Data Mining Project: Student Performance

ISM 6136

Tracyann Burton

Florida Atlantic University

April 26, 2022

**Report Title**: Student Performance

**Prepared By**: Tracyann Burton

**Synopses**

The data set shows student academic performance in Math, Reading and Writing and includes demographic data points, including gender, parent’s educational attainment, socioeconomic standing and test preparation services utilized. Using the data provided, the assessment that follows will use the socioeconomic and test indicators to determine student gender. The data will determine if gender can be determined by academic performance and demographic indicators. The data contained 1000 observations, 50 of which were separated and used for new data modeling.

**Dataset Cleanup: Dummy Creation and Data Cleanup**

**Missing Data Handling**

All rows and columns of the dataset were assessed to determine if there were any empty cells that should be deleted. This step was done in order to ensure the most accurate information was given to the algorithm to run, and to reduce possible issues with the data during analyses.

Graphical user interface, text, application, email

Description automatically generated

**Dummy Creation**

Dummies were created for all categories, except outcome (gender), that did not contain continuous numerical data. Creating dummies allowed for the assessment to use numerical data for all variables. Dummies are numerical figures assigned to categorical data to allow the algorithm to assess the data. Dummy was not created for gender, which is the outcome. Outcome is never assigned a dummy.

Graphical user interface

Description automatically generated

**Data Mining Techniques: Classification**

The output that the dataset looks to ascertain is categorical, i.e., non continuous numerical data, and as such, classification methodology is needed. Classification algorithm is used when the output data does not contain continuous numerical information. For the analysis that follows, KNN and Neural Network algorithms were used.

KNN and Neural Network algorithms can be used for both classification and regression analyses. KNN, nearest neighbor, uses random number K and is effective at determining complex relationships among indicators. KNN was selected because the dataset is optimal when there are less indicators to consider. In the dataset studied, there were only seven indicators utilized to determine gender. Neural Network, similarly, captures complex relationships amongst indicators, and it can build accurate models even if the data is noisy.

**KNN**

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Neural Network**

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**Model Construction: PCA and Partitioning**

**PCA: Dimension Reduction at 95%**

PCA technique was selected to reduce dimensions/ variables ahead of running the data. Typically, PCA helps to identify the salient variables that impact the assessment. In the dataset, the 95% variance was selected to ascertain the smallest number of variables to give the highest variance. The Covariance Matrix was selected in order for the data to be automatically normalized. The data needed to be normalized because the data set had varying decimal places. Additionally, this prevented the system applying bias based on perceived importance of one number over the other. Looking at the variables included in the top 5 components (writing score, test prep, lunch, race and level of education), I expanded this to include also math and reading. As a result of the limited number of dimensions and the results of PCA, all the variables in the dataset were identified as being important and did not warrant dimension reduction.

Graphical user interface, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Partitioning**

Three partitioning techniques were used in both the KNN and Neural Network algorithms. In the KNN and Neural Network algorithms, 60/40 and 80/20 partitions were utilized. In Neural Network, a 90/10 split was done to achieve a third data model. In each partition, all indicators, including the outcome, were used in the data partitioning.

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated

**Model Selection**

Seven models were built and the top model for each algorithm highlighted in the model table below. KNN’s best model is Model 3, which had a validation accuracy of 62%, which was higher than the other KNN models, despite being lower than the 70% “good model” data point. Additionally, the best KNN model had lower overall validation error and higher validation F1 and precision values, than the other KNN models. The better Neural Network Concept (NNC) model is Model 6 which used a 80/20 partition. Model 6, similar all NNC models, had a validation accuracy over 70%, indicating that they were all good models. Model 6 had the higher validation accuracy (86.8%), along with the lowest validation error (13.16%) and highest validation F1 (.88).

Application

Description automatically generated

**Deployment: Scoring**

**KNN—The best KNN model, model 3 was deployed and the results screenshot below.**

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

**NNC Scoring- the best NNC and best overall model, model 6 was deployed and the screenshot posted below.**

Table

Description automatically generated

Table

Description automatically generated

**Conclusion**

In the case of this dataset and the algorithms deployed, NNC performed better than KNN. The NNC algorithms built more accurate models. The automatic NNC method was selected for the NNC model creation and that resulted in an easier model building process. KNN algorithm, though theoretically a simple model, it did not build reliable models for this current dataset as all the models were below 70% in their validation accuracy. NNC was able to quickly develop more reliable and accurate models. In the results for the deployed models for KNN and NNC, despite differences in the outcomes, the first and last two predictions in each deployed model are the same.