CMPT459 Project Proposal

Group members: James Qian (301416296), Luna Sang (301540164), Seulah Kim (301292814)

Database Name: Predict Students' Dropout and Academic Success

Database URL: UCI Machine Learning Repository - Predict Students' Dropout and Academic Success

Dataset description

For our project, we have selected the **Predict Students' Dropout and Academic Success** dataset from the UCI Machine Learning Repository. This dataset captures student-level information that can be used to predict academic outcomes such as dropout risk and success rates. Information is collected at the time of student enrollment and also includes the students' academic performance at the end of the first and second semesters from a higher education institution.

This dataset includes 35 attributes, a mix of numerical and categorical features. Some key features include demographic information like age, nationality, and gender, socio-economic factors such as parents' qualifications and occupation, and academic performance indicators like curricular units credited, grades, and dropout status. The target variable is whether a student graduates, drops out, or remains enrolled. The dataset includes a total of 4,424 samples, significantly more than the required 1,000 samples for this project, ensuring a robust data mining process.

Problem definition

At universities like SFU, student dropout rates are a concern, and early identification of at-risk students can help institutions provide timely support. Is there a way to know beforehand if a student will likely drop out of school or not? This project aims to predict whether a student is likely to drop out, stay enrolled, or graduate based on various demographic and academic factors. By training classification models using this dataset, we aim to help educational institutions identify at-risk students early and provide interventions that could improve retention and academic success. We will employ data mining techniques to develop predictive models. These models can help administrators and educators better understand the underlying factors that contribute to student dropout, enabling them to focus on the most critical factors that affect student retention.

Dataset Selection & Justification

This dataset is suitable for our project as it contains a good mix of numerical and categorical features, with over 4000 records, making it ideal for data mining tasks such as preprocessing, exploratory analysis, and classification. The target variable clearly differentiates between students who drop out, remain enrolled, or graduate, making it a good fit for classification models such as Random Forest, Support Vector Machines (SVM), and k-Nearest Neighbors (k-NN).

Furthermore, this dataset provides valuable insights into student retention, a critical issue in education. The inclusion of features such as parents' qualifications and occupation adds an important layer of socioeconomic context, as research has shown that parental background significantly impacts a child's educational achievements. This real-world relevance ensures that our project will not only fulfill academic requirements but also contribute to understanding factors influencing student success, making it a meaningful and impactful project.