**Project Proposal:** Deep Learning Based Model for Predictive Analysis of School Admissions

**Course:** CSCE 5218

# **Members:**

* Pooja Goyal
* Roshan Sah
* Sourabh Yadav
* [Sanjib Paudel](https://unt.instructure.com/groups/116357/users/195576)

# **Abstract**

Student educational score records consist of valuable information to help make fast and effective admission decisions. Every year making admission decision by going through single application to get score an input in the system manually and going through statement of purposes and letter of recommendations makes it really time consuming and tedious for the administrative staff that delays the decision-making process. In the era of digitalization, these problems can be solved by employing machine learning based models. Our role comes in here, we are proposing Deep Learning based model for predictive analysis of student profile for admissions decisions to help deal with scores data efficiently and fast.

# **Problem Statement:**

Graduate applications to graduate schools have risen significantly due to competitive job market, and students applying for admission, often face the issue of long processing times and delay in decision making processes. Admission committee members of the universities must have to deal with thousands of applications per admission cycle and respond to each application. The purpose of our project is to implement and compare performance to transformer-based models on labeled dataset based on applications sent by students to the universities. This model can be employed to ease down the complexity of graduate admission process. Furthermore, the proposed model might have capabilities to bring more transparency and reliability to the admission process for the students who puts their applications.

# **Objectives:**

1. To extract the student admissions data from public source
2. Pre-train the Deep Learning based model on small cohort of educational specific data
3. Fine-tune to predict student admission probability to any school
4. Evaluation of other multiple baseline models to compare performance on various benchmarks.

# **Dataset Employed:**

The proposed project revolves around building the predictive model for classification of graduate applications either accepted for admission or not. The dataset required for making the classification must contains the previous undergraduate records of the student, such as High School marks, number of subjects in undergraduate studies, cumulative GPA, total number of projects, any public project, number of internships, past work experience, etc. These attributes are mostly numerical, but if there will be any text-based attribute which can contribute to the classification, will be converted into the numerical form, i.e., vector form, by using the bag of word approach. The required dataset will be firstly analyzed for cleaning, which will be considered for exploratory analysis, and result of analysis will make the dataset fit for further usage. The model and model parameter selection will be selected on the basis of dataset exploratory analysis.

# **Proposed Approach/Solution:**

Diagram

Description automatically generated

The above-quoted figure gives a glimpse of the proposed approach to predict the probability of an admit school based on the scores and application materials.

For performing the analysis, the Anaconda framework, which uses python as language, can be used. Anaconda framework offers a lot of visualization features and a bulk of pre-defined libraries which can help to perform analysis in a better way.

Once the dataset is ready for training and utilization, it is very crucial to prepare some baseline models, which can further be optimized to prepare some hybrid and ensemble models. For the baseline perspective, Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, etc. can be prepared. Furthermore, few baseline neural network models or perceptron can be prepared. Also, training perspectives can be optimized by splitting the dataset into multiple sub datasets and performing separate training for each dataset. These baseline results can be further refined by collaborating the outputs from multiple baseline models. With this approach hybrid model can be prepared which gives final output based on results of multiple models.

For evaluation of model, accuracy parameter and confusion matrix can be used for binary classification.