MSCI 436: Decision Support System

Dev. Status Update

Group 6

Name	**Email**
Aashan Meiyappan	a2meiyap@uwaterloo.ca
Karthik Prasad	k3prasad@uwaterloo.ca
Rishega Selvaraj	r9selvar@uwaterloo.ca
Sam Walker	sa2walke@uwaterloo.ca

Problem:

This project will be focusing on operationalizing a machine learning model which is focused on improving student study habits in order to achieve higher scores. This DSS will be solving the issue of students not knowing what aspects of their study habits have the most significant impact on their final grade.

Value Proposition: The model will identify which of the students behaviors must be tackled explicitly in order to see improvement in their grade and what this grade will be. This problem necessitates the use of a DSS as each student's lifestyle can greatly vary in the level of discipline they display in particular categories. The implementation of a DSS will make the process of improvement more efficient and the students can effectively work on the identified factors which have a direct impact on their grades.

Data:

Dataset: https://www.kaggle.com/datasets/devansodariya/student-performance-data
Code: https://www.kaggle.com/code/aryanmsr/complete-ml-analysis/notebook

A student performance dataset from Kaggle has been selected for use in this project. The dataset measures 30 student characteristics such as gender, study time, internet usage, and more. All 30 of the features in the data will be used against three measured target values: G1, G2, G3 (test scores). The data will be formatted and processed with the help of libraries like sklearn, numpy and pandas to prepare the features to be inputted easily into the model. This dataset is appropriate for the problem at hand as it provides a variety of factors that impact a student's grade and it outputs 3 test scores which can then be averaged to provide a more accurate representation of each student.

There are some limitations to this dataset that impede the results. For example, the participants in this data set range in age from 15-22 however, the distribution is not even. There are only 4 participants over the age of 20 which skews the data as students in high school and post-secondary school practice different lifestyles. The data is also limiting as it does not accommodate a variety of family dynamics. Information regarding the traditional mother and father occupation/education is collected but a lot of students are raised in a household that does not look as such. Categories such as health are very relative and it is difficult to ensure that participants answer on the same basis.

Model:

The model will be developed using the 'scikit-learn' library which consists of many models/tools that will be used to model the student performances. This task will use a supervised regression model in order to learn a function mapping the numerical target to labeled inputs. As a result, a linear regression model was used to process the inputted data.

The above model was assessed on the basis of a variety of constraints but mainly focused on two, accuracy and interpretability. The model needed to provide insights that were accurate so the students were able to implement the suggested improvements and the suggestion needed to be easily digestible as anyone with next to no ML experience could use the tool. As the model listed above aligned with our desired results, it was implemented in the final model. Some advantages of linear regression are that it can be regularized in order to avoid overfitting. It's also scalable as new data can be easily added utilizing stochastic gradient descent. Some disadvantages consist of the model being unable to compute non-linear relationships showcasing more complex patterns.

User Interface:

The interface runs on Streamlit.io and is extremely simple to use. Users of the model are encouraged to use a multiple choice selection to input their parameters (such as age, gender, etc.) on the left hand side of the screen. Once inputted, the user can 'run' the model which will then output results containing text as well as visualizations using matplotlib. Visuals will display the key factors which were identified to highly impact the score of the students. Thus, the student will be able to clearly identify which student behaviors need to be improved in order to increase their grade as well as receive an easy to read estimation of their expected final grade based on their behavior.

The user will have the control to alter what input parameters they choose to fill in as well as the value of those parameters. For example, a user may fill in their average study time parameter then run the model to receive an expected final grade. The user may then choose to raise or lower that study time parameter and run the model again in order to gauge the impact of that parameter on their output. Through this experimentation of multiple parameters, the user will be aided in making decisions on the optimal study times, absences, alcohol consumption, etc. that they should exhibit in order to receive their desired final grade. From the previous example, the user may realize higher study times correlate to a higher final grade. Through the model, they may then determine that if they maintain all their current behaviors and only increase their study time by 20%, they can get their desired grade.

Insights:

After setting up the model and running through the data set, our team was then able to analyze results as to which features will have the largest impact on a student grade and predict what grade they will receive based on their study habits . We found the largest correlation between the study hours a student inputs to how high their grade will be.

Work Cited

Aryanmsr. (2021, August 6). *Complete_ml_analysis*. Kaggle. Retrieved July 14, 2022, from https://www.kaggle.com/code/aryanmsr/complete-ml-analysis/notebook

Ansodariya, D. (2022, May 26). *Student performance dataset*. Kaggle. Retrieved July 14, 2022, from https://www.kaggle.com/datasets/devansodariya/student-performance-data

caro_caro. (2022, March 26). *Web app for linear regression by Streamlit*. Step-by-step to a Data Scientist. Retrieved July 14, 2022, from https://machine-learning.tokyo/web-app-for-linear-regression-by-streamlit/

Satyavishnumolakala. (2020, June 12). *Linear regression -Pros & Cons*. Medium. Retrieved July 14, 2022, from https://medium.com/@satyavishnumolakala/linear-regression-pros-cons-62085314aef0