PREDICTING HOMESICKNESS IN INTERNATIONAL STUDENTS AND FINDING POTENTIAL FACTORS

Sushmitha Suresh and Yingnan (Jordan) Guo,
Department of Management Sciences,
University of Waterloo

Abstract

Homesickness is a problem among students at many universities. Few studies have examined this phenomenon and its association with student's relationship with their family, friends, other international students and academic stress. Employing a cross-sectional survey methodology with stratified sampling at the University of Waterloo, we try to predict whether or not a student is prone to Homesickness based on a number of demographic variables about the student and environmental variables about the University. We also perform algorithms to find out the variables that show strong association with the feeling of Homesickness.

1 Introduction

With Canada's immigration rate increasing exponentially every year, the number of international students joining the Canadian Universities has also increased proportionally. One of the underrated emotional problems that International students often face in the University is the feeling of homesickness. Staying away from family and friends and not able to meet them regularly causes a sensation of grief and feeling of longing for one's home. While a lot of Universities in Canada have a dedicated department to cope with general psychological issues of students, homesickness is often seen as a smaller problem in comparison to issues like Anxiety and Depression [4]. In general, students who experience homesickness often face challenges academically, socially and emotionally as they wrestle with their new life [5]

In this study, we try to predict the level of homesickness that a student experiences based on the variables that explain about the student and his/her relationship with the University, parents, siblings and friends. In addition to predicting the level of homesickness, we try to find out the variables that largely contribute to the higher levels of homesickness. The variables that we use in our models are both demographic and environmental variables in order to explain the variations in levels of homesickness better. Understanding environmental factors that affect homesickness would benefit teachers and administrators who are constantly challenged with students experiencing symptoms of homesickness.

2 Related Work

For International students, the transition from their homeland to a foreign country not only poses a physical adaptation but also increases the need for psychological and mental ability to adapt to the new situation. Feeling of homesickness can lead to other mental issues like Anxiety and Depression. Studies show that this also causes students to poorly focus in their academic performances, leading to increased number of dropouts among international students [4].

Over the last decade the number of students studying abroad has increased 150% to more than a quarter of a million. The programs in which these students participate are no longer seen as simply a campus extension of academic exercise, but as an overall educational experience that develops holistic life skills in the participants. Psychological adjustment as measured by homesickness is an indicator of the success of abroad experience and student development [3]. A key component in addressing homesickness is social connectedness [2]. Socializing with local students prevents homesickness, as the more inter-national students mingle with local students, lesser is the homesickness they experience [7].

In addition, extracurricular involvement during the first year of college decreases homesickness among underrepresented (i.e., first generation, underrepresented ethnic/racial minority, and low income) students attending an elite, predominantly White institution [8].

Also, the demographics variables play a moderation role in emotional intelligence, homesickness and the development of mood swings in University Students. Among the external factors affecting the homesickness in International students, there is important internal factor that contributes to the grief, which is the cell phones [5].

3 Methodology

3.1 Sample

We collected data through cross-sectional survey in which the participants were students from University of Waterloo. The survey was distributed to a total of 131 students pursuing master's degree and enrolled in three courses from the department of Management Sciences, MSCI 623 (Big Data Analytics), MSCI 641 (Text Analytics) and MSCI 719 (Operational Analytics). In order to have diversity in data and to have undergrad and PhD students, the survey was also shared with members of University pages on social platforms, reddit (r/uwaterloo) and twitter (@UWaterloo). The total number of responses we received at the end of the survey deadline was 116, in which 43% were female, 54% were male and 3% were from other categories. In the category of educational level there were 52% undergraduate students, 44% graduate students and 4% PhDs, among whom, 48% were international students and 52% were local students.

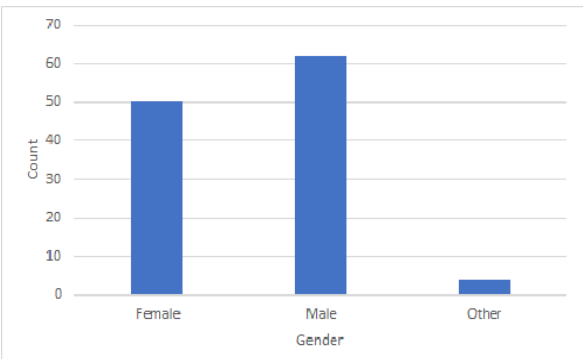

Fig 1: Distribution of Gender

Among the participants, 76% were between the age 18 and 25, 20% were between 26 and 30 and 4% were above 30.

3.2 Variables

The questions in the survey were framed to obtain a specific set of variables based on the previous work done in the area of homesickness. The two sets of variables collected were demographic and environmental. The demographic set included variables such as age, gender, number of siblings and friends, number of hometown visits in a month and number of times students cook their authentic food to feel like home. The environmental set of variables were included to understand the role of University in student's life. The variables in this set included number of events at the University to network with other students, financial aid provided to International Students and rate of difficulty of the program.

In order to predict homesickness in students, we needed a labelled dataset. The target variable homesickness, which was a part of demographic set, had four labels based on how often the student thought of their home in a week.

4 Data Cleaning and Pre-Processing

The variables collected were all categorical and required exhaustive pre-processing before feeding them into the models. We started with the elimination of NULL and NaN records from the dataset. There was only one record with NaN which was removed. There were several questions in the survey that had the same options.

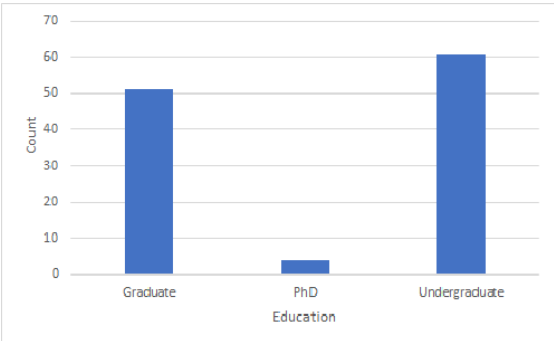

Fig 2: Distribution of Educational level

Fig 3: Distribution of Status

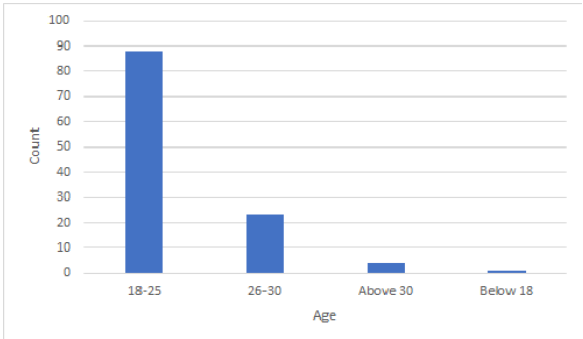

Fig 4: Distribution of Age

For example, questions like *How many siblings do you have?* and *How many friends do you have in the University?* had the same options which were *[0,1,2,3,4 or more]*. In order to uniquely distinguish among the variables, we replaced the answers with the corresponding variable name. That is, number of siblings were changed **from ['0' '1' '2' '3' '4' '5 or more'] to [‘Single child’, 'One sibling' 'Two siblings' ‘Three siblings’ 'Large family', 'Large family’]** respectively and number of friends were changed from **[0, 1, 2, 3, 4, ‘5 or more’] to ['Lonely', 'Lonely', 'Introvert', 'Friendly', 'Friendly', 'Extrovert']**. This process is neither one-hot encoding nor binning (the numbers were received as characters) but a simple change of texts to uniquely represent the variables they belong to.

Similarly, answers to *How difficult is your coursework?* was changed from **[1 2 3 4 5] to [Easy, Easy, Moderate, Moderate, Difficult].** The answers to *How often the University conducts events to network with other students?* was changed from **[‘I don't know’, 'Once', 'Twice', 'Thrice or more'] to ['Does not know about networking', 'Networking happens once', 'Networking happens twice', 'Networking happens often']** respectively.

For questions *How often do you cook your authentic food?* and *How often have you visited your home since your program started?* the answers were changed from **['Thrice' , 'Thrice or more', '2 to 3 times' , 'Everyday', '4 to 6 times', 'Never'] to [‘Cooks thrice’, 'Cooks very often', 'Cooks twice', 'Cooks everyday', ‘Cooks very often’, 'Never cooks']** for cooking and from **['Once', 'Never', 'Twice', 'Thrice or more'] to ['Never visits', 'Visits twice', 'Visits once', 'Visits often']** for hometown visits. Homesickness is more prevalent in International students who stay away from their home than the local students. The locals, irrespective of their place of stay, have the luxury to visit their family anytime as they like. Hence, from the dataset we filtered out the records of local students. Since 52% were local students, filtering out these records reduced the data size from 116 to 55.

All variables are converted to unique values; however, they still are categorical. In order to perform exploratory data analysis, we performed encoding and converted the categorical data into numerical values. After the encoding the data in the dataset looked like Table 1.

5 Exploratory Data Analysis

After the data was cleaned and preprocessed, we went on to explore and analyze the relation of each of the variables with the outcome variable homesickness.

We observed (Fig 5) that majority of men (1) lie on the level 2 of Homesickness which denotes *missing home every day*. And women (0) were seen equally on levels 2 and 0, indicating that most women *miss their home between 3 and 7 days a week.* (Fig 6) Students who never visited their house ever since their program started with feel homesick *all the time* (2) *or rarely* (0).

Other demographic variables like Siblings (Fig 7) and Friends (Fig 8) showed that students with two siblings (4) were spread across all levels of Homesickness equally and students who are extroverts (0) missed their home almost every day (2). Also, most students with zero or one sibling were extroverts (Fig 9).

Although new friends promote the adaptation process, keeping feelings of homesickness at bay, our datapoints showed that students from a smaller family tend to be extroverts and in turn missed their home every day.

	Categories	Encodings
Age	18-25	0
	Below 18	3
	Above 30	2
	26-30	1
Gender	Male	1
	Female	0
	Other	2
Education	Grad	0
	Undergrad	2
	PHD	1
Course difficulty	Moderate	2
	Easy	1
	Difficult	0
Network events	Does not know	0
	Happens once	2
	Happens twice	3
	Happens often	1
Financial aid	No aid	2
	Aided	1
	Doesn't know	0
Friends	Friendly	1
	Lonely	3
	Introvert	2
	Extrovert	0
Siblings	One sibling	1
	Two siblings	4
	Single Child	2
	Large Family	0
Hometown visits	Never	0
	Visits twice	3
	Visits once	2
	Visits often	1
Cooking	Cooks very often	2
	Cooks twice	1
	Cooks everyday	0
	Never cooks	3
Homesickness	Twice	4
	4 to 6 times	1
	2 to 3 times	0
	Everyday	2
	Never	3

Table 1: Categories and Encodings

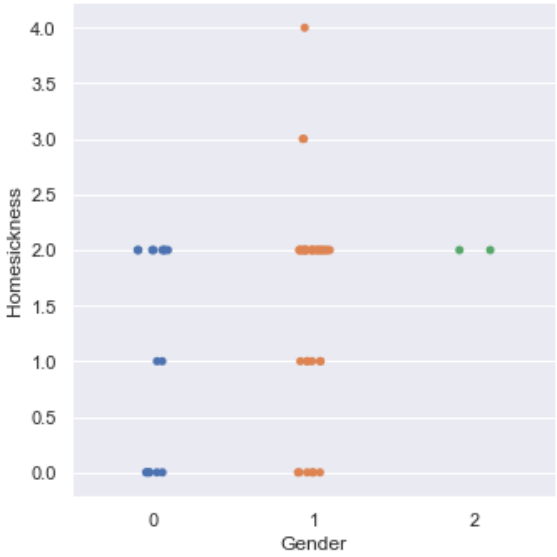

Fig 5: Homesickness vs Gender

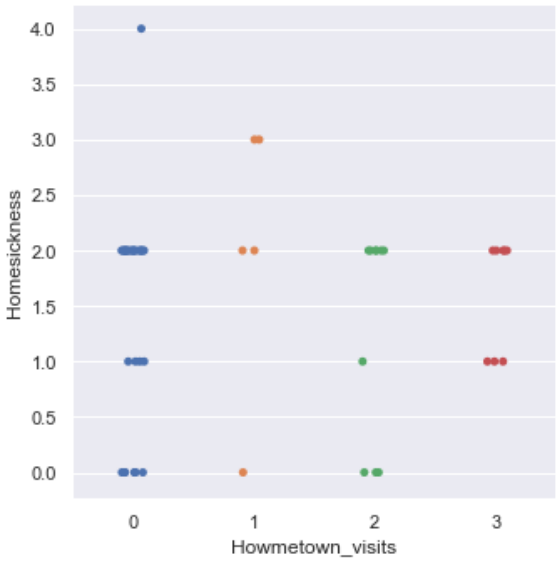

Fig 6: Homesickness vs Hometown_Visits

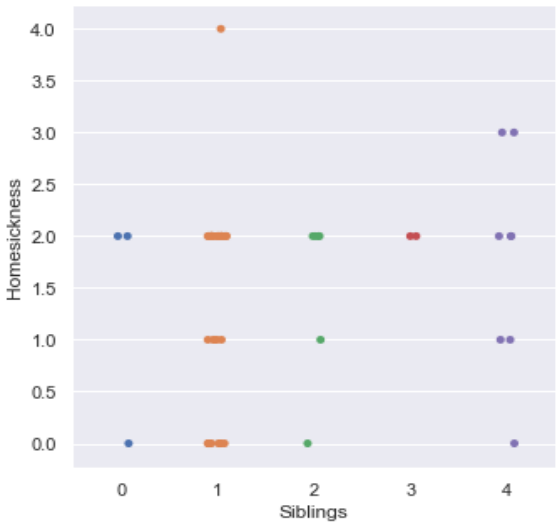

Fig 7: Homesickness vs Siblings

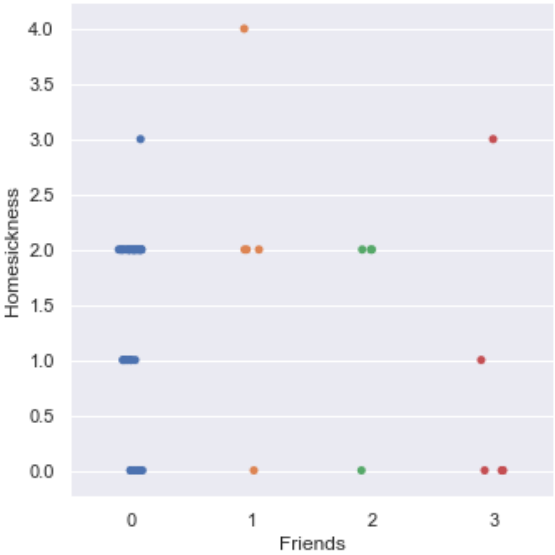

Fig 8: Homesickness vs Siblings

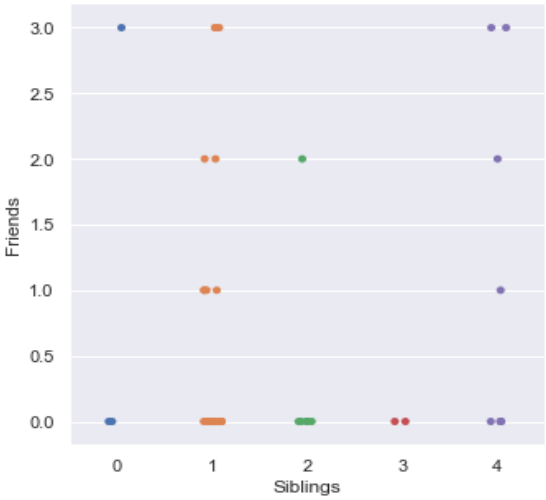

Fig 9: Friends vs Siblings

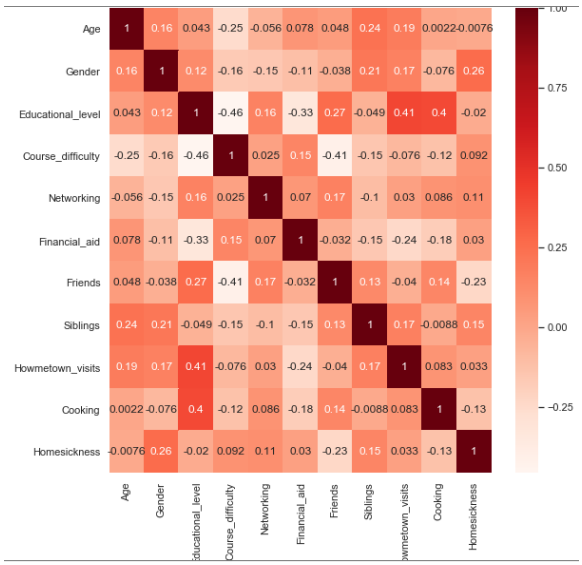

Fig 12: Heatmap of all the variables

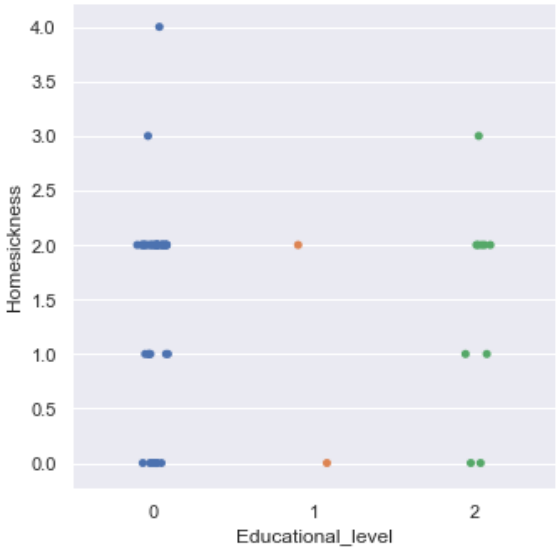

Fig 10: Homesickness vs Course_difficulty

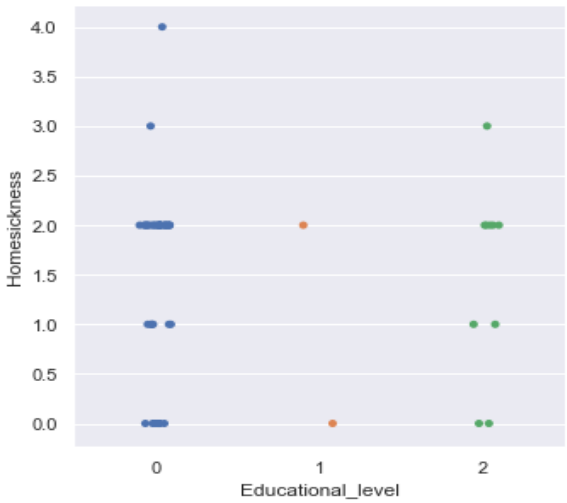

Fig 11: Homesickness vs Educational_level

It looks, from the graph that there could be a possibility that students tend to have more friends because they miss their home every day, thereby showing a strong positive relation between extroverts and increased levels of homesickness. However, it is hard to state anything conclusive given the limited data we have.

The environmental variables like difficulty of the course (Fig 10) and the level of Education (Fig 11) also affect the feeling of homesickness. It is interesting to note that students with most difficult courses fall in the levels 0 and 3 of homesickness, which means they never or rarely miss their homes. It is possible that a huge academic load might take student's mind off of their family and

make them focus completely on the coursework. Likewise, the PhD students are seen only in two levels, either miss home every day (2) or very rarely (0).

With each of the variables showing varying levels of association with homesickness, we needed a better visualization of the variables and their corresponding strength of relation with homesickness to decide which variables are the most suited to predict the outcome variable in our models. The heatmap map (Fig 12) showed that almost all variables show minimal correlation between each other.

6 Model Creation

6.1 Baseline Model

We used 4 classifiers for our predictive analysis:

- Logistic Regression
- Decision Tree Classifier with Gini
- Decision Tree Classifier with entropy
- Multinomial Naïve Bayes Classifier.

All the models were used without any parameters. Before training the models on the dataset, the rows were all shuffled to fetch better accuracy. In order to keep the results consistent, seed value was set to 300. The dataset was split into train and test set in the ration 80:20 respectively.

Each model was trained on the train set to evaluate the corresponding 10 cross validation score. The model that gave the highest cross validation score was run on the test set to predict the values of homesickness. This was our baseline model.

Classifiers	Accuracy
Logistic Regression	77
DT with Gini	78
DT with entropy	80
Naive Bayes	77

Table 2: Cross validation score before parameter tuning

Naïve Bayes makes the naïve assumption that the variables are independent of each other (no multicollinearity). It is no wonder that accuracy of NB is lesser than Decision tree. However, the accuracy is not the lowest and is same as the regression because there was never a strong correlation between any of the predictors (Fig 12). Gini and Entropy gave different accuracy which is very unlikely. However, given the size of data and the variation in the percentage of each of the classes (Table 3), entropy might have performed better as Gini impurity is maximal if the classes are perfectly mixed. Gini tends to find the largest class, and entropy tends to find groups of classes that make up ~50% of the data. Level 2 of Homesickness variable forms 59% of the data, by being the largest class. We would be able to gauge the performance of the criteria better with a larger data. But for now, we had considered the criterion that gave better accuracy which is the Entropy.

Homesickness	% of each class
2 to 3 times	21.82
4 to 6 times	14.55
Everyday	58.18
Never	3.64
Twice	1.82

Table 3: % of each class

Now that we had our models without any parameters and the corresponding validation scores, we ran these models on the test set which gave the accuracy as in Table 4.

Baseline model	Accuracy
Logistic Regression	64
DT with Gini	72
DT with Gini	72
Naive Bayes	54

Table 4: Test set accuracy in baseline model

6.2 Hyperparameter Tuning

Given the accuracy of baseline model, we then tried to tune the parameters of Logistic Regression, Decision Tree and Naive Bayes using GridSearchCV. At the end of Grid Search, four sets of parameters one for each mode was obtained.

	Parameters	Values
Logistic Regression	Penalty	L2
	C	1
	class_weight	{1: 0.5, 0: 0.5}
Decision Tree with Gini/Entropy	max_depth	3
	min_samples_split	10
Naïve Bayes	Alpha	0.73

Table 5: Hyperparameters of Classifiers

6.3 Ablation Study

The final models we had after parameter tuning were:

- ***LogisticRegression (random_state=0, solver='saga', multi_class='multinomial', penalty = 'l2', class_weight = {1: 0.5, 0: 0.5})***
- ***DecisionTreeClassifier (criterion = "gini", max_depth = 3, min_samples_split = 10)***
- ***DecisionTreeClassifier (criterion = "entropy", max_depth = 3, min_samples_split = 10)***
- ***MultinomialNB(alpha=0.73)***

The above models were again run on the training sets in 10 cross validation to check if the cross-validation score had improved after the hyperparameter tuning. At the cross validation, we observed the following scores for each of the models.

Classifiers	Accuracy
Logistic Regression	79
DT with Gini	80
DT with entropy	75
Naive Bayes	75

Table 6: Cross validation score after parameter tuning

Decision Tree with Gini and Logistic Regression models showed significant improvement in cross validation score after the parameter tuning while Decision Tree with Entropy and Naïve Bayes performance dropped by 5% and 2% respectively.

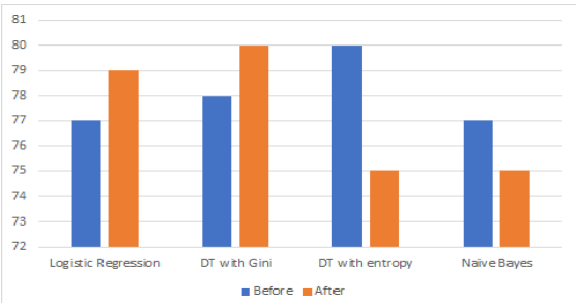

Fig 11: Model performance before and after parameter tuning

7 Potential Factors

7.1 Association

In order to find out the potential factors causing the feeling, we conducted Association rule mining to find possible association rules and patterns. We used a support value of 0.5 and a confidence of 0.9. The algorithm resulted in four association rules:

- ***If a student is graduate, then his workload is moderate.***
- ***If a student never visits home, then he is a graduate.***
- ***If a student is an extrovert and a graduate, then his workload is moderate.***
- ***If a student is a graduate and never visits home, then his workload is moderate.***
- ***If a student never visits home and his workload is moderate, then he is a graduate.***

The above rules show redundancy and can be shrunk into two distinct rules:

- ***If a student is graduate his workload is moderate.***
- ***If a student is a graduate with a moderate workload, he never visits home.***

Although the above two rules show that here might exist a relationship among the educational level, course load and feeling of homesickness, we were not able to state anything conclusive because the data was limited and heavily skewed, 78% of the international students were graduates.

7.2 Clustering

We conducted a simple KMeans clustering on the encoded values of the variables of the dataset. We did not encode the class variables, but had it converted to numerical value, same as in the Table 1. The results of clustering showed results similar to our previous exploratory analysis of some of the variables.

In the category of Educational_level (Fig 12), there was a clear distinction of two clusters on the extreme values of x-axis (Categories of program: Grad, Undergrad, PhD) spread across all values of y-axis (Homesickness).

Similarly, course_difficulty also clustered throughout the values of Homesickness but the clusters remined predominantly at the tails of the x axis.

The category of Friends, however, did not give a proper clustering results, owing to the improper distribution of data and classes. While clustering gives better results in unsupervised learning with numerical variables, our dataset was labelled data with categorical variables. Although the encoding gave an approximate visualization of the clusters, the plots during EDA with raw data gave a much better understanding of the clusters.

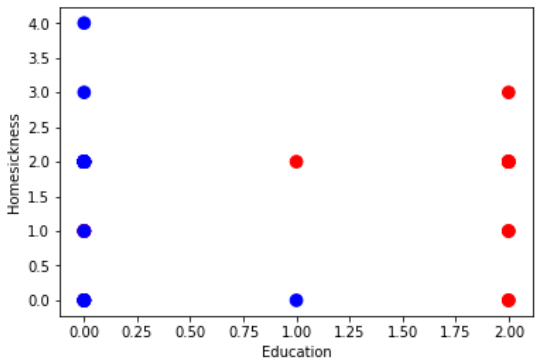

Fig 12: Clustering of Educational_level

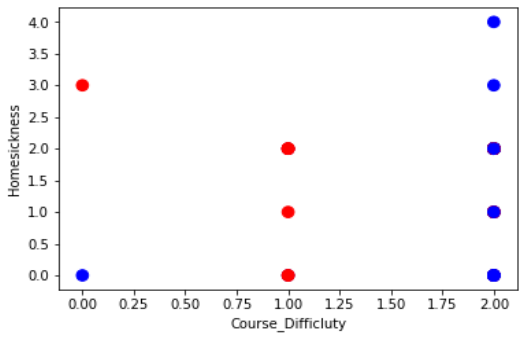

Fig 13: Clustering of Course_difficulty

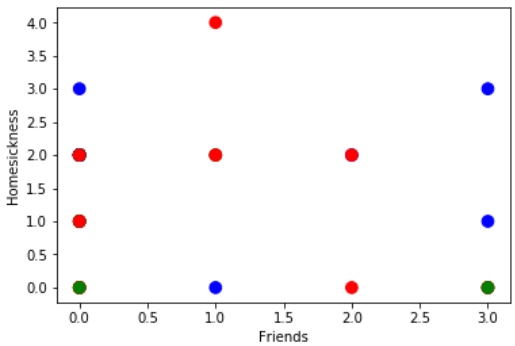

Fig 14: Clustering of Friends

8 Results and Discussions

The ablation study showed that out of the four models, only Logistic Regression and Decision Tree with Gini showed an improvement of 2% in cross validation score. Between the two, the Decision Tree with Gini gave a higher accuracy of 80%. Although the Decision tree with Entropy gave a better accuracy without parameters, it tends to yield an overfitting model without maximum depth and splits specified. Hence, we conclude that, for the collected data, the Decision Tree with Gini criterion predicted the classes of Homesickness better than the rest of the classifiers. By running an association mining, we concluded that Homesickness levels were very low for students who never visits home or have moderate workload. In addition, the data showed that the more extroverted the student is, the more is the feeling of homesickness. The causation may or may not hold, but there is a correlation between the variables. Finally, students who are occupied with a heavy course load focus more on their academics, there by showing a little or no feeling of homesickness.

9 Future Work

This entire study was made to show the possible methods and techniques to predict homesickness in international students. We received results and factors associated with homesickness, however the accuracy and values published are only with respect to the limited dataset we were able to collect. The scope of this project could go far and beyond and fetch more accurate results with much stronger models for larger datasets. One possible extension of this work that can be considered for future is to understand the stand of University in dealing with student's psychological issues better by including a greater number of environmental variables.

10 References

[1]. "Determinants of Homesickness Chronicity: Coping and Personality." *Personality and Individual Differences*, Pergamon, 27 Dec. 1999, https://www.sciencedirect.com/science/article/pii/S0191886998002621

[2]. Duru, Erdinc, and Senel Poyrazli. "Perceived Discrimination, Social Connectedness, and Other Predictors of Adjustment Difficulties among Turkish International Students - Duru - 2011 - International Journal of Psychology - Wiley

Online Library." *International Journal of Psychology*, John Wiley & Sons, Ltd, 25 July 2011, https://onlinelibrary.wiley.com/doi/full/10.1080/00207594.2011.585158

[3]. Harrison, et al. "The Impact of Cultural Intelligence and Psychological Hardiness on Homesickness among Study Abroad Students." *Frontiers: The Interdisciplinary Journal of Study Abroad*, Frontiers Journal. Dickinson College P.O. Box 1773, Carlisle, PA 17013. Tel: 717-254-8858; Fax: 717-245-1677; Web Site: https://eric.ed.gov/?id=EJ991042

[4]. Tinto, Vincent. "Dropout from Higher Education: A Theoretical Synthesis of Recent Research - Vincent Tinto, 1975." *SAGE Journals*, https://journals.sagepub.com/doi/10.3102/00346543045001089

[5]. Thomas, Darrin. *Cellphone Addiction and Academic Stress Among University Students in Thailand*. Oct. 2016, scholar.google.com/scholar_lookup?title=Cellphone addiction and academic stress among university students in Thailand&publication_year=2016&author=D. Thomas. http://journals.aiias.edu/iforum/article/view/187/191

[6]. "The Moderating Role of Gender Inequality and Age among Emotional Intelligence, Homesickness and Development of Mood Swings in University Students." *International Journal of Human Rights in Healthcare*, https://www.emerald.com/insight/content/doi/10.1108/IJHRH-11-2017-0071/full/html

[7]. Tochkov, Karin, et al. "Variation in the prediction of cross-cultural adjustment by Asian-Indian students in the United States." College Student Journal, vol. 44, no. 3, 2010, p. 677+. Gale Academic Onefile, Accessed 27 July 2019. https://go.galegroup.com/ps/anonymous?id=GALE%7CA238474689&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=01463934&p=AONE&sw=w

[8]. Wittrup, Audrey, and Noelle Hurd. "Extracurricular Involvement, Homesickness, and Depressive Symptoms Among Underrepresented College Students - Audrey Wittrup, Noelle Hurd." *SAGE Journals*, https://journals.sagepub.com/doi/abs/10.1177/2167696819847333