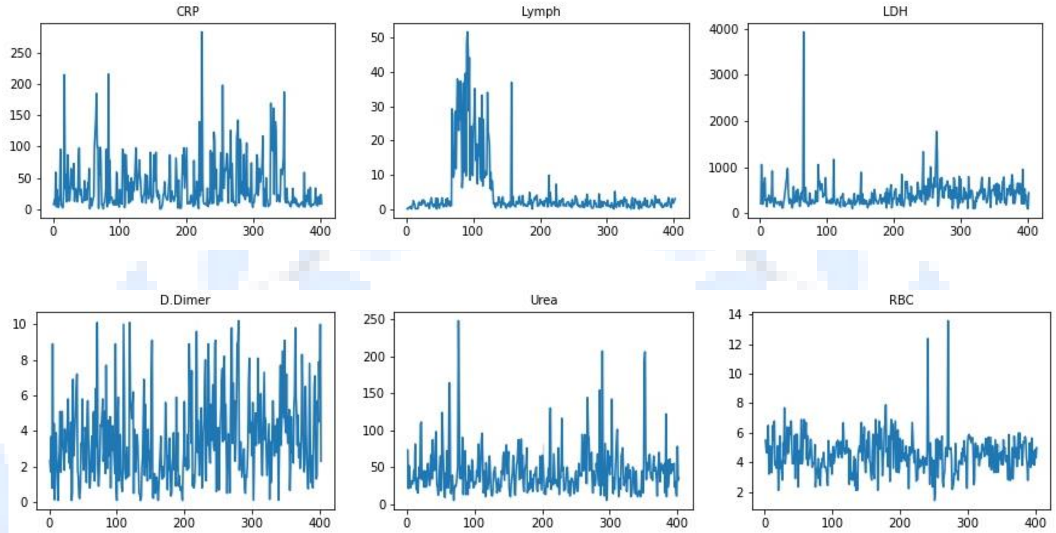

Figure1: Numerical columns are not from the same scale

The graph above clearly shows that the 6 numerical columns are not from the same scale (figure 1).

In our study, we have used 3 popular Scaling techniques to build the prediction model and compared the results to find out the best Feature Scaling that can be used for the prediction from this kind of problem and data sets

3.2.3.1 Min-Max Scaling

This is the most used technique to scale the data to a specific range using each feature's minimum and maximum value. By default, the Min-max Scaling technique returns a value between 0 and 1, using the equation:

$$Xscaled = \frac{x - xmin}{xmax - xmin}$$

As it is obvious from figure2, shows numerical features in our data set with min-max Scaling

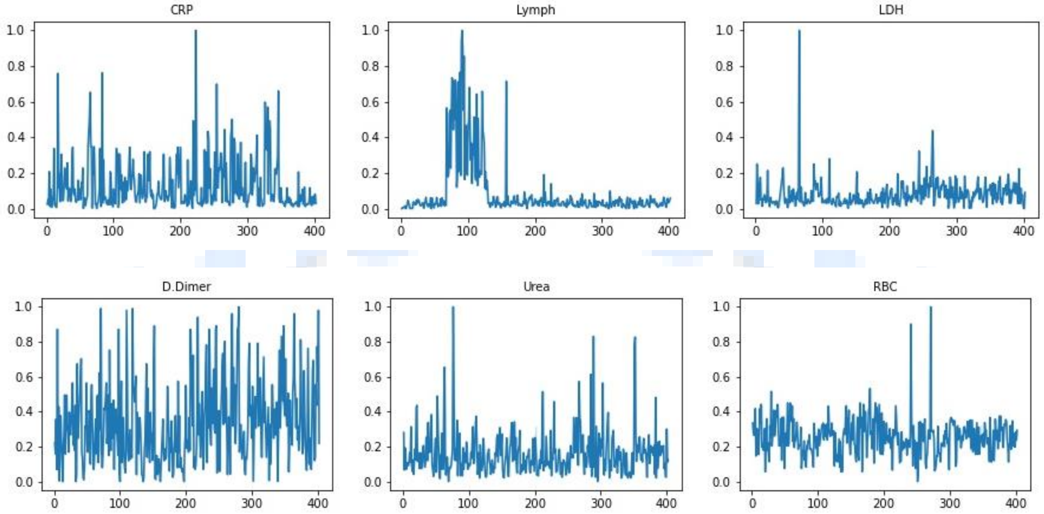

Figure2: Numerical features in our data set with Min-Max Scaling

3.2.3.2 Standardization Scaling

Standardization involves Scaling the features by subtracting the mean and dividing by the standard deviation to shift the distribution, a value is standardized as follows:

$$y = \frac{x - mean}{standard\_deviation}$$

Figure3 shows numerical features in our data set with Standardization Scaling

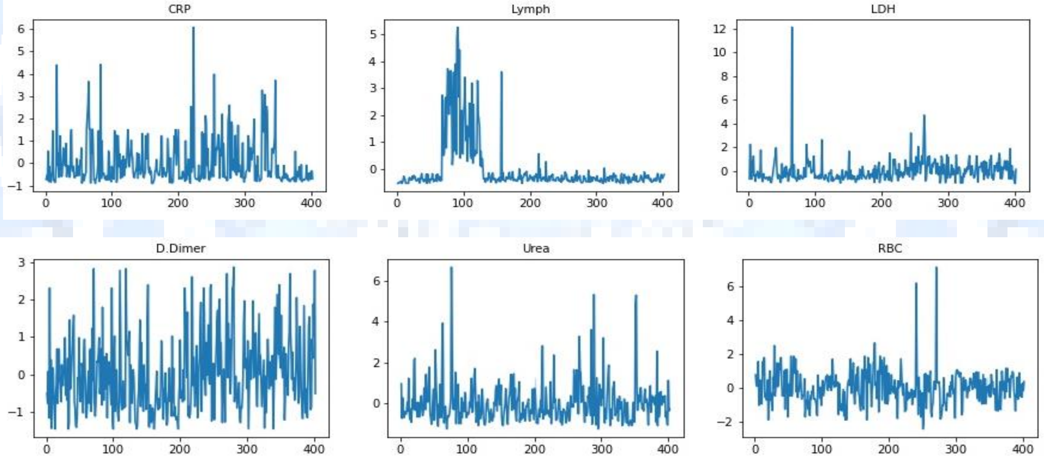

Figure3: Numerical features in our data set with Standardization Scaling

3.2.3.3 Robust Scaling Data

This can be achieved by calculating the median (50th percentile) and the 25th and 75th percentiles. The values of each variable then have their median subtracted and are divided by the interquartile range (IQR) which is the range between the 75th and 25th

$$value = \frac{\text{value} - \text{median}}{(p75 - p25)}$$

As we can see in figure4 numerical features in our data set with robust Scaling

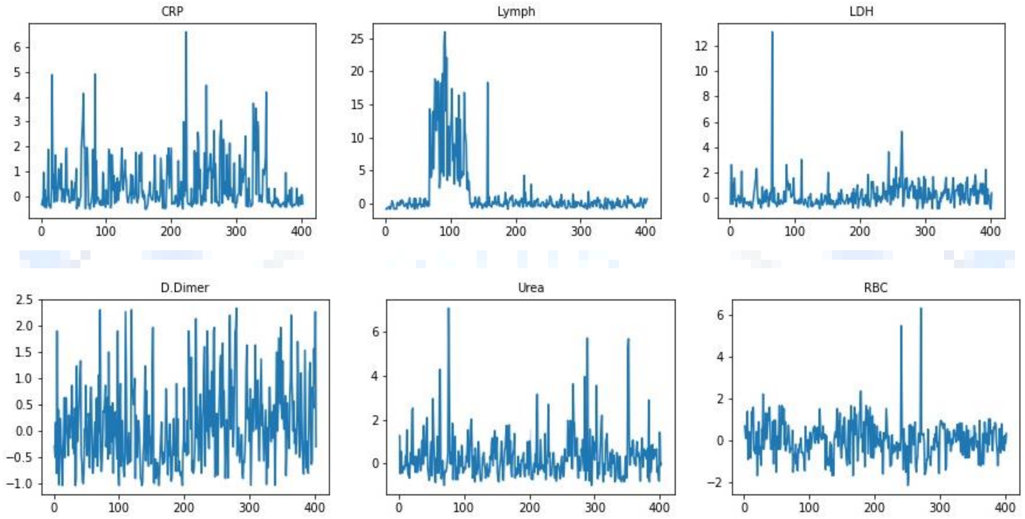

Figure4: Numerical features in our data set with Robust Scaling

3.4 Evaluation Measures

To evaluate the results, we used the confusion matrixes used to evaluate the quality of binary classification or multiple classes. It shows the number of correct and incorrect values which were identified by the classification model compared to the actual target value in the dataset. The matrix is NxN, where N is the number of target classes and each positive (Target) and negative (Non-Target). Accuracy can define as the total number of predictions that were correct. It is the average of precision & recall, where precision = true positives ÷ predicted positives, and recall = true positives ÷ all positives,

However, achieving a good F-measure requires the classifier to have a good precision and a good recall on the positive class.

To achieve our goal, we have used three scale methods in four classification ML. First, we will implement the Support Vector Machines, Naive Bayes, Decision Tree, and K-Nearest Neighbors

without using feature scaling methods in the data pre-processing stage. Second, three scaling methods Min-Max Scaling, Standard Scaling, and Robust Scaling were used for the classification algorithms in the data pre-processing stage. Finally, we measured the improvement level of the models with feature scaling methods.

4. Results and discussion

The experiments have conducted with four classification tasks to predict the severity of SARS-CoV-2 in the patients. In this section, all experimental results with a short discussion will be presented to show how the models performance and stability will vary with three feature scaling methods while using it in the data pre-processing stage

4.1 Support Vector Machines(SVM) performance

A comparison of the SVM performance without using feature scaling methods and with using the three feature scaling methods presents in table 1

Table 1: A summary of the SVM performance with and without using feature scaling methods.

without Scaling	Min-Max Scaling	Standard Scaling	Robust Scaling
78%	87%	92.3%	88%

Figure 5 clearly shows that there is an increase in SVM performance while applying the three feature scaling methods in the data pre-processing stage. Furthermore, the classifier performs the best performance while using standard scaling method

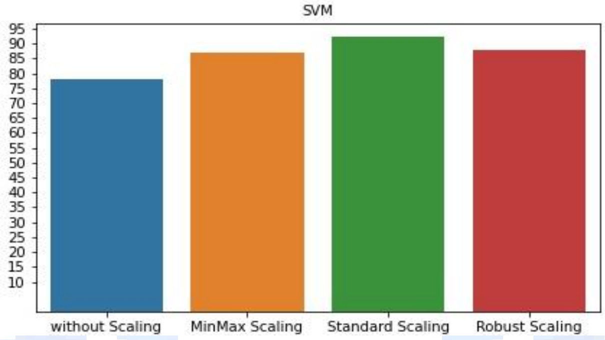

Figure 5 : SVM performance without and with using feature

4.2 Naive Bayes performance

A summary of Naive Bayes performance without using feature scaling methods and with using the three Feature scaling methods are shown in Table3

Table 2: A summary of Naive Bayes performance with and without using feature scaling methods

without Scaling	Min-Max Scaling	Standard Scaling	Robust Scaling
78.5%	81%	81%	81%

By comparing the accuracy of the Naive Bayes algorithm, we can conclude that there are improvement in the classifier accuracy As shown in figure 6. In addition, the improvement does not look different with using the three feature scaling methods.

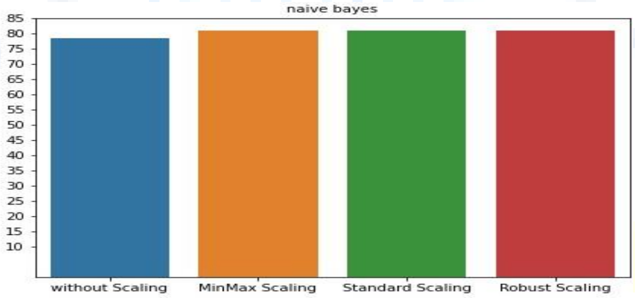

Figure 6 : Naive Bayes performance without and with using feature

4.3 Decision Tree (DT)performance

A comparison of DT classifier performance without using feature scaling methods and with the three feature scaling methods as shown in table3

Table 3: A summary of DT classifier performance with and without using feature scaling methods

without Scaling	Min-Max Scaling	Standard Scaling	Robust Scaling
82%	83.2%	84.3%	84.3%

Figure 7 shows there is an improvement in DT classifier performance by using feature scaling methods at the data pre-processing stage. In addition, the improvement does not look different from Standard Scaling and Robust Scaling which gave the same results.

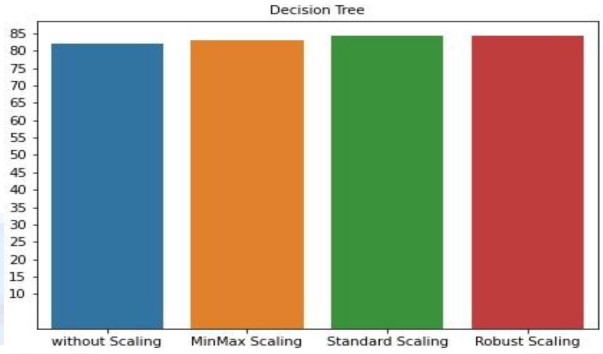

Figure 7 : DT performance without and with using feature

4.4 K-Nearest Neighbors (KNN)performance

Table 4 lists KNN performance without using feature scaling and after applying the three different feature scaling methods

Table4: A summary of KNN performance with and without using feature scaling methods

without Scaling	Min-Max Scaling	Standard Scaling	Robust Scaling
75.2%	76.9%	79.3%	81.8%

From figure8, it has been noticed there are improvement in the KNN classifier accuracy to predict the severity of (SARS-CoV-2) on the patients

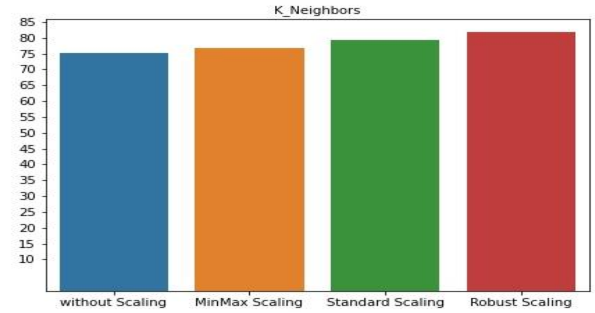

Figure 8 : KNN performance without and with using feature

5. Conclusion

The study evaluated the effect of feature scaling methods at the data pre-processing stage on the accuracy of classifier models. Three methods of feature scaling were used for four classification algorithms at the data pre-processing stage to predict the severity of (SARS-

CoV-2) on the patients. The results are indicating to significant advantages of the presented feature scaling methods. These methods **help the algorithms to better understand and learn the patterns in the dataset which** help making accurate models. However, The SVM classifier with using standard scaling method has outperformed all the other techniques in the four experiments, while the lists KNN showed the lowest performance. Moreover, The best techniques in this case, turned out to be Standard Scaling and Robust Scaling.

Reference

[1] M. M. Abualhaj, A. A. Abu-Shareha, M. O. Hiari, Y. Alrabanah, M. Al-Zyoud, and M. A. Alsharaiah, “A Paradigm for DoS Attack Disclosure using Machine Learning Techniques,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 3, 2022.

[2] D. A. P. Delzell, S. Magnuson, T. Peter, M. Smith, and B. J. Smith, “Machine learning and feature selection methods for disease classification with application to lung cancer screening image data,” *Front. Oncol.*, vol. 9, p. 1393, 2019.

[3] M. Kang and N. J. Jameson, “Machine learning: fundamentals,” *Progn. Heal. Manag. Electron. Fundam. Mach. Learn. Internet Things*, pp. 85–109, 2018.

[4] R. Nisbet, G. Miner, and K. Yale, “Handbook of Statistical Analysis and Data Mining Applications.” Academic Press, Inc., 2017.

[5] M. Kuhn and K. Johnson, *Applied predictive modeling*, vol. 26. Springer, 2013.

[6] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O’Sullivan, “A review of feature selection methods for machine learning-based disease risk prediction,” *Front. Bioinforma.*, vol. 2, p. 927312, 2022.

[7] D. S. W. Ho, W. Schierding, M. Wake, R. Saffery, and J. O’Sullivan, “Machine learning snp based prediction for precision medicine. Front Genet. 2019; 10: 267.” 2019.

[8] Y. Xu, K. Hong, J. Tsujii, and E. I.-C. Chang, “Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries,” *J. Am. Med. Informatics Assoc.*, vol. 19, no. 5, pp. 824–832, 2012.

[9] Ü. Çavuşoğlu, “A new hybrid approach for intrusion detection using machine learning methods,” *Appl. Intell.*, vol. 49, no. 7, pp. 2735–2761, 2019.

[10] T. M. Ma, K. Yamamori, and A. Thida, “A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification,” in *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, 2020, pp. 324–326.

[11] P. Wang, Y. Zhang, and W. Jiang, "Application of K-Nearest Neighbor (KNN) Algorithm for Human Action Recognition," in *2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2021, vol. 4, pp. 492–496.

[12] H. Elaidi, Y. Elhaddar, Z. Benabbou, and H. Abbar, "An idea of a clustering algorithm using support vector machines based on binary decision tree," in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2018, pp. 1–5.

[13] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, 2021.

[14] W. Xu *et al.*, "Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases," *Cancer Cell*, vol. 19, no. 1, pp. 17–30, 2011.

[15] Y. Tang and I. Sutskever, "Data normalization in the learning of restricted Boltzmann machines," *Dep. Comput. Sci. Univ. Toronto, Tech. Rep. UTML-TR-11-2*, pp. 27–41, 2011.

[16] Q. Munisa, "Pengaruh kandungan lemak dan energi yang berbeda dalam pakan terhadap pemanfaatan pakan dan pertumbuhan patin (Pangasius pangasius)," *J. Aquac. Manag. Technol.*, vol. 4, no. 3, pp. 12–21, 2015.

[17] F. R. F. Padao and E. A. Maravillas, "Using Naïve Bayesian method for plant leaf classification based on shape and texture features," in *2015 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*, 2015, pp. 1–5.

[18] A. Ambarwari, Y. Herdiyeni, and I. Hermadi, "Biometric analysis of leaf venation density based on digital image," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 16, no. 4, pp. 1735–1744, 2018.

[19] L. Shahriyari, "Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma," *Brief. Bioinform.*, vol.

20, no. 3, pp. 985–994, 2019.

[20] A. Ambarwari, Q. J. Adrian, and Y. Herdiyeni, “Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification,” *J. RESTI (Rekayasa Sist. Dan Teknol. Informasi)*, vol. 4, no. 1, pp. 117–122, 2020.

[21] K. Balabaeva and S. Kovalchuk, “Comparison of temporal and non-temporal features effect on machine learning models quality and interpretability for chronic heart failure patients,” *Procedia Comput. Sci.*, vol. 156, pp. 87–96, 2019.

[22] K. Balabaeva and S. Kovalchuk, “Post-hoc interpretation of clinical pathways clustering using Bayesian inference,” *Procedia Comput. Sci.*, vol. 178, pp. 264–273, 2020.

[23] S. Dong, B. Tang, and R. Chen, “Bearing running state recognition based on non-extensive wavelet feature scale entropy and support vector machine,” *Measurement*, vol. 46, no. 10, pp. 4189–4199, 2013.

[24] T. Pranckevičius and V. Marcinkevičius, “Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification,” *Balt. J. Mod. Comput.*, vol. 5, no. 2, p. 221, 2017.

[25] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, and M. Dey, “A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews,” in *2020 International Conference on Contemporary Computing and Applications (IC3A)*, 2020, pp. 217–220.

[26] L. Jiang, L. Zhang, L. Yu, and D. Wang, “Class-specific attribute weighted naive Bayes,” *Pattern Recognit.*, vol. 88, pp. 321–330, 2019.

[27] K. L. Priya, M. S. C. R. Kypa, M. M. S. Reddy, and G. R. M. Reddy, “A novel approach to predict diabetes by using Naive Bayes classifier,” in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, 2020, pp. 603–607.

[28] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, “Variable

selection for Naïve Bayes classification," *Comput. Oper. Res.*, vol. 135, p. 105456, 2021.

[29] K. P. Murphy, "Naive bayes classifiers," *Univ. Br. Columbia*, vol. 18, no. 60, pp. 1–8, 2006.

[30] M. Rakhra *et al.*, "Crop price prediction using random forest and decision tree regression:-a review," *Mater. Today Proc.*, 2021.

[31] T. R. Prajwala, "A comparative study on decision tree and random forest using R tool," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 1, pp. 196–199, 2015.

[32] R. Caffrey, "Using the Decision Tree (DT) to Help Scientists Navigate the Access to Space (ATS) Options," in *2022 IEEE Aerospace Conference Proceedings*, 2022.

[33] M. Brijain, R. Patel, M. R. Kushik, and K. Rana, "A survey on decision tree algorithm for classification," 2014.

[34] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of improving k-nearest-neighbor for classification," in *Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007)*, 2007, vol. 1, pp. 679–683.

[35] H. A. Abu Alfeilat *et al.*, "Effects of distance measure choice on k-nearest neighbor classifier performance: a review," *Big data*, vol. 7, no. 4, pp. 221–248, 2019.

[36] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Ann. Transl. Med.*, vol. 4, no. 11, 2016.

[37] M. M. Ali, "Dealing with Missing Values in Classification Tasks," in *Special Issue for 5th International Conference for Basic Sciences and Their Applications (5th ICBSTA, 2022), P:------ , 22-24/10/2022 https://ljbs.omu.edu.ly eISSN 2707-6261*, 2022.

[38] S. Gnat, "Impact of Categorical Variables Encoding on Property Mass Valuation," *Procedia Comput. Sci.*, vol. 192, pp. 3542–3550, 2021.

[39] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *Int. J. Comput. Appl.*, vol. 175, no. 4, pp. 7–9, 2017.

[40] C. T. T. Thuy, K. A. Tran, and C. N. Giap, “Optimize the Combination of Categorical Variable Encoding and Deep Learning Technique for the Problem of Prediction of Vietnamese Student Academic Performance,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, 2020.

[41] S. Kotsiantis, “Feature selection for machine learning classification problems: a recent overview,” *Artif. Intell. Rev.*, vol. 42, no. 1, pp. 157–176, 2011.

[42] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A survey on evolutionary computation approaches to feature selection,” *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, 2015.