2023 AS&DA

Unraveling the Threads of Success

A PREDICTIVE ANALYSIS OF STUDENT DROPOUT & ACHIEVEMENT

GROUP MEMBERS :
BUYAN LI
RUOTONG SUN
YUNNAN HE

Background & Summary of work

This dataset provides a comprehensive view of students in higher education institution in Portugal. The dataset comprise 4424 observations with each recording his **academic achievement** and **36 other attributes**, including family status, previous level and learning condition, of the student.

Our research is committed to **predict** the students' academic achievement as well as identifying the most influential factors. To achieve this, we applied a wide range of methods including regression, decision tree, factor analysis, natural language processing(NLP), K-NearestNeighbor(KNN) and support vector machines(SVM).

Exploratory Data Analysis & PCA

Figure 1 presents a **heatmap** of the entire dataset, featuring a hierarchical cluster that signifies the dataset's multicollinearity. From the figure we can see that the course-related data are highly corelated.

Figure 2 illustrates the correlation among course-related data from the first semester.

Figure 3 is the p-values of Chi-square test conducted between all discrete variables and the labels from which we can see that some attributes are **not significantly related**.

Figure 4 depicts the count and portion of each label, highlighting the presence of label imbalance.

Figure 5 reveals the results of a Principal Component Analysis (PCA) performed on 19 'continuous' attributes. The analysis suggests that distinguishing and **predicting the 'enrolled' students may be challenging**.

Combining the finding from Figure 4 and Figure 5, the **"enrolled" label was omitted** from some parts of the prediction process.

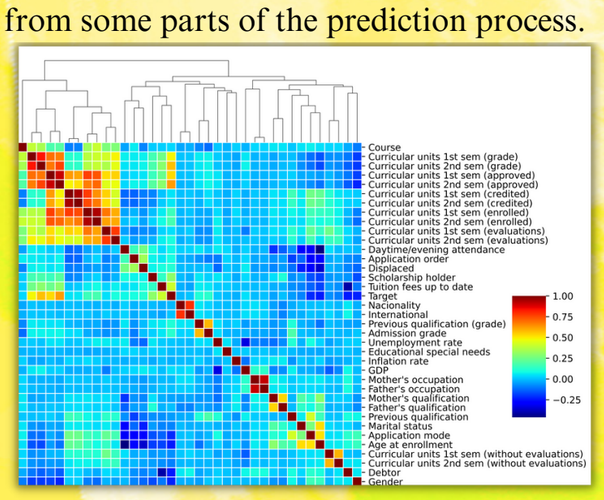

FIGURE 1

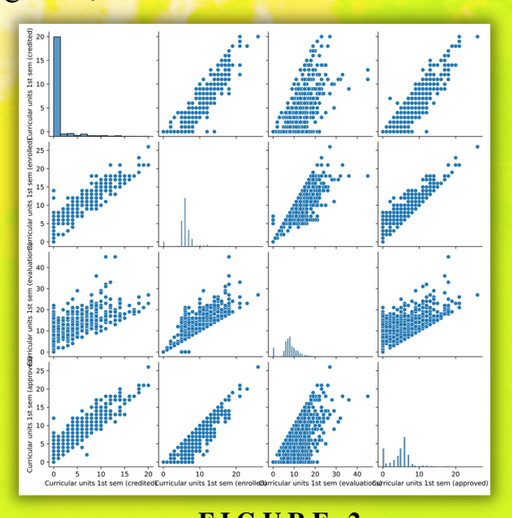

FIGURE 2

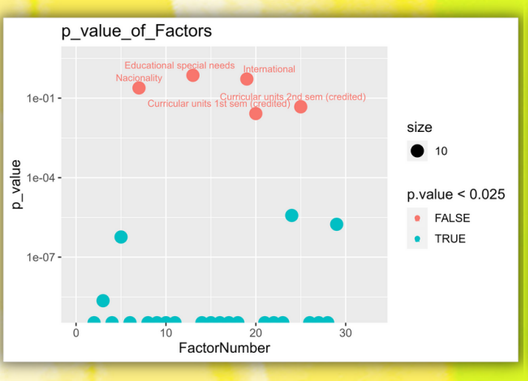

FIGURE 3

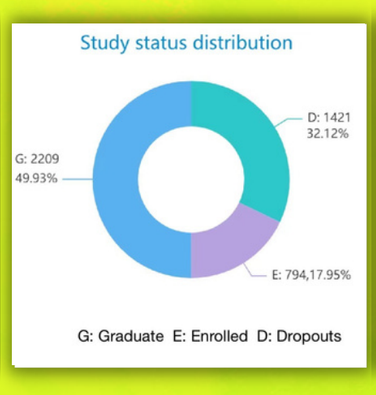

FIGURE 4

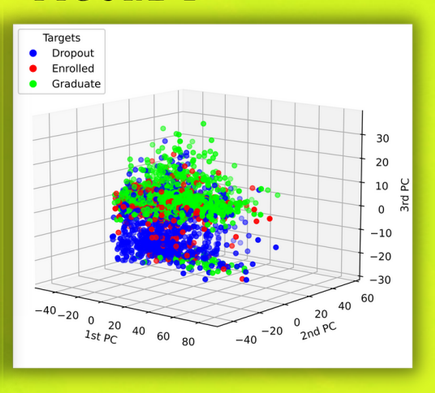

FIGURE 5

Decision Tree & Smote

We employed a **decision tree model**, shuffling the data and partitioning it into a 70% training set and a 30% test set. Hyperparameters were then tuned to optimize test accuracy, achieving a peak of approximately 75%, while the training score was 80.4%. As illustrated in Figure 5's confusion matrix, the lower accuracy is primarily attributed to the **imbalanced data**, a latent issue given the minimum data and highest confusion rate within the 'enrolled' group. To address this, we implemented **Synthetic Minority Over-sampling Technique (SMOTE)** to balance the labels, subsequently increasing the test accuracy to 83%. The updated confusion matrix is displayed in Figure 6.

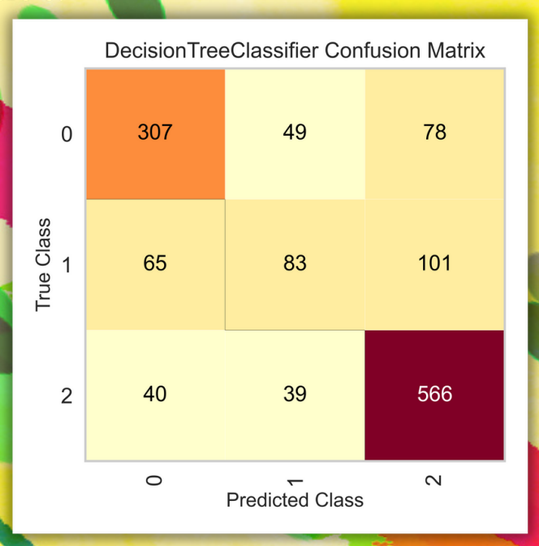

FIGURE 6

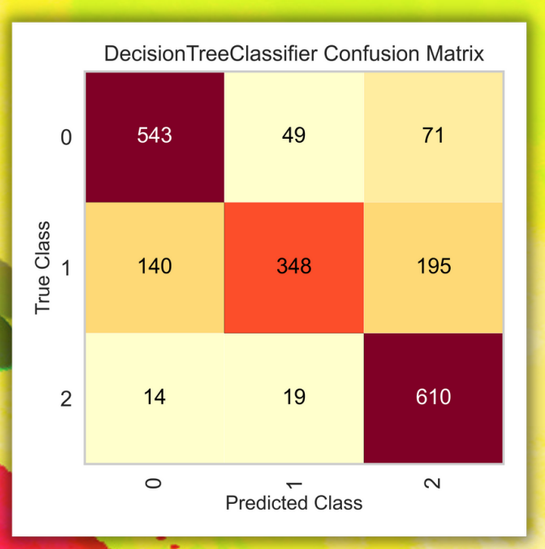

FIGURE 7

Regression & variable selection

We utilized a regression model to identify the most relevant variables. We **omitted blindly adding dummy variables** since the **"distance"** between the categories are obviously different. For example, fisher is more "close" to farmer than managers. To address this issue, we applied word vector model to construct **word vectors** based on their semantic meaning, and then use t-SNE method to reduce them to 2-D vectors. The word vectors of course type and marriage condition are shown in Figure 8 and Figure 9 as examples.

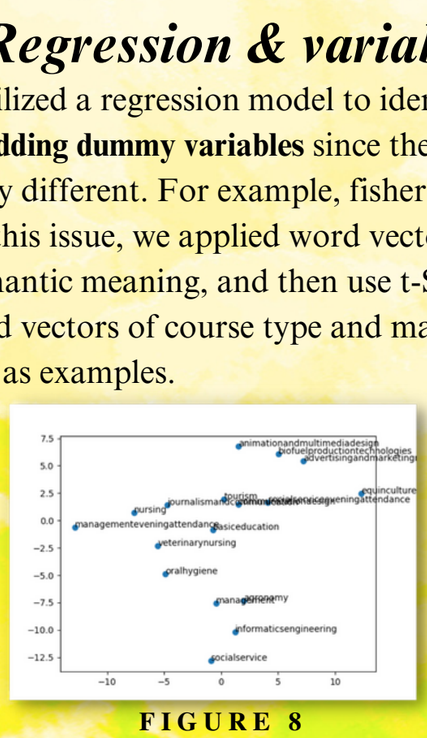

FIGURE 8

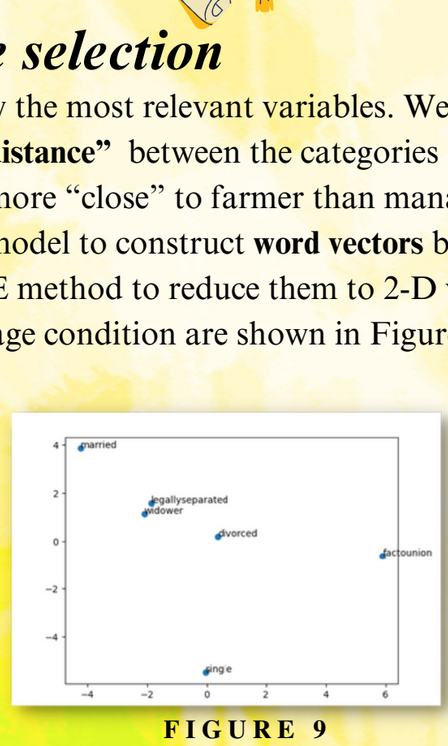

FIGURE 9

The selected variables imply the prominent influence of **students' behaviors during semesters** on their academic outcomes. Additionally, attributes such as **age, and economic status** also demonstrate significant associations with results. Besides, father's qualification emerges as a noteworthy factor. Consistent with our expectations, macroeconomic data exhibit weak correlations with students' academic behaviors.

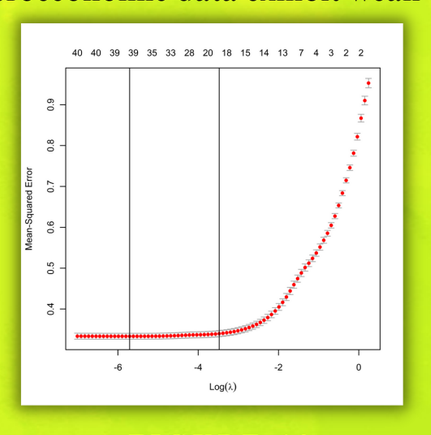

FIGURE 10

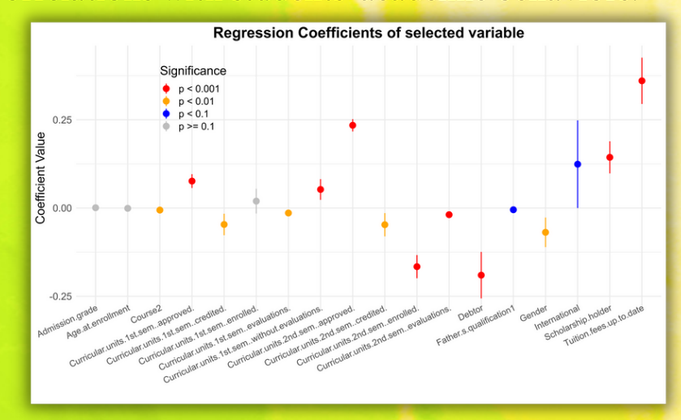

FIGURE 11

To select the variables, we applied the **Lasso** method. Since we focus on dropout and achievement, we excluded all data points with the 'Enrolled' label, which has an ambiguous meaning. The lambda parameter and the significance level for the selected variables are shown in Figure 10 and Figure 11, respectively.

Factor analysis

Factor analysis, employing **MLE**, was conducted to identify underlying constructs within the data. Six factors were extracted and rotated to maximize variance of loadings. We use **WLS** method to get Bartlett scores. The sixth factor was excluded due to ambiguous interpretability. The remaining five factors, namely "Credits," "Grade," "Age & Study Status," "Parental Status," and "Course Difficulty," were consistent with the regression analyses. In addition, figure 12 and 13 depicted data points 40 and 104 respectively, serving as examples of the derived factor scores.

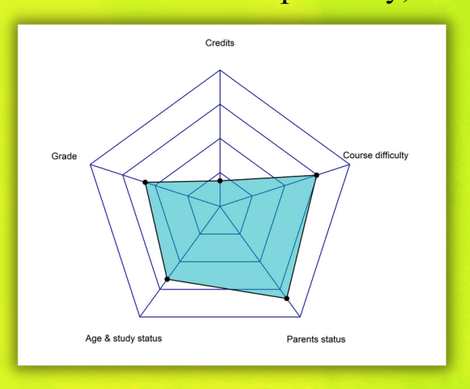

FIGURE 12

FIGURE 13

Some other attempts

We also tried some other common clustering methods in machine learning.

We applied **K-NearestNeighbor (KNN)** to the **standardized** data, and used **Grid Search** to get a nice parameter set. But the score of this model is just around 70% which is not so satisfying.

Since we had "played" with regression before, we now considered the clustering problem in a nonlinear way. We applied **support vector machines(SVM)** with Gauss kernel function to the standardized data with only 2 types "Dropout" and "Graduate" , and we used **cross validation** in parameter adjustment. And the model turns out to be quite good with score over 90%.

However, we just simply applied those models without considering choosing appropriate predictors, which may be a feasible way to improve model accuracy.

To sum up, instead of just a decision of our own, dropout is probably related to complex factors (sad story huh). Hope that we'll never have to think over this problem throughout our college life!