Detection of Risks to Student Health from Educational Data

Introduction:

According the the Dave Nee Foundation, symptoms of depression affect between 20-30% of adolescents today, and in the past year around 9% of high school students have attempted suicide [1].
The American College Health Association's (ACHA) Spring 2022 National College Health Assessment surveyed over 54,000 undergraduate students. It revealed that approximately 77% were experiencing moderate to serious psychological distress.

- 79% of surveyed students reported they had experienced moderate or high stress levels within the last 30 days.
- 54% met the criteria for experiencing loneliness.
- 29% met the criteria for suicidal ideation, while 3% reported attempting suicide in the past year.
- 12% had intentionally injured themselves within the year.

Early diagnosis of depressive symptoms can greatly increase a teenager's chance of leading a productive life, but many cases of these symptoms go unreported. Similarly, substance abuse remains a large problem among adolescents and according to the CDC, 3.7% of deaths among adolescents can be attributed to drug overdose [2]. However, just like with depressive symptoms, most substance abuse goes unreported until it's too late. Often schools can provide early intervention, or access to helpful resources, for these common threats to student mental health but lack the means of knowing when a student's health is at risk.
We thus explore the possibility of predicting student Mental health from data readily available to schools, which would allow schools to identify students at risk of threats to their mental health and provide intervention or resources as necessary. The input is educational records containing features such as student grades, absences, mother and father occupations, and relationships with friends. We then use Decision Trees, Random Forests, to output a prediction for student mental health on a scale from 0 to 5.

Related work:

Here, two Portuguese Higher Secondary School students have predicted the score G3, i.e in the final exam based on the previous 2 exams & other given features,

For the five-way classification, DT and RF performed the best in predicting Portuguese grades [3]. Class imbalance was not considered and the hyperparameter search was limited. It will be more difficult to achieve high accuracy in predicting student health, as first and second-period grades were by far the most important features in predicting third-period grades. Many other studies aim to predict some aspect of student performance or health given features collected from administrative and survey data.

Some students of Stanford University predicted student health from their educational records.

Dataset and Features:

We can download the dataset from: https://archive.ics.uci.edu/dataset/320/student+performance

Predicting student health from their educational records is an interesting and potentially valuable task. With a dataset containing 34 features, you have a wide range of information to work with.

Additionally, we are merging `student_math1 & student_por` to have a larger dataset.

One thing to be noted is that we do not have any label for mental_health_status.

To approach this problem effectively, we can follow these steps:

Data Preprocessing:

a. Data Cleaning: Begin by inspecting the dataset for missing values, outliers, and inconsistencies. Handle missing data using techniques like imputation or removal, and address outliers appropriately. b. Data Exploration: Explore the dataset to understand the distribution of features, relationships between variables, and any patterns that might emerge. Here, we are combining both the student_math and student_por datasets, where some Portuguese students study both subjects, so we have to drop the two tuples, which have the same values for all features.

Feature Selection:

Identify the features that are most relevant for predicting student health. We can use techniques like correlation analysis, feature importance scores, and domain knowledge to select the most informative features. Now we have a health_status which is the most important feature, & some other important features like scores in G1, G2, & G3. Also, we are going to add a new column to the merged dataset i.e. mental_health_status via random integer values 0 to 5.[which is our target column]. Here the correlation doesn't seem to work as we have imputed the target column, so instead, we are using Ensemble tree-based models like random forest to obtain the feature_importance, & we can also apply the Recursive feature elimination, using GradientBoostingClassifier to select the top k important features. So finally we are selecting those k top features which are common to both above feature selection methods.

Feature Engineering:

Create new features or transform existing ones if required to improve the predictive power of the model. First, we count the number of unique values of each feature, so that we can check for class imbalance of each feature. Now we can check for categorical columns, & apply one hot encoding to map with numerical columns, & drop the original categorical columns, & drop the nth numerical to avoid redundancy in the numerical columns, we are directly applying smote oversampling to overcome class imbalance, to have a good number of different values for each feature in training as well as a testing dataset,

Target Variable: "Mental_Health_status." It is a 5-way classification task (e.g., state of being good Mental_health_status increases as classification values rise towards 5.

Data Split: Split the dataset into training, validation, and test sets to evaluate model performance effectively. Before splitting we are taking those top k features of the student dataset and splitting the dataset into training and testing, further splitting the training dataset into training and validation.

Model Selection: Choosing appropriate machine learning or statistical models for our prediction task. So we are opting for the Decision tree classifier and then Random forest.

Model Training: Train the selected models on the training data and fine-tune their hyperparameters for optimal performance. Monitor for overfitting and use techniques like cross-validation to assess model generalization.

Model Evaluation: Evaluate model performance using appropriate metrics. Metrics like accuracy, precision, recall, F1-score, and ROC AUC can be useful. For regression, metrics like mean squared error (MSE) or mean absolute error (MAE) are common. So we do not have a good metrics score like accuracy, & model is not able to classify the dataset to predict the actual mental_health_status, we are imputing various hyperparameters of random forest classifier, to have a better result for selected sets of hyperparameters like `n_estimators`, `max_depth`, & `min_samples_split`.

Interpretability: Consider model interpretability techniques to understand which features are domains where actionable insights are needed. We can apply PCA or LDA to merged_dataset having numerical columns only, to reduce the number of features, and to have some new features also.