Team members:
Sushil Deore
Charan Kanwal Preet Singh
Max Solko
Vishadh Vilas Sawant

# STATS 5010-002    Project Proposal

1. What is the data that you plan to work with?
Ans) The data which we are working on is "Student Performance" dataset that pertains to
the academic performance of students in Mathematics and Portuguese from two
secondary schools in Portugal. The dataset comprises attributes like test scores, study habits,
demographic information, and family background.

2. Where did the data come from? Are they experimental or observational?
Ans) This data was collected from an online machine learning dataset repository called UCI
(University of California Irvine). This dataset is a combination of experimental and observational
data. While grades obtained in mid-term and final exams are experimental, demographic
information and family background are observational.

3. Why is this data interesting to you? What questions do you hope to answer about it?
Ans) The "Student Performance" dataset is fascinating because it can provide insights into the
factors affecting academic performance and can help build predictive models to
forecast academic success. Some of the questions that can be investigated using this dataset
include identifying the most influential factors on student performance, the impact of various
study habits on student performance, and whether academic performance can be predicted
using demographic and family background information.

4. What are the relationships between the variables? Does theory suggest that they are related
in some way?
Ans) Various relationships exist between the variables in the "Student Performance" dataset.
For instance, a strong positive correlation between mid-term and final exam grades
indicates that good performance in mid-term exams leads to good grades in final exams.
Moreover, family background variables, including parental education and occupation,
are positively correlated with student performance, indicating that children of
highly educated parents with prestigious jobs tend to perform better academically.

5. What random components are present (e.g., measurement error)?
Ans) The "Student Performance" dataset may have measurement errors due to inaccuracies in
recording student grades or demographic information. There could also be other random factors affecting
student performance that are not captured by the variables included in the dataset.

6. What prior research on your topic might be helpful to consider?
Ans) Previous studies on student performance and academic outcomes, such as the effects of
teaching methods, parental involvement, and socio-economic status, could be valuable
to consider while analyzing the "Student Performance" dataset.

7. What methods might be useful in analyzing this data?
Ans) To analyze the "Student Performance" dataset, statistical and machine learning
methods such as linear regression, logistic regression, decision trees, and random forests can
be used based on the research questions. Exploratory data analysis techniques like data
visualization can also be utilized to identify patterns and relationships in the data.