USING MACHINE LEARNING AS A TOOL TO HELP GUIDE UNDECLARED/UNDECIDED FIRST-YEAR ENGINEERING STUDENTS TOWARDS A DISCIPLINE

Shashank Iyer and Chirag Variawa
shashank.iyer@mail.utoronto.ca, chirag.variawa@utoronto.ca

Abstract – Supervised Machine Learning classification algorithms are used to analyze the potential inclination of the undecided/undeclared first-year engineering students. The data exploration task is possible by building a dataset that comprises of questions based on significant attributes. These attributes hover around different disciplines of engineering being offered at the University of Toronto. This qualitative survey is distributed to upperclassmen students (3rd, 4th year and graduate students, N = 54) and undecided first-year engineering students (N = 29).

Multi-class classification is a technique that is used to categorize the data into two or more classes, in this case, the different disciplines of engineering at University of Toronto. The dataset that is built, based on the answers provided by the upperclassmen, is programmed into different classification algorithms such as Logistic Regression, KNN (K-nearest neighbors), Decision Tree and Random Forest classifier. The algorithms are compared so as to identify the most appropriate one that can determine the specific class label of the upperclassmen based on the answers provided in the qualitative survey.

The accuracy of the various algorithms is an indicator of the favorable algorithm that can serve as a tool to suggest the potential majors that could be pursued by the undecided/undeclared students. Moreover, the answers given by the upperclassmen is visually analyzed for identifying the patterns of inclination of the students belonging to different disciplines of engineering.

Keywords: Multi-class classification, potential inclination, engineering education, supervised machine learning, logistic regression, k-nearest neighbors, decision tree, random forest

1. INTRODUCTION

1.1. Background

An interview with Fritze Grupe at the University of Nevada reveals that 80% of matriculates enter college undecided on a major. [1] The process of choosing a major is often very complicated for most undecided students either because they are unable to identify their innate passion and skills, or they try to base their decision solely on the income generating capability associated with a specific major post-graduation. This paper investigates undecided/undeclared first-year engineering students, called the “TrackOne” program at the University of Toronto Faculty of Applied Science and Engineering. At a study conducted at Purdue University in 2005, enjoyment, interest and satisfaction are major factors associated with students’ success in Engineering. [2] A similar study suggests approximately 13% of First Year Engineering students leave the program before completing their first-year. [3] The results of these two surveys have been the primary motivation to determine where the interests of the students’ lie in order to hypothesize the potential majors they could pursue.

1.2. Problem Definition

Machine learning is the science of getting computers to act without being explicitly programmed. Machine learning is used in a wide range of applications such as image recognition, self-driving cars and practical speech recognition. [4]

Classification is a concept of supervised machine learning, that uses pattern recognition to identify a set of classes a new observation belongs to. This is determined on the basis of a training set whose categories are known. The classification algorithms being used are Logistic Regression, KNN, Decision Tree and Random Forest. These are trained on the answers to the qualitative survey given by the upperclassmen whose majors are known. The accuracies of the various models have been generated based on a comparison between the true labels available in the dataset and the predicted labels of the models. Finally, when a new observation from the dataset containing the answers of the undecided students is fed into the most accurate model, a hypothesized list of potential majors that a student could pursue is available.

Current literature found that the majority of students made the decision to declare their major based upon the resources within the university, family member
encouragement and knowing someone in the related field. [5] This research paper aims at finding an alternative technique to the current method used by the undecided students to choose a discipline of engineering that is based on multiple extrinsic factors. The problem definition that this paper focuses on is:

**Can supervised machine learning classification algorithms be a tool to guide undecided/undeclared first-year engineering students towards choosing a discipline?**

The answer to this question helps to hypothesize the probabilities of the choices of major an undecided student could pursue. In addition, it also throws light on the intrinsic inclination of the students pursuing different disciplines of engineering by analyzing key patterns in the binary response questionnaire.

### 1.3. Selected Methodology

The standard approach used in Machine Learning is to split the data into training and testing sets. The former comprises of an already known output and the model is trained on this data in order to fit the unknown data later on. The test dataset is made available to test the accuracy of the classification on this subset. The two possible outcomes of this process are either overfitting the model or underfitting the model. Either of the two outcomes can reduce the accuracy of the classification/prediction and is therefore not desirable.

In this case, the outcome is that we overfit the model because of the over-complex nature of it. A model is called complex when there are too many features (in our case, the range of questions in the qualitative survey) as compared to the number of observations. Basically, when this happens, the model learns or describes the “noise” in the training data instead of the actual relationships between variables in the data. [6]

To overcome this limitation, the technique of k-fold cross-validation is used. In this method, the dataset is split into k subsets (in our case, k = 10) and trained on k – 1 subsets. The subset which is held back is used for testing purpose. This process is repeated for each of the subsets and the average of the model against all the subsets is the predicted accuracy of the final model. In the later section a detailed description is provided regarding the enhanced accuracy of each model by using k-fold cross validation as compared to the standard train/test split.

### 2. METHODOLOGY

**Figure 1** outlines the steps that have been adopted to process the data, visualize and gather insights and trends in the decision-making process adopted by students from different disciplines of Engineering to answer the qualitative questions. Thereafter, the data is implemented in different classification models and the most appropriate one is chosen to test and hypothesize the probabilities of majors undecided students could possibly pursue.

![Fig. 1. Flowchart for data processing, data visualization, model implementation and testing](image)

#### 2.1. Qualitative Survey

The support of the work-study students in the First-Year Engineering Education Research team is taken to reach out to upperclassmen (3\textsuperscript{rd}, 4\textsuperscript{th} year and graduate students) and undecided first-year engineering students through social media applications such as Facebook. The survey is built using google form and it contains a set of 32 questions that hover around important attributes from different disciplines of Engineering (Chemical, Mechanical, Civil, Electrical & Computer, Industrial and Material Science Engineering).

The ethics approved survey being sent out to the upperclassmen students as a link to the google form contains only one personally identifiable question: *What is your major?* In case of the undecided/undeclared students, this question is not valid as they have not yet declared a major. Hence, it is omitted from the google form being sent to them, however, the remaining 32 questions are the same. The body of the survey that consists of the 32 questions requires a binary response (Yes/No) by both the upperclassmen and the undecided students.
2.2. Data Collection

Table 1 describes the sample questions from each discipline of engineering built around associated attributes of the various disciplines. The disciplines per se do not rigidly compartmentalize the questions, as these questions can address an intersection of the attributes of two-three disciplines. The benefit of this being the ability to portray an option of two-three probable disciplines that are based on the innate interests and skill-set of the students.

Table 1: Different disciplines, associated attributes and sample questions

| Discipline of Engineering | Attributes                                                                 | Sample Question based on the specific attributes |
|---------------------------|---------------------------------------------------------------------------|--------------------------------------------------|
| Mechanical Engineering    | Physical principles of objects design; appearance, working of individual components, manufacturing process, safety | Q. Does the working of the gear box of your bike interest you? |
| Electrical & Computer Engineering | Hardware, Control, Machine Learning, Computer vision, Electromagnetism, Bioelectronics | Q. Would you ever disassemble a computer and rebuild it from the scratch? |
| Civil Engineering         | Design, construction & maintenance of structures and infrastructure, transportation, alternative energy systems | Q. When you see articles on the growing natural disasters, do you wonder about the possibilities in the next generation building design? |
| Chemical Engineering      | Environmental footprint, sustainable processes, human health, renewable fuels and materials | Q. When you shop for clothes do you consider issues such as ecological footprint before making the buy? |
| Material Science Engineering | High efficient energy storage, conversion technology, sustainable energy sources, alternative materials for industry manufacturing | Q. Do you think of ways to harness from readily available materials for healing homes? |
| Industrial Engineering    | Human interactions with technology and systems, organizational safety, efficiency & profitability | Q. Do you think ergonomically designed libraries attract students better? |

2.3. Data Cleaning

The data collected from the qualitative survey includes input variables (Yes/No) and the output variables (Mechanical, Industrial, Chemical, Civil, Material Science, Electrical & Computer). As these variables contain specific labels, they are called categorical variables. Categorical variables are of two types: Nominal & Ordinal. In this case, these represent nominal categorical variables as they do not have a natural ordering of values.

For the data cleaning process in this project, it is imperative to convert the categorical variables (input & output) to numerical variables (integer encoding/dummy variables) because many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. [7]

This qualitative data is converted into a series of 0’s and 1’s by using the pandas and scikit learn libraries in Python in order to quantify and compare the data easily. For the input variables (Yes/No), 1 signifies Yes and 0 signifies No. For the output variables, the integers assigned are: Chemical Engineering – 0, Civil Engineering – 1, Electrical & Computer Engineering (ECE) – 2, Industrial Engineering – 3, Material Science Engineering (MSE) – 4, Mechanical Engineering – 5.

Figure A1 in the appendix represents a sample of the qualitative data collected before data cleaning. Figure A2 in the appendix illustrates the output of the process of converting the categorical variables to numerical variables using the above-mentioned libraries.

2.4. Data Visualization

Data exploration is a valuable step that needs to be performed before implementing the data in various models. It is important to perform this step so that valuable insights and trends can be deduced that would lead to a deeper understanding of the data collected.

Figure 2 shows a sample question that was built using an attribute of Mechanical Engineering ‘working of individual components.’ Based on the results of the survey and using the seaborn library in Python, the responses (Yes: 1/No: 0) for each discipline of engineering is plotted, where the disciplines are shown on the horizontal axis and the counts of the Yes/No are shown on the vertical axis. The blue bars in the figure for each discipline of engineering indicate the total count of the Yes response and the red bars represent the total count of the No response.

An obvious analysis that can be drawn from a close study is that, students pursuing Mechanical, ECE and Industrial Engineering have largely responded to this question in Yes. In contrast, those studying Civil Engineering have responded with an overwhelming No response. The gap between the Yes and No response is miniscule in the case of MSE. However, for Chemical Engineers the response is at par with each other. It can be deciphered from this that students in Mechanical Engineering responded to this question along the attribute mentioned in figure 1 ‘working of individual components.’ Similarly, students studying ECE and Industrial Engineering used the attributes ‘hardware’ and ‘human interactions with technology and systems’ while responding to this question. Another conclusion that can be drawn is that, these disciplines of engineering are at the intersection of their respective attributes which can be seen on the basis of their responses.
Figure 2. Bar plot representing the responses of different disciplines to a question associated with mechanical engineering.

Figure 3 portrays a sample question that was built using an attribute associated with ECE, ‘hardware’. It can be observed that in case of ECE the response received has been distinctly in favor of the Yes option which goes hand in hand with its associated attribute. Thus, proving their clear-cut inclination. Along the same lines, students studying Mechanical Engineering have also rated the Yes response at a higher count proving that any hands-on work is of interest to them too. In contrast, students studying Industrial, Civil and Chemical have responded to this with more No’s than Yes and in the case of MSE students have just responded with a No.

Figure 4 contains the sample question that constitutes an attribute associated with MSE, ‘sustainable energy sources.’

In accordance with this, it is visible that the students of MSE are largely in agreement with the Yes response. In addition, the Industrial Engineering students as well, follow the same pattern of answering, which is congruent with the attribute ‘human interactions with technology.’

However, those pursuing ECE, Civil and Mechanical show a reversal of trend for this question. The students of Chemical Engineering have given a neutral response to this question indicating that there could be an overlap with its associated attribute ‘sustainable processes.’

2.5. Model Overview

Logistic Regression is the most commonly used classification algorithm for two-class classification (binary classification) and multinomial classification. In this case, because the target variable (majors) consisted of multiple classes, it is a multinomial classification scenario. Logistic Regression works on the premise of estimating the relationship between dependent binary variables and independent target variables.

K-Nearest Neighbor classifier is one of the most suitable choices for a classification study where there is limited knowledge about the data. Being a non-parametric algorithm, there is no assumption for underlying data distribution. Thus, the model structure is determined from the dataset. [8]
Decision Tree classifier can be represented in a tree-like structure where the internal node denotes the feature, the branch depicts the decision rule and each leaf node is the outcome. The node at the apex of the decision tree is called the root node and it learns to partition on the basis of the feature/attribute value. [9]

Random Forest Classifier is an ensemble technique of decision trees and it is called a forest as it comprises of a collection of decision tree classifiers. In a classification problem, each tree votes and the most popular class is chosen as the final result. [10]

2.6. Model Implementation

The dataset post cleaning and visualization is implemented in four different models: Logistic Regression, K-Nearest Neighbor, Decision Tree Classifier and Random Forest Classifier. For each model, the dataset containing numerical variables (input & output) is implemented using the train/test split methodology and 10-fold cross validation. In each case, it is shown that the model performs better using 10-fold cross validation over train/test split method. Figure A3, A4, A5 and A6 in the appendix, depict the enhanced accuracy of each model using the 10-fold cross validation method.

2.7. Model Testing

Table 2 shows the different models used and the variability in the accuracy of predicting the majors for the dataset containing the information of the upperclassmen students.

Table 2: Comparison of the accuracies of the different models

| Model                  | Accuracy with Train/Test split | Accuracy with 10-Fold Cross Validation |
|------------------------|--------------------------------|----------------------------------------|
| Logistic Regression    | 36.36%                         | 48.78%                                 |
| K-Nearest Neighbor Classifier | 36.36%                         | 50.78%                                 |
| Decision Tree Classifier | 27.27%                         | 42.53%                                 |
| Random Forest Classifier | 36.36%                         | 57.53%                                 |

It is evident from Table 2 that, the Random Forest classifier performed the best of all the models because it is a collection of the number of decision trees. The accuracy is the highest as it does not undergo the overfitting problem. This is because it takes into consideration the average of all the predictions, thereby, cancelling out the biases.

3. RESULTS

In order to suggest the possible majors that undecided/undeclared students could pursue, the random forest classifier is the best choice as it achieved an accuracy of approximately 58% using the 10-fold cross validation method. The predict_proba syntax in Python is used to achieve this. A snippet of the code is shown in figure A7 in the appendix. From the sample set of 29 undecided/undeclared students, student number 27 is randomly chosen to test the hypothesized probabilities of majors using the above syntax. Table 3 represents the outcome of the result.

Table 3: Probable majors and percentage probability

| Student number 27 | Percentage probability of the majors |
|-------------------|--------------------------------------|
| Industrial Engineering | 70%                             |
| ECE                | 20%                                 |
| Mechanical Engineering | 6.67%                            |

It can be observed from Table 3 that the hypothesized probable choice of majors from highest to lowest on the basis of the findings are Industrial Engineering, ECE and Mechanical Engineering. This reflects the pattern of choice adopted by student number 27 to answer the set of 32 questions which is largely along the along the lines of the attributes of Industrial Engineering.

4. LIMITATIONS

The accuracy of 57.53% which has been achieved using the random forest classifier has greater room for improvement by using a larger window for prediction, using more precise and accurate predictor data and with a further complex trend analysis.

The prediction window can be enlarged by recruiting a greater number of participants (upperclassmen students) and by using more relevant data. Inclusion of other powerful and accurate predictors would result in the random forest algorithm getting closer to a model with greater accuracy. Furthermore, a deeper analysis of the data would enable the machine learning algorithm to estimate better predictions.

5. CONCLUSION

In this exploratory study, the methodology adopted for addressing the problem is detailed along with the description of the qualitative survey to be distributed among various participants. (54 upperclassmen students and 29 undecided/undeclared students). Thereafter, the
process being adopted to collect, clean and visualize the data is discussed. Subsequently, the models being used are outlined along with the steps needed to implement and test the model. Before conclusion, the results of the report are discussed along with certain limitations while using this approach.

With respect to the future work it is suggested that more data should be acquired along with its deeper analysis in order to increase the accuracy of Machine Learning algorithms. By achieving a greater level of accuracy, this study could prove to be a beneficial tool to be used by the advisors of undecided students and the students themselves. Thereby, making the decision of choosing a major at the end of the first-year a less daunting process.

6. ACKNOWLEDGEMENTS

This Master of Engineering project has been a very beneficial experience to me as it helped to sharpen my data collection, analysis and machine learning skills. I would like to express my gratitude to Dr. Chirag Vartiwa, Director of the First Year Engineering curriculum for his constant support and inspiration. The acknowledgement will be incomplete without thanking the First Year Engineering Research team for extending their valuable time and helping me recruit the right participants.

7. REFERENCES

[1] Ronan, G. B. (2005, November 29). College freshmen face major dilemma. Retrieved January 8, 2019, from http://www.nbcnews.com/id/10154383/ns/business-personal_finance/t/college-freshmen-face-major-dilemma%26U_hRs9Kk6Y

[2] Follman, D., Bodner, G., & Hutchison, M. A. (2005). Shaping The Self-Efficacy Beliefs Of First-Year Engineering Students: What is the Role We Play? Proceedings of the American Society for Engineering Education Annual Conference. Portland, OR. Retrieved January 8, 2019, from https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1011&context=eneg

[3] Ngambeki, I., Dulrymple, O., & Evangelou, D., Dr. (2009). Decision Making In First Year Engineering: Exploring How Students Decide About Future Studies And Career Pathways. Retrieved January 8, 2019, from https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1003&context=eneg

[4] Ng, A. Machine Learning. Retrieved February 14, 2019, from https://www.coursera.org/learn/machine-learning

[5] Pasquarella, K. (7/8/2013). Undecided students: A study of decision-making styles and choosing a college major at Rowan University. Retrieved February 14, 2019, from https://rdw.rowan.edu/cgi/viewcontent.cgi?article=1390&context=etd

[6] Bronshtein, A. (2017, May 17). Train/Test Split and Cross Validation in Python – Towards Data Science. Retrieved February 14, 2019, from https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

[7] Brownlee, J. (2018, May 19). Why One-Hot Encode Data in Machine Learning? Retrieved February 14, 2019, from https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[8] Navlani, A. (2018, August 2). KNN Classification using Scikit-learn. Retrieved February 14, 2019, from https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

[9] Navlani, A. (2018, August 2). Decision Tree Classification in Python. Retrieved February 14, 2019, from https://www.datacamp.com/community/tutorials/decision-tree-classification-python

[10] Navlani, A. (2018, August 2). Random Forests Classifiers in Python. Retrieved February 14, 2019, from https://www.datacamp.com/community/tutorials/random-forests-classifier-python

8. APPENDIX

A.1 Qualitative data collected before performing data cleaning

| What is your major?          | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 | ... | Q23 | Q24 |
|-----------------------------|----|----|----|----|----|----|----|----|----|-----|-----|-----|
| Industrial Engineering      | Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| ... | Yes | Yes |
| Civil Engineering           | Yes| No | Yes| Yes| Yes| Yes| No | No | No | ... | Yes | Yes |
| EOE                         | Yes| No | Yes| No | Yes| No | Yes| Yes| Yes| ... | No  | No  |
| MSE                         | Yes| No | Yes| No | Yes| No | Yes| No | No | ... | Yes | Yes |
| EOE                         | Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| Yes| ... | Yes | Yes |
| EOE                         | No | No | No | Yes| No | Yes| No | No | No | ... | No  | No  |
| EOE                         | Yes| No | Yes| Yes| No | Yes| Yes| Yes| Yes| ... | No  | No  |
| EOE                         | No | No | No | Yes| No | Yes| No | No | No | ... | No  | No  |
| EOE                         | Yes| No | Yes| No | Yes| No | Yes| No | No | ... | No  | No  |
| Chemical Engineering        | Yes| Yes| Yes| No | No | No | No | No | No | ... | No  | No  |
| Industrial Engineering      | Yes| Yes| No | Yes| Yes| Yes| Yes| Yes| Yes| ... | No  | Yes |
| Industrial Engineering      | Yes| No | No | Yes| No | Yes| No | Yes| No | ... | Yes | No  |
A.2 Qualitative data after performing data cleaning

| MAJOR | Q1_score | Q2_score | Q3_score | Q9_score | Q24_score | Q1_No |
|-------|----------|----------|----------|----------|-----------|-------|
| 3     | 1        | 1        | 1        | 1        | 1         | 0     |
| 1     | 1        | 1        | 0        | 0        | 1         | 0     |
| 2     | 1        | 1        | 1        | 0        | 0         | 0     |
| 4     | 1        | 0        | 1        | 0        | 1         | 0     |
| 2     | 1        | 1        | 1        | 1        | 1         | 0     |
| 2     | 0        | 0        | 0        | 0        | 0         | 1     |
| 2     | 1        | 1        | 0        | 1        | 0         | 0     |
| 2     | 0        | 0        | 1        | 1        | 1         | 0     |
| 2     | 0        | 0        | 1        | 0        | 1         | 0     |
| 0     | 1        | 1        | 0        | 0        | 0         | 0     |
| 3     | 1        | 1        | 1        | 1        | 0         | 0     |
| 3     | 1        | 0        | 0        | 1        | 0         | 0     |
| 5     | 1        | 1        | 1        | 0        | 0         | 0     |
| 2     | 1        | 1        | 1        | 1        | 0         | 0     |
| 5     | 1        | 1        | 1        | 1        | 0         | 0     |
| 2     | 1        | 1        | 1        | 0        | 1         | 0     |
| 2     | 1        | 1        | 1        | 1        | 1         | 0     |
| 5     | 1        | 1        | 1        | 1        | 0         | 0     |
| 3     | 1        | 1        | 1        | 0        | 1         | 0     |
| 2     | 1        | 1        | 1        | 1        | 0         | 0     |
| 3     | 1        | 1        | 1        | 0        | 1         | 0     |
| 3     | 1        | 1        | 1        | 0        | 1         | 0     |

A.3 Logistic Regression Accuracy

Logistic Regression

```python
from sklearn.metrics import accuracy_score
accuracy_score(Y_test, predictions) * 100
36.30636363636363
print('Accuracy: 36.306% (% accuracy score: predictions, Y_test) * 100)
Accuracy: 36.30%
```

```python
from sklearn.cross_validation import cross_val_score
print(cross_val_score(lgm_model, X, y, cv=10, scoring='accuracy').mean())
0.4878571428571429
```

A.4 KNN Classifier Accuracy

KNN Classifier

```python
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)
metrics.accuracy_score(Y_test, y_pred)
0.36363636363636365
```

```python
print(cross_val_score(knn, X, y, cv=10, scoring='accuracy').mean())
0.5078571428571428
```

A.5 Decision Tree Classifier Accuracy

Decision Tree Classifier

```python
from sklearn import tree
model = tree.DecisionTreeClassifier()
model.fit(X_train, Y_train)
y_predict = model.predict(X_test)
print(accuracy_score(Y_test, y_predict))
print(cross_val_score(model, X, y, cv=10, scoring='accuracy').mean())
```

```python
0.36363636363636365
```

A.6 Random Forest Classifier Accuracy

Random Forest Classifier

```python
from sklearn.ensemble import RandomForestClassifier
random_forest = RandomForestClassifier(n_estimators=30, max_depth=10, n)
random_forest.fit(X_train, y_train)
y_pred = random_forest.predict(X_test)
print(accuracy_score(Y_test, y_pred))
print(cross_val_score(random_forest, X, y, cv=10, scoring='accuracy').mean())
```

```python
0.36363636363636365
```

A.7 Probable majors for Student Number 27

Hypothesized probability for student number 27

```python
random_forest.predict_proba(np.array([X_new_test.iloc[27]]).reshape([1,1]))
```

```python
array([[0.0000000, 0.0133330, 0.2 , 9.7 , 0. , 0.0666667]])
```