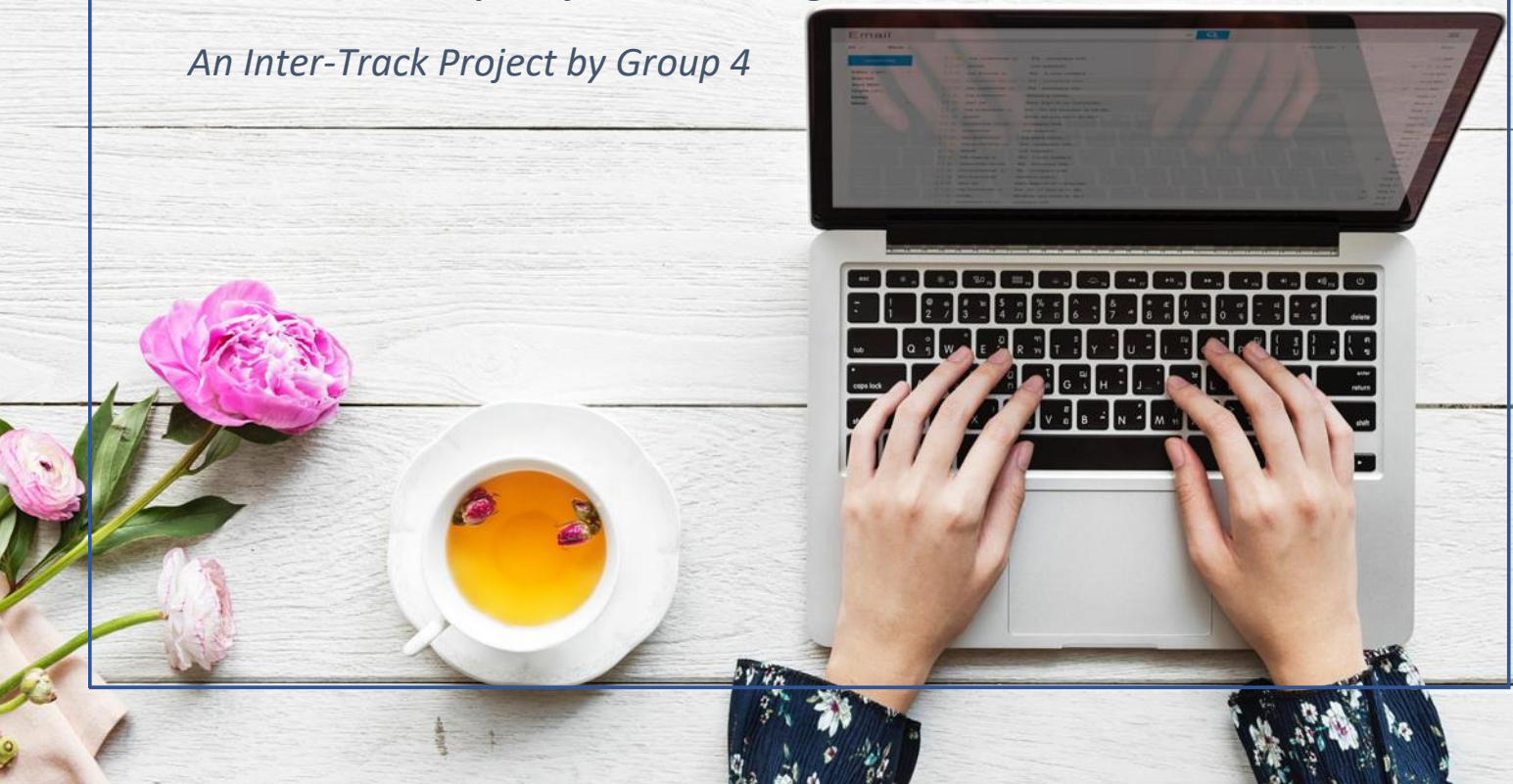
# *TECHNOLOGY JUSTIFICATION DOCUMENT*

JOY JOSEPH - DATA SCIENTIST

**Inter-Track Group Project: Predicting Student's Academic Performance**

*An Inter-Track Project by Group 4*



# TABLE OF CONTENT

# INTRODUCTION

Academic performance is the extent to which a student, teacher or institution has achieved their short or long-term educational goals. Cumulative GPA and completion of educational benchmarks such as secondary school diplomas and bachelor's degrees represent academic achievement. Wikipedia

Academic performance is the process of learning in an academic surroundings. Academic performance is seen to be the end result of the skills and intellectual abilities of a student in an academic surroundings. There is more to learning, it is more than Performance. Learning is defined as the expansion of a student's knowledge and skills that result from instruction and experience in an academic environment. Performance also consist of a student's ability to show that knowledge and skill in different settings and situations within and outside the classroom.

The Academic performance of students play an important role in the education system as well as the learning process. It is considered to be a major yardstick to judge one's total potentialities and capacities Nuthana & Yenagi (2009), which are frequently measured by the examination results. It is used to pass judgment on the quality of education offered by academic institutions.

Predicting students' performance is mostly useful to help the educators and learners improving their learning and teaching process.

# THE RATIONALE BEHIND THE TOOLS AND LANGUAGES USED ON THE PROJECT

**The Tools used in this project are;**

1. Jupyter -- Jupyter supports multiple languages like Julia, Python, and R. It is a web-application tool used for writing live code, visualizations, and presentations. Jupyter is a widely popular tool that is designed to address the requirements of Data Science.
2. **Python libraries such** as;

**Pandas** -a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

**Numpy** -is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

**Scikit-learn** -a key library for the Python programming language that is typically used in machine learning projects. It is very widely used across all parts of the bank for classification, predictive analytics, and very many other machine learning tasks.

**Matplotlib -**a plotting and visualization library developed for Python. It is the most popular tool for generating graphs with the analyzed data. It is mainly used for plotting complex graphs using simple lines of code. Using this, one can generate bar plots, histograms, scatterplots etc.

**The language used in this project** is Python. The reason python was chosen is because Python is a powerful, flexible, and easy-to-use language. Python has a lot to offer in terms of flexibility when it comes to implementing changes as machine learning practitioners need not recompile the source code to see the changes.

# MODEL PHASE

The modelling phase consists of applying different machine learning techniques to the dataset. The goal of the prediction is to create a model based on the students' current activities and accomplishments that attempts to predict learner failure and future performance. It is a typical classification problem, which a multi classification model can solve in order to predict whether a student can complete the program or not, that is, whether a student will graduate or not.

We will choose a model that is relevant to the task at hand. Since it is a classification problem, the model selected will be one suited for the categorical data.

Six models will be used to compare their accuracy scores. The model with the best score will then be chosen. The selected Models are;

1. **Gaussian Naive Bayes**
2. **Linear SVC**
3. **Decision Tree Classifier**
4. **Logistic Regression**
5. k-Nearest Neighbors
6. CatBoost

### Evaluating the Model Performance

Understanding how well a machine learning model will perform is the main purpose behind working with evaluation metrics. Metrics like **accuracy, precision, recall and f1 score** will be used to evaluate classification models for our balanced datasets and help it perform better in evaluating the model performance.