RECOMMENDER SYSTEM

END TERM REPORT

*by*

**Group 11**

**Jayanth Tamminidi**

(Section: K18PG) (Roll Number: 40)

**Satya Chandra Reddy Sabbella**

(Section: K18PG) (Roll Number: 57)

**Dileep Reddy Digivinti**

(Section: K18PG) (Roll Number: 36)



**Department of Intelligent Systems**

**School of Computer Science Engineering**

**Lovely Professional University, Jalandhar**

**APRIL-2020**

**Student Declaration**

This is to declare that this report has been written by me/us. No part of the report is copied from other sources. All information included from other sources have been duly acknowledged. I/We aver that if any part of the report is found to be copied, I/we are shall take full responsibility for it.

Satya Chandra Reddy Sabbella

57

Jayanth Tamminidi

40

Dileep Reddy Digivinti

36

Punjab

05-04-2020

TABLE OF CONTENTS

1. Background and objectives of project assigned

2. Description of Project

3. Description of work division in terms of roles among students

BONAFIDE CERTIFICATE

Certified that this project report “RECOMMENDATION SYSTEM”is the bonafide work of “Jayanth, Satya, Dileep” who carried out the project work under my supervision.

<<Signature of the Supervisor>>

<<Name of supervisor>>

<<Academic Designation>>

<<ID of Supervisor>>

<<Department of Supervisor>>

1. Background and objectives of the project assigned -

Our project is the recommender system for the students, this project helps students analyse their academics. This project shows students where they’re lacking behind.

1. Description of project –

Modules used are- pandas, matplotlib, seaborn and sklearn.model\_selection

There are a total of 633 observations with 33 variables. Each row is one student with each column containing a difference characteristic. The Grade column is our target variable (also known as the response), which makes this a **supervised, regression** machine learning task. It’s supervised because we have a set of training data with known targets and, during training, we want our model to learn to predict the grade from the other variables. We will treat the grade as continuous which makes this a regression problem (technically the grade only takes on integer values so it is a [nominal variable](https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/)).

The grades are close to normally distributed with a mode at 11 (the grading scale in this school goes from 0–20). While the overall grades do not have a noticeable skew, it’s possible that students from certain categories will have skewed grades. To look at the effect of categorical variables on the grade we can make [density plots](https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0)of the grade distribution colored by the value of the categorical variable. For this we use the [seaborn](https://seaborn.pydata.org/) library and the kdeplot function.

As we saw from the plots, we don’t expect every variable to be related to the final grade, so we need to perform[feature selection](https://en.wikipedia.org/wiki/Feature_selection) (also called dimensionality reduction) to choose only the “relevant” variables. This depends on the problem, but because we will be doing linear modeling in this project, we can use a simple measure called the [Correlation Coefficient](https://onlinecourses.science.psu.edu/stat501/node/256) to determine the most useful variables for predicting a grade. This is a value between -1 and +1 that measures the direction and strength of a linear relationship between two variables.

These correlations seem to make sense at least by my rudimentary social science knowledge! failures is the number of previous class failures and is negatively correlated with the grade, as is absences, the number of absences from school. This negative correlation indicates that as these variables increase, the final grade tends to decrease (although we can only say this is a correlation and not that one variable causes another to decrease). On the other hand, both studytime, the amount of studying per week, and Medu the mother’s level of education, are positively correlated with the grade.

Correlations can only be calculated between numerical variables, so to find the relationship between categorical variables and grade, we have to one-hot encode the categorical variable and then calculate the correlation coefficient. [One-hot encoding](https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f?gi=af9695dcf290) is a process that creates one column for every category within a categorical variable.

We again see relationships that intuitively make sense: higher\_no represents the student does not want to go on to higher education and is negatively correlated with the grade with higher\_yes indicating the student does want higher education and showing a positive correlation.Mjob\_at\_home means the mother stays at home, and is negatively correlated with the grade while Mjob\_teacher indicates the mother teaches and has a positive correlation.

In this problem we will use these results to perform feature selection by retaining only the 6 variables that are most highly correlated with the final grade. 6 is sort of an arbitrary number that I found works well in the model, which shows that a lot of machine learning is just experimentation!

1. Description of work division in terms of roles among students-

Satya – Research, report and code design

Jayanth – Research and code design

Dileep – Research and code design