Model Enhancement Report
- Student Classification Task -

Ji-seong Yang (Kaggle Account: **Michael**)

I. **Overview** As per the standard procedure of data science, the dataset went through the preprocessing as the very first step, and several classifier methods have been tested. Repetitive experiments have been conducted to find the best performing model. The main methodology of choosing the "better" model in this paper was an A/B test, in which one that is superior to the prior model is selected as the baseline for the following experiment.

II. **Data Preprocessing** Data preprocessing roughly consists of dealing with missing values, data encoding, and scaling. It turned out that the given dataset does not contain any missing values. As for the categorical data encoding, "famsize" needed special treatment since it was an ordinal variable. "LE3" and "GT3" have been labeled as respectively 0 and 1, so that the size difference information could be maintained. Nominal features[1] have been converted into dummy variables. One-hot encoding has been applied to features with more than two value categories.[2] Numerical features[3], along as numerically converted nominal values in some trials, have been standardized with the mean of 0 and the standard deviation of 1 with an exception of the "ID" variable.

III. Experiments

No.	Preprocessing	Scaling	Methods	Accuracy
1	Nominal values → dummies + one-hot encoding[4]	Minmax Scaler	Naïve Bayes	0.12977
2	Nominal values → dummies + one-hot encoding	Standard Scaler	Naïve Bayes	0.12977
3	Nominal values → dummies + one-hot encoding	Standard Scaler	Decision Tree	0.32824
4	Nominal values → dummies + one-hot encoding	Standard Scaler	KNN (k=1)	0.32824
5	Nominal values → dummies + one-hot encoding	Standard Scaler	KNN (k=2)	0.30534
6	Nominal values → dummies + one-hot encoding	Standard Scaler	KNN (k=5)	0.36641
7	Nominal values → dummies + one-hot encoding	Standard Scaler	KNN (k=7)	0.36641
8	Nominal values → dummies + one-hot encoding	Standard Scaler	KNN (k=10)	0.35877
9	Nominal values → dummies + one-hot encoding	Standard Scaler	SVM (linear)	0.38931
10	Nominal values → dummies + one-hot encoding	Standard Scaler	SVM (RBF)	0.39694
11	Some numerical features[5] → one-hot encoding	Standard Scaler	SVM (RBF)	0.38931

[1] 'school', 'sex', 'address', 'Pstatus', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery', 'higher', 'internet', 'romantic', 'Mjob', 'Fjob', 'reason', 'guardian' (sixteen features)

[2] 'Mjob', 'Fjob', 'reason', 'guardian' (four features)

[3] 'goout', 'famsize', 'Dalc', 'traveltime', 'famrel', 'failures', 'health', 'Fedu', 'studytime', 'freetime', 'age', 'Walc', 'absences', 'Medu' (fourteen features)

[4] Pandas method to get dummies automatically performs one-hot encoding on multi-class nominal features.

[5] Features such as "Medu", "Fedu", "famrel", "health" that have ordinal values were experimentally converted

12	All features scaled altogether[6]	Standard Scaler	SVM (RBF)	0.41984
13	All features scaled altogether	Standard Scaler	SVM (RBF) (C=100)	0.41984
14	All features scaled altogether	Standard Scaler	SVM (RBF) (C=100, gamma=1000)	0.46564
15	All features scaled altogether	Standard Scaler	SVM (RBF) (C=100, gamma=5000)	0.46564
16	Some multi-class nominal features → scaled[7]	Standard Scaler	SVM (RBF) (C=100)	0.44274
17	Some multi-class nominal features → scaled	Standard Scaler	SVM (RBF) (C=100, gamma=1000)	0.46564

IV. Summary

a. **Seventeen** submissions in total.

b. Naïve Bayes, Decision Tree, k-Nearest Neighbor, and Support Vector Machine

c. The Best Solution

1. The accuracy score is as high as **0.46564.**
2. The best when **all features were scaled altogether** or when **some multi-class nominal features** were scaled.
3. Standard scaler
4. **Support Vector Machine** using **RBF** function with C of 100 and gamma value of 1000 or 5000.

V. Major Takeaways

a. **Size Matters!** The implication of the trial **No. 11** is that unlike nominal features whose values are equivalent to one another, the size information ordinal features reflect should be retained. The loss of information leads to poor performance.

b. **Scaling Brightens Your Model** In the initial stage of the experiment, only numerical features were considered as the target of scaling. However, it proved that even scaling numerically encoded values would enhance the model. The difference in the score of trial No. 11 and 12 reflects "scaling effect". There is no significant difference between minmax scaling and standardization in terms of the result.

c. **Methods vary in their performance** In general, the model has poor performance in the following order: Naïve Bayes, Decision Tree, k-Nearest Neighbor, and Support Vector Machine. However, it is not necessarily the case depending on the parameters of each method. The success of Support Vector Machine may be attributed to kernels that allows the model more accurately than using **just a linear model**. Though not tried in this experiment, Softmax method with Deep Learning is worth to be adopted in the subsequent project since it is known to be a very powerful multinomial classifier.

d. **Optimal Number of neighbors in KNN** There is an optimal k value. The larger number does not guarantee the higher accuracy. In the case of this dataset and model, the value of k ranging from 5 to 7 has resulted in the best outcome. It is imperative to find adequate value by conducting multiple times of experiments.

to strings as if they are mere nominals before one-hot encoding.

[6] This means dummy variables including one-hot encoded values were all scaled as well as their numerical counterpart, as if they are arithmetic numbers.

[7] Features such as 'Mjob', 'Fjob', 'reason', 'guardian' that have non-ordinal multiclass values were scaled as if they are numerical values after being mapped to their corresponding numbers.

e. **Tinkering with Support Vector Machines** Support Vector Machine supports several different kernel functions and among which RBF, also known as the Gaussian kernel, scored the highest accuracy. C parameter, whose default value is one, does not significantly affect the model performance despite its increase up to one hundred. On the other hand, the change in gamma value has boosted the model accuracy remarkably when it was set to one thousand. However, a further increase in this value did not do much for the model. No more optimization is expected due to C and gamma value.