Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230152

Salary grades prediction using machine learning

Xiaotian Liu

Detroit Green Technology Institute, Hubei University of Technology, Wuhan, Hubei, 430012, China

2011621109@hbut.edu.cn

Abstract. According to the reference, big data is the kind of data with a great amount of capacity that cannot be analysed with the traditional database system. And the process of analysing big data helps in discovering significant hidden patterns and other regularities. Thus, machine learning algorithms are introduced to replace conventional methods in big data analysis. In this essay, the support vector machine (SVM) algorithm is applied to analyse the salary classification dataset by calculating the impact of all features in annual wage and doing the prediction based on these results. Sense for companies looking for new employees, a reasonable wage can be provided referring to their personal conditions. While for jobhunters, this prediction process can help estimate is the salary provided is worth it or not. Compared to other machine learning methods, the accuracy of SVM can be over 80% or even 90% in classification missions with both low and high-dimensional data. Additionally, varieties of kernel functions can be qualified for specific tasks. However, the running efficiency would decrease as the amount of data for training increases. And the data needs a stricter normalization step to ensure that the result would not be affected by the errors in the dataset, which could be costly. The prediction results can also be harder to interpret in the end.

Keywords: Salary grades prediction, support vector machine, machine learning.

1. Introduction

According to Abhinav Rai, big data is the kind of data with a great amount of capacity that cannot be analysed with the traditional database system. And the process of analysing big data helps in discovering significant hidden patterns and other regularities. The types of big data include structured, unstructured, and semi-structured. Structured ones could be processed, stored, and retrieved in a fixed format. Unstructured ones refer to those “lack any specific form or structure”. And semi-structured data that “although has not been classified under a particular repository (database), yet contains vital information or tags that segregate individual elements within the data” [1]. Thus, machine learning algorithms are introduced to replace the conventional methods in big data analysis.

The salary prediction has been a hot topic in recent years, especially during the epidemic situation. Thousands of jobs were in great need, people also had the demand to find jobs with reasonable wages. Nowadays, finding jobs online has become the mainstream, due to its advantages of a large amount of information access, the ignorance of space and time limits, and also, the comparatively low cost. However, these can also bring up side effects. For example, tons of information would raise the

DOI: 10.54254/2755-2721/8/20230152

difficulty of verifying, updating, or analysing [2]. By predicting annual wages using machine learning algorithms, useless or misleading information could be screened out, thus the rate of being cheated in the salary market would be declined. For companies looking for new employees, a reasonable wage can be provided referring to their personal conditions. Additionally, the salary prediction could also help companies update their salary structure, in turn, raise the rate of salary satisfaction among employees. "The shortcomings of the salary structure have also been exposed. Second, the current salary structure is unreasonable. The basic salary accounts for 75%, while the performance salary is only 20%. The proportion of performance salary is low, which cannot motivate employees [3]." determined Hongyan Zhang and Chunmei Ni. While for job hunters, this prediction process can help estimate whether the salary provided is worth it or not. For the government, after the prediction data is collected and analysed, related salary or job delivering policies can be established. Finally, for the whole nation, the result of salary prediction could be used to evaluate the economic system. Specifically, a wage-prediction application based on the SVM algorithm, and compatible with both mobile phones and computers could be designed. The salary data could also be collected and sent to the official platform for further studying, raising the success rate of recruitment or policies making.

James Otto, HAN Chaodong, and TOMASI Stella at Towson University used the Neural Network to predict wages based on workers' skills in the year 2021. They created the architecture using "freely available standardized skills and wage data" to determine the wage values according to varieties of combinations of general working skills [4]. Professor PENG Yichun at Yulin Normal University also applied linear regression, SVM, decision tree, random forest, and other algorithms to job salary prediction in 2021. In his experiments, the average absolute error, mean square error, and the R square error were set to be the evaluation index. Then it came with the result that "the algorithm of random forest could provide reasonable and scientific reference for job seekers to publish recruitment information and search for suitable position." [2]. Studies mostly focus on algorithms comparing and selecting. In this experiment, the SVM algorithm is applied to analyse the salary classification dataset by calculating the impact of all features in annual wage and doing the prediction based on these results. The influence of training and testing set division to the accuracy was selected to be the variable in the research.

Compared to other machine learning methods, the accuracy of SVM can be over 80% or even 90% in classification missions with both low and high-dimensional data. Additionally, varieties of kernel functions can be qualified for specific tasks. However, the running efficiency would decrease as the amount of data for training increases. And the data needs a stricter normalization step to ensure that the result would not be affected by the errors in the dataset, which could be costly. The prediction results can also be harder to interpret in the end.

2. Method

In this section, the data source and the pre-processing steps are firstly demonstrated. Then the mechanism of the SVM is introduced.

2.1. Dataset and pre-processing

The dataset, named Salary Classification, was downloaded from Kaggle, a website that provides thousands of datasets in different fields. It includes 32562 individuals as samples, each sample consists of 14 features as columns. Due to its large sample size and multiple characteristics, this dataset was believed to perform well in training and testing.

After being downloaded, the next step is to do the pre-processing. This step is been divided into four parts: missing data supplement, data type transformation, dataset division, and the columns dropping part. Then, SVM algorithm is applied to analyse the dataset prepared. When using SVM, the division rate and the gamma value are switched, in order to improve the evaluation scores.

Recent years have witnessed a dramatic rise in the demand for the application of big data analytics techniques. It has the power to boost the realization of analysis and simulations in data-related fields like financial engineering and electrical engineering. Data quality plays an important role in

DOI: 10.54254/2755-2721/8/20230152

determining the quality of big data analytics products, argued Caihua Liu [5]. Hence after having access to the dataset, it was not used for training the algorithm immediately. As for data pre-processing, the experiment contains four steps. The first step is to fill up the missing data. When checking the dataset, many question marks were found in "workclass", "occupation" and "native-country" columns, which means these data were lost in the process of collecting. Thus, as the most frequent value of every feature could mostly approach the original version of the dataset, it was chosen to replace the data with question marks. Next, the encoder function was used to transform the data in form of description into values. For example, work class, education, occupation, and relationship in the family were converted into integers. The following step is to delete some features that have slight influence on the result. Skimming the dataset, most sample individuals were from the United States, the majority education background was high school. Additionally, the 0-capital gain and 0-capital loss made up 96% of all the data. So, these columns should be ignored to get rid of errors caused by them. Last, the remaining data was divided as the ratio of 7: 3 for SVM training and testing respectively.

Since the label of annual salary is whether an individual's wage could surpass 50k or not, it can be referred to as 1 and 0. To begin with, the dataset was imported by applying the pandas.read_csv function. Then, the label was set to be the Salary column, while other columns were set to be the parameters. Next, both the label column and other columns were divided at the ratio of 7:3, for training and testing respectively. The following step was to apply the data prepared using the SVM model importing from the sklearn library for prediction. The ratio of dataset division was switched for 4 times, each time the accuracy was calculated and printed for comparison.

2.2. SVM model

Support Vector Machine (SVM) is a theory that studies the regulation of machine learning in the case of a limited sample amount based on the statistical learning theory including Vapnik-Chervonenkis Dimension (VCD) and Structural Risk Minimization (SRM). Support Vector Machines are simple Supervised Machine Learning Algorithms applied for multidimensional classification or regression. Basically, the task of SVM is to find a hyper-plane as a boundary dividing data into several kinds. For example, in a two-dimensional plane, the hyper-plane is a straight line across the dots plotted. As for three-dimensional space, it could be an irregular curve [6]. Firstly, all the data will be plotted as dots in a multidimensional space, in which the number of dimensions equals the number of features of each sample. After that, SVM would find a hyper-plane, using as a barrier, which has the longest distance to the nearest dots through calculating. Support Vector Machine has extraordinary advantages in the task of processing datasets with a small number of samples, nonlinear features, and high dimension [6].

The classification and prediction accuracy are based on the parameter choice. In a Support Vector Machine algorithm, there are three main parameters that would affect the accuracy, the Kernel, Gamma value, and the C parameter. The kernel, also called the kernel function, is always selected according to the type of data. If no certain kernel function is chosen, the default one, named Radial Basis Function Kernel, would be used. In this function, the similarity between two samples in the transformed feature space is regarded as an exponentially decaying function of the distance between the vectors and the original input space [6]. The gamma value is responsible for deciding the influence of one specific sample in the training set when calculating. One step further, it will influence the judgment of the similarity of several samples by calculating the distance between them. The smaller the gamma value, the further points would be considered as one group [6]. Consequently, a large gamma value may cause the accuracy to be low. While extremely small gamma values can result in overfitting. The C parameter determines the amount of regularization when processing the data. A high C value means a low standard of regularization, which may lead to high accuracy. However, a large C value may also cause overfitting [6]. On the contrary, a small C value means more tolerance to the errors, which would result in low accuracy.

DOI: 10.54254/2755-2721/8/20230152

3. Result

In order to grasp the characteristics of the data, three diagrams were plotted to illustrate the salary distribution, the age-salary relation and the occupation-salary relation respectively. During the process, abnormal samples that does not make sense were found and eliminated. In the following steps, model parameters were altered to see if the accuracy would increase.

From the salary distribution in Figure 1, the number of individuals whose annual wage was under fifty thousand far exceeded those whose wage was beyond fifty thousand, which was almost 25 thousand. The age-salary distribution chart, shown in Figure 2, illustrated that most employees were 25 to 35 years old, among which only few of them got a highly paid job. While over 5500 people went to work before 25 years old, and only less than a hundred were paid over fifty thousand. The occupation-salary bar chart, displayed in Figure 3, showed that professors' salary was most likely to go beyond fifty thousand. On the contrary, only few individuals chose to join the army, whose salary were less likely to go over fifty thousand.

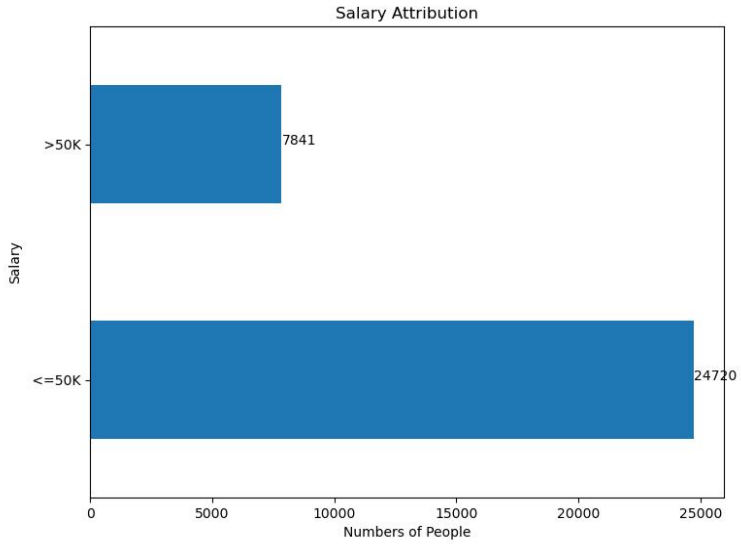

Figure 1. Salary distribution.

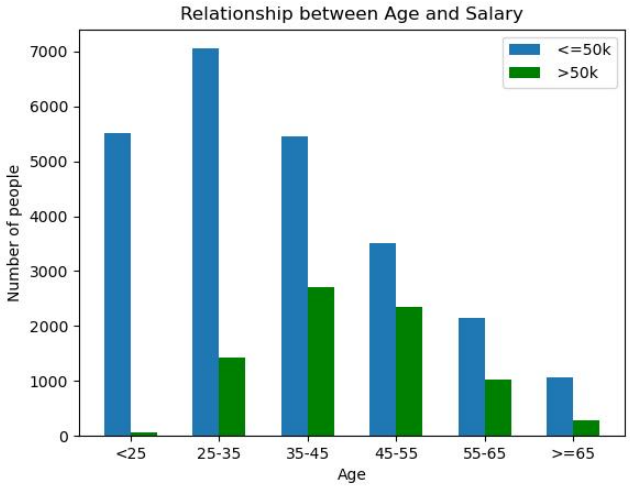

Figure 2. Relationship between age and salary.

DOI: 10.54254/2755-2721/8/20230152

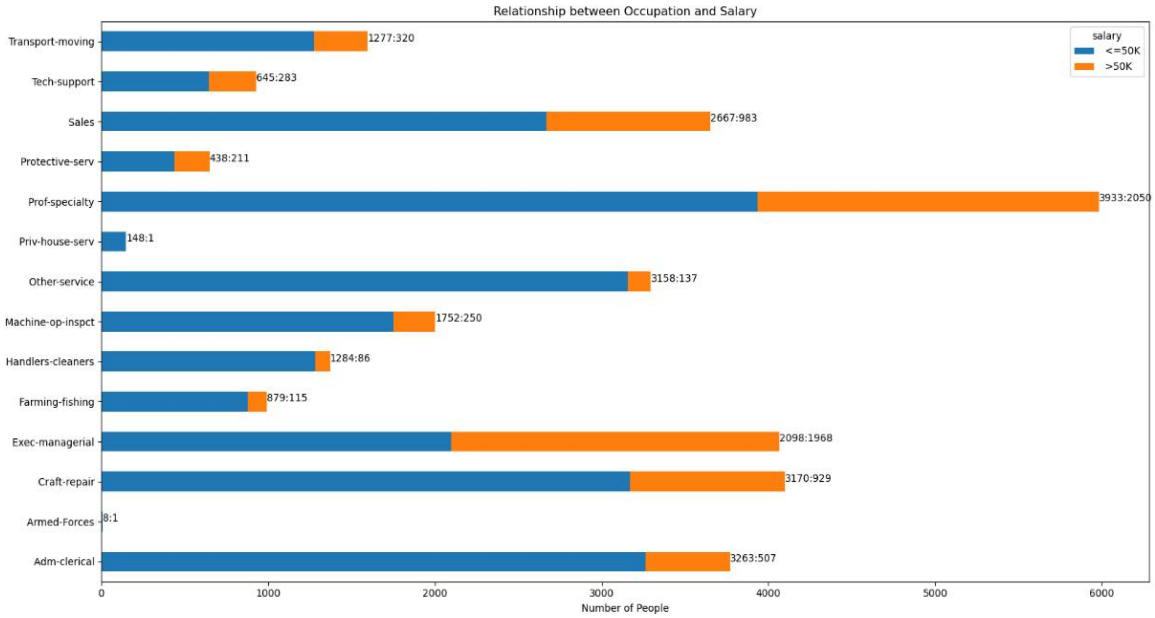

Figure 3. Relationship between occupation and salary.

During the plotting step, it is found that there existed seven people who never worked but weekly working hours were not zero, and several individual whose age were over 80 but worked 99 hours per week. These strange data was deleted in order to avoid accuracy decrease caused by them.

When applying the SVM algorithm, the proportion of training and testing set was switched from 6:4 to 9:1, the gamma value was changed from 1 to default value. The results are demonstrated from Table 1 to Table 5. The accuracy calculated shows that the accuracy of both training and testing set increased steadily as the testing set became larger. However, the difference between accuracy of training and testing set was almost 0.2. Sense the gamma value controls the weight of each sample in the training set in calculating, different gamma values might also cause the change in accuracy. As the gamma value was switched from 1 to the default value, the accuracy of both datasets was around 0.8.

Table 1. SVM classification evaluating index (test rate=0.4)

	Training Dataset	Testing Dataset
Accuracy	0.964	0.783
F1 score	0.921	0.306
Precision score	0.966	0.682
Recall score	0.881	0.197

Table 2. SVM classification evaluating index (test rate=0.3)

	Training Dataset	Testing Dataset
Accuracy	0.961	0.785
F1 score	0.916	0.326
Precision score	0.959	0.676
Recall score	0.876	0.215

DOI: 10.54254/2755-2721/8/20230152

Table 3. SVM classification evaluating index (test rate=0.2)

	Training Dataset	Testing Dataset
Accuracy	0.959	0.785
F1 score	0.910	0.343
Precision score	0.958	0.682
Recall score	0.866	0.229

Table 4. SVM classification evaluating index (test rate=0.1)

	Training Dataset	Testing Dataset
Accuracy	0.957	0.780
F1 score	0.906	0.351
Precision score	0.951	0.680
Recall score	0.865	0.236

Table 5. SVM classification evaluating index.

	Training Dataset	Testing Dataset
Accuracy	0.801	0.802
F1 score	0.423	0.436
Precision score	0.700	0.707
Recall score	0.303	0.315

4. Discussion

The goal of predicting the annual salary accurately was achieved by doing pre-process and applying SVM algorithm to 14 characteristics of 32562 individuals. The accuracy of the architecture built was around 80%.

The result of this experiment could be considered satisfied, sense the accuracy was comparatively high, while the time cost of the experiment in each condition was close to 30 seconds, which could be regarded quick. Due to the high accuracy and efficiency, this method could be generalized at a relatively low cost, which helps the job hunters to estimate the rationality of salary provided quickly so that they do not have to wait for the interview results for weeks. And for companies looking for employees, it could be more likely to raise the efficiency of evaluating the conditions of interviewees, then offer the reasonable wage in a short time. As for the government, this algorithm might help evaluate the situation of salary market, in turn, potential problems could be found and solved, related policies would also be established.

Though the dataset was cleaned before the experiment, parameters of the model and rate of training set were switched for many times, the accuracy seemed not to exceed 90%. One reason might be that some abnormal data, for example, the lost data in the column “occupation”, “workclass” and “native-country”, were replaced with the most frequent ones, nevertheless, this process still affected the accuracy, which means it might not reflect the origin situation. Another influential feature could be that dealing with 32562 samples might cause the overfit of model. When processing a great number of samples, the noise existing in each column might be amplified, this may also affect the weight of each feature. Consequently, the model would be misguided, which would cause the accuracy to be low. Additionally, when checking the dataset, 4 features that related weakly to the salary was observed, hence, the “education”, “relationship”, “capital-gain” and “capital-loss” column were dropped to avoid errors. However, it was not sure that whether other features like “race” and “gender” were strongly connected to the salary. Considering this, Ankit Gupta developed a confusion matrix to visualize the connection coefficients of each element in his research[7]. If more weakly-related columns were eliminated, the accuracy might be higher.

DOI: 10.54254/2755-2721/8/20230152

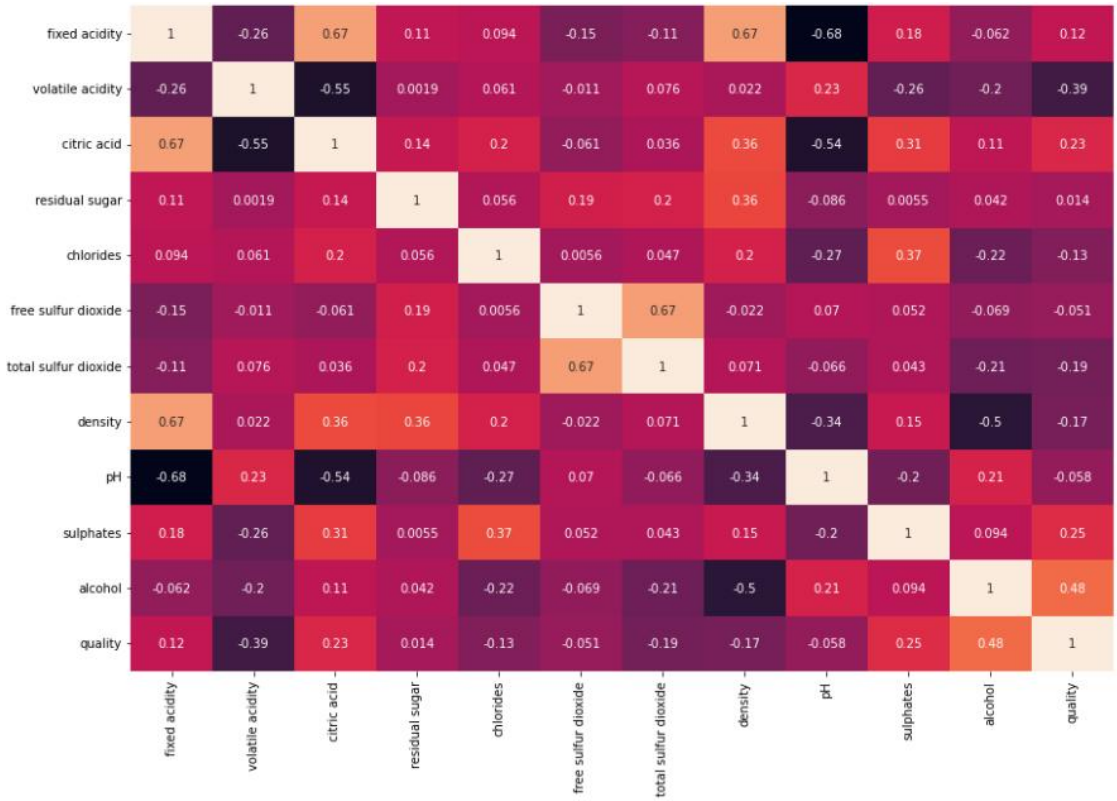

Figure 4. Feature correlations.

To make the improvement in accuracy of prediction, there are came out several methods. In the pre-processing part, when the data in the type of description was convert into numbers, the function sorted the serial numbers according to the value as a default step. However, the values bear no meaning, so the error could not be avoided in the algorithm. Sense, the improvement in the data type transforming function might avoid this problem. In the training set selection part, there still exists room for progress. Jakub Nalepa and Michal Kawulok demonstrated that "SVM algorithm suffers from the important shortcomings of memory training complexities, which depend on the training set size."[8] They designed an algorithm whose main task is to minimize the cardinality of the training set. "To make this measure easier to interpret for datasets of different size," they pointed, "it is very often presented as the reduction rate (R)."

$$R = \frac{t}{t} \tag{1}$$

The parameter "t" refers to the original size of training set, and "t" is the refined one [8]. Though the training set might be refined and some less important vectors would be eliminated, those comparatively significant vectors which might influence the weight calculating must be kept.

Another method of training and testing set division is that dividing the original dataset into K disjoint subsets, one of which could be selected as testing set and others be the training sets. Each training set matches the testing set as one group, then the accuracy could be calculated. "The K models can be obtained in this way, and the average of the final classification accuracy of the K models as the performance index of the K-CV classifier" [9] argued Xuemei Yao in Application of Optimized SVM in Sample Classification, "K is greater than or equals to 2. But in practice, K is taken from 3. K will try to take 2 only when the original data is very small."

One further point to make, algorithm with more complicate structure or more advanced architecture may perform much better than SVM in annual salary classification and prediction task. Ignacio Martin,

DOI: 10.54254/2755-2721/8/20230152

Andrea Mariello, Roberto Battiti, and Jose Alberto Hernandez compared 5 models including logistic regression, nearest neighbours, MLPs, SVMs, random forests, adaptive boosting and voting classifiers based on all or part of them. According to their results, "ensembles based on decision trees behave generally better and that a voting committee based on them leads to an accuracy of about 84%." [10].

5. Conclusion

The salary prediction has become a spotlight in the past few years, especially when the pandemic attacked the world and resulted in a record-low rate of employment. The requirement of searching for a job with a salary that fits employees' capability rocketed up.

Many researchers have investigated the influence of individuals' features on their annual salary by comparing and selecting the relatively best machine learning algorithm. While in this experiment, the focus was on the ratio of training, testing sets division, and the gamma value in the SVM algorithm. Data were pre-processed before analysing. The pre-processing part contains missing data supplement, data type transformation, training set division and columns drop 4 steps. After data cleaning, SVM was applied to the Salary Classification dataset. According to the result, when the gamma value was set to 1, the accuracy of the training set could reach over 90%, comparatively, the accuracy of the testing set was only near 78%, which means the model might be overfitted. The other group where the gamma value was the default performed much better. The accuracy of both the training and testing set were around 80%, which can be considered high. Still, there stands a chance for the improvement of the data cleaning step and the algorithm parameter adjustment.

References

[1] Abhinav Rai. (2018) What is Big Data – Characteristics, Types, Benefits & Examples. Accessed: 09/29/2022. ULR: https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples.

[2] Gao, X., Wen, J., & Zhang, C. (2019). An improved random forest algorithm for predicting employee turnover. Mathematical Problems in Engineering.

[3] Wang, X., & Zhang, L. (2020). A Survey on Job Satisfaction and Countermeasures in Shipping Companies. Journal of Coastal Research, 106, 486-489.

[4] James, O., Chaodong, H. A. N., & Tomasi, S. (2021). Using Neural Networks to Predict Wages Based on Worker Skills. Studies in Business & Economics, 16(1).

[5] Liu, C., Peng, G., Kong, Y., Li, S., & Chen, S. (2021). Data Quality Affecting Big Data Analytics in Smart Factories: Research Themes, Issues and Methods. Symmetry, 13(8), 1440.

[6] Kecman, V. (2005). Support vector machines–an introduction. In Support vector machines: theory and applications 1-47.

[7] Bocca, F. F., & Rodrigues, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Computers and electronics in agriculture, 128, 67-76.

[8] Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. Artificial Intelligence Review, 52(2), 857-900.

[9] Yan, X., & Jia, M. (2018). A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing. Neurocomputing, 313, 47-64.

[10] Martín, I., Mariello, A., Battiti, R., & Hernández, J. A. (2018). Salary prediction in the IT job market with few high-dimensional samples: A Spanish case study. International Journal of Computational Intelligence Systems, 11(1), 1192-1209.