Exploring the Role of Supervised Learning in Predicting the Happiness Index of Elite College Students

Aashrith Chejerla

2023

Word Count: 4506

Introduction

The past decade of technological advancement has been primarily dominated by the introduction and development of new data analysis tools, such as machine learning. The surge in machine learning has fostered an incredible amount of change in the world due to its large versatility. By utilizing algorithms to analyze large datasets, machine learning is capable of identifying relationships and patterns normally invisible to traditional data analysis methods, plagued with issues such as human bias, enabling it to become a useful tool for making future predictions (Tirmizi, 2023). Its versatility and ability to be applied to extremely nuanced or complicated situations has led to its implementation in many different sectors. Of the many applications, education has become one of the most prominent areas in which predictive analytics through machine learning has begun to be employed. Over the past few decades, "Educational Data Mining" has already enabled researchers to gain quick, valuable insight into students through new algorithms and models tailored to specific student bodies, creating further insight that enables schools and administrators to tailor their services to their students' needs (Yagci, 2022). From predicting student performance in lower-income schools to evaluating the mental health of specific student populations, machine learning has revolutionized the conventional approach to "Educational Data Mining". While predicting aspects like student performance are key to maximizing performance, the mental health of students, specifically in environments that place large emotional demands on students, have always been understudied and underprioritized (Billings, 2020). Harnessing machine learning's ability to identify what demographics or specific types of students are at risk of lower happiness levels is crucial for schools to be able to create healthier environments for students that fosters even more learning. Additionally, it can set major precedents for the future of machine learning as an educational data

mining tool and promote the further research of nuanced student populations. Therefore, the primary goal of this research paper is to evaluate the extent to which machine learning can be used to predict self-reported happiness levels in a specific, nuanced educational population—undergraduate students attending T20 universities.

Literature Review

Machine Learning to predict Happiness Index

Happiness is a generally vague concept that has been inherently difficult to quantify or standardize. As a result, The Happiness Alliance aimed to quantify the concept of happiness and create a standardized measurement by splitting it into multiple domains through the development of the official Happiness Index. Researchers at Arizona State University justified the methodology behind the development of the Happiness Index through a series of studies in different regions, both geographically and demographically diverse, that attempted to find the major factors that would contribute to one's happiness. They successfully created a measurement tool for researchers throughout the world that measured happiness as a combination of satisfaction, the feeling of happiness, health, psychological status, time balance, social support, communities, culture, education, environment, materials, governance, and work (Musikanski et al. 2017). These "10 Domains of Happiness" are defined as follows: psychological well-being refers to one's personal sense of accomplishment and purpose; health refers to one's ability to perform daily tasks with adequate effort; time balance refers to one's free-time, stress, and enjoyment; community refers to one's sense of belonging and safety; social support refers to one's satisfaction with their community and feelings of love and loneliness; education, arts, and culture refers to one's experience with cultural diversity; environment refers to one's ability to

experience nature and clean environments; governance refers to one's confidence in their governing body; material well-being refers to one's possessions, needs, and financials; and work refers to one's independence, compensation and productivity (Musikanski et al. 2017).

The concept of using machine learning to predict the happiness index data of individuals was first introduced by Lexin You, who built a general machine learning model that considered individual socio-economic factors and demographics when predicting (2021). You's research encompassed roughly two to three thousand U.S. participants and surveyed them for information regarding their educational, relationships, age, and income and utilized a random forest algorithm to classify individuals in a binary fashion: "happy" or "not happy". You's model was able to achieve a 77.4% classification accuracy rate and identified income as the most prominent factor, proving that machine learning was capable of predicting happiness to a relatively large extent. Additionally, You set the precedent of utilizing random forest algorithms, a popular machine learning model that employs multiple decision trees simultaneously, when predicting happiness (IBM, 2024). The factors of happiness that You chose to examine was mainly contextualized by researcher Nam, with the Korean Academy of Nursing Administration,who contextualized many modern ML studies surrounding happiness indexes, specifically You's by identifying the primary contributors to one's happiness by examining high-stress careers, like nursing (Nam et. al, 2013). The significance of You's research, although they didn't directly implement the Happiness Alliance's standardized Happiness Index and chose to implement their own version, lies in the precedents it set by establishing that machine learning was mathematically viable as a general predictor of happiness.

This precedent that You established would be utilized in studies like Maria Fernanda Duron-Ramos's, a doctorate in social sciences, examination of 266 university students in

Mexico, which produced models with a 76.7% classification accuracy, utilizing a random forest algorithm, by focusing on how certain academic activities engaged students' happiness orientations (Duron-Ramos et. al, 2018). The main significance of this study was that it was one of the first, any only, studies that applied the precedent established by You in the specific context of students. Additionally, it carried the precedent of using random forest algorithms when predicting happiness. Similar to You, Duron-Ramos didn't choose to utilize the standardized Happiness Index, possible creating regional bias.

The final major key study concerning the happiness index of students was conducted in 2023 by Venkata Sailaja, assistant professor at the Vignana Jyothi Institute of Engineering and Technology. Because Sailaja's study was focused on impoverished schools in India, his model used features that were centered around the conditions of the school itself, including cleanliness, boards, environment, etc. Although Sailaja's study was heavily focused on specific schools and was, therefore, not generalizable to other countries or student populations, his study still set major precedents and established clearer frameworks for predicting the happiness of students through his analysis of specific populations (Sailaja, 2023).

Elite College Students

When analyzing the happiness index of different populations, specifically elite university students, one must be aware of the unique circumstances that influence their happiness indexes. This study defines "elite university students" as any student attending one of the top twenty national universities in the U.S. News's university ranking system (U.S. News, 2023). Students at these universities experience a unique environment that facilitates unique factors of happiness, specifically regarding their mental health.

Researcher Song, whose results align with the literature's consensus surrounding the topic conducted a thorough analysis of the costs and benefits of attending such universities and concluded that there was a substantial toll on mental health as stress levels were high and self-esteem levels were low (Song, 2017). This converses with PhD Candidate Katie Billings's research that examines how mental illnesses are stigmatized at these elite universities, creating an environment that encourages students to avoid treatment, further perpetuating the mental toll (Billings, 2020). Together, these two sources exemplify the unique conditions that surround elite university students and their mental health.

Another significant contributor to the lower happiness index of elite college students is the prevalence of imposter syndrome, defined as the ideological framework that one is incompetent (McLean, 2023). Many elite students view their admission into these top universities as a mistake and not a proper indicator of their success, perpetuating the idea that they're not accomplished and don't deserve their success (Burroughs, 2019). Additionally, Bravata finds that imposter syndrome is the key condition that contributes to professional/academic burnout and has no official treatment (Bravata et al., 2019). Bravata also finds that conditions like academic burnout have implications for one's perception of their own accomplishments.

Gap - Unique Factors of Happiness

Much of the modern literature surrounding machine learning as a predictor of one's happiness index quantifies happiness through general factors and implements the models on general populations. This results in a research gap as the literature fails to account for the nuanced factors of happiness in certain subpopulations and fails to utilize machine learning's

unique capability to examine nuanced situations, as defined earlier. The major studies, done by You and Duron-Ramos, examine general populations. Although Duron-Ramos examines a specific college, they're not trying to account for any specific unique factors and, therefore, their sample population is simply a microcosm of the general Mexican student population.

This student chooses elite college students as its target population must push themselves to their academic limits in environments that perpetuate poor mental health (Song, 2017). Being able to utilize modern technologies, specifical machine learning models, to not only predict which students are at risk of having lower happiness levels at these colleges, but also what factors are the primary cause of the happiness decline would enable educators and policymakers to construct healthier school environments. This project aims to prove the exclusive capabilities of machine learning technologies in creating predictions for nuanced populations, specifically in the context of schooling, in order to set precedents for the examination of other nuanced populations, such as marginalized populations, special education students, and more.

Methods

Data Collection

The machine learning model I have developed needs data on elite college students and their unique circumstances. Since such data is often sparse, sensitive, and difficult to access, a survey must be conducted.

The questionnaire I created focused on asking the college students a specific set of questions that would produce information that could be reasonably used to measure one's happiness index. Therefore, this study will utilize the Happiness Alliances' official Happiness Index Survey. The survey that was produced as a result of their research study contained a series

of questions for each of the 10 domains of happiness. The questions would either ask the user to agree with a statement related to the domain in the manner of "1-10" or "Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree". Adopting this style of assessment for my own questionnaire will enable me to quantify all the answers the user puts, further enabling me to input the values into the machine learning model as they require quantitative data to train with. Additionally, in order to create a more concise dataset, I chose to significantly reduce the amount of questions being asked. In order to still preserve the validity of the test, I had to ask the participant every question in the respective sections but average out their answers to produce a singular, weighted value for each section. Although, I had to remove certain work-related questions because it is not safe to assume that all college students have reliable jobs or internships that reflect their accomplishments. This generated exactly 10 feature variables for each participant. The final data point I collected through my survey was the participant's self-reported happiness, answered on a scale of 1 to 10. Having the user self-report their happiness enabled me to utilize a supervised learning approach.

Supervised Learning to Fill Research Gap

The premise of this project will rely on the concept of supervised learning, a subcategory of machine learning that relies on a pre-existing, labeled data set for the model to learn from to make predictions for future data points (IBM, 2023). The data I collected on each student was split into feature variables, Happiness Index Survey, and outputs, self-reported happiness. The model was fed information that will attempt to isolate a correlation between the variables and the happiness, enabling it to make future predictions on students given only their feature variables.

Artificial Intelligence

The next step was to use all this data to train the machine learning model. The model was built with the Scikit-Learn machine learning python library. In programming, libraries are existing pieces of code that are published in an "open-source" manner for anyone to implement in their own projects. Coding a machine learning model raw without any libraries is a very difficult and unreliable process. Additionally, the project will be coded in Google Colab, which implements Jupyter Notebooks. This enables programmers to isolate certain "blocks" in their code and run them independently from the others, further enabling them to isolate errors and try different algorithms with ease. In the context of machine learning, this allows users to test different machine learning algorithms independent of each other. While this project's machine learning models could have been built from scratch, Scikit-Learn's libraries coupled with Google Colab's interface provides a free, reliable framework, backed by a professional team of artificial intelligence engineers and presented in a very digestible, "low-code" manner, that can make this project easy to replicate and share. Additionally, most of the optimization of the chosen algorithm was handled by the platform itself.

Once the data had already been fed into the model, multiple machine learning algorithms were tested. The leading options to implement were a linear regression algorithm, random forest algorithm, and a multi-layered perceptron algorithm. The linear regression algorithm and the multi-layered perceptron algorithm were chosen because they are standard for the type of data I am using and are widely accepted as some of the most versatile options. I chose to use the random forest algorithm because much of the surrounding literature already set it as a precedent (You, 2021, Duron-Ramos et al., 2018).

Typical data sets are expected to be at least 10 times as large as the number of feature variables for every data point in order to ensure reliable training (Smolic, 2024). Therefore, my 10 feature variables, the domains of happiness, will require 100 students.

Most data sets are split into two subsets: a training set, usually comprising 70%, and a testing set, usually the remaining 30%. This enables the model to learn from the training set and determine its accuracy by using the testing set. An easy way of thinking about this would be through viewing the larger portion of the dataset as study material and viewing the smaller portion of the dataset as a self-administered quiz. This is crucial to ensure that models are not overfitting, the act of an algorithm tailoring its learning to the data it trained on and being unable to generalize results to any new data points. Unfortunately, this study's data set is already too small to split into sufficient subsets and, therefore, will require K-fold cross-validation. This technique splits the data into a certain number of folds, training itself on all but one and using the final one to test it (Lin et al, 2022). It then repeats this time for every permutation of training and testing folds and declares the overall accuracy as the average of all the accuracies of the tests it runs. This enables the model to sufficiently use all of the data for both training and testing without overfitting, the concept of a model becoming overly tailored to the dataset and unable to generalize (IBM, 2024).

Results

Data Collection and Cleaning

The survey received a total of 161 responses, all from students from T20 Universities. In order to collect the responses, the survey was posted on different social media platforms, specifically Reddit. It was posted in every subreddit, smaller, stratified digital communities within the larger Reddit community, dedicated to T20 Universities. Unfortunately, certain

subreddits removed the survey due to certain community guidelines. Additionally, a few of the larger colleges, such as Stanford and Harvard, had significantly larger digital communities and, therefore, disproportionately more activity on the survey, creating a possible bias in the data set.

Additionally, one of the participants from MIT sent the survey out to all of MIT's undergraduate student population through their email network. Despite the fact that this generated a large amount of useful data points, it created a heavy bias in the data set towards MIT undergraduate students as MIT students now comprised at least 27% of the survey population. It was impossible to analyze what other schools held large portions of the survey population and if there were any schools that were not represented because the survey avoided asking for too much personal information and many students chose to use their personal email address, instead of their university email address.

As a result of the two potential biases created, it becomes important to be cautious when generalizing the results of this study to larger populations as the sample is not representative. Future research, with more resources, should aim to create a more balanced data set. That being said, 161 responses is more than sufficient to conduct a rudimentary machine learning analysis.

Before conducting any machine learning analysis, the data had to be cleaned. Since the data was collected by a survey filled with multiple choice problems, there were no numerical errors present as all inputs were forced to fit the necessary parameters. In order to condense the raw data, as mentioned previously, all questions in every domain of happiness were averaged out into one value, creating 10 parameter values and one output value per data point. The data was then randomly split into a training set and a test set, consisting of 70 percent and 30 percent of the total data set, respectively.

Model Building

To build the model, the Scikit-Learn (Sklearn), the popular python library that specializes in machine learning, library was imported into my workspace. The existing frameworks were called and implemented with the data I collected. I utilized Sklearn's Linear Regression, Random Forest Regressor, and MLPRegressor algorithms.

Linear regression is a foundational algorithm in supervised learning for predictive modeling tasks. It aims to find the best-fitting linear relationship between the feature variables and the target variable. The relationship can be represented with the equation: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots \beta_{10} x_{10} + \varepsilon$. Y is the target variable, β_0 is the constant, x_1 to x_{10} are the feature variables, β_1 to β_{10} are specific weights assigned to each variable, and ε represents the variable due to chance. The typical training process for a linear regression model aims to fine tune the coefficients, or weights (β_1 to β_{10}), to minimize the difference between the predicted and actual values in the training data.

The other algorithm that was tested was a random forest regression model. It's a very versatile algorithm that allows us to utilize the concept of decision trees. It functions by creating many decision trees while training itself and combining their outputs to create an accurate model for regression analysis. It is a form of ensemble learning, a machine learning technique that utilizes many different models at once and combines their results. Each decision tree in a random forest algorithm is created independently with a randomly selected subset of the training data and a randomly selected subset of the feature variables. The algorithm then begins placing a bias on certain trees that have produced better values and aggregates all of the results. This method is quite effective for handling complex relationships between many variables, which is suitable for our survey involving many fluid factors.

The final algorithm that was used was a multi-layered perceptron (MLP) regression model, also known as a “neural network”. Its structure consists of multiple layers of neurons, including an input layer, a few hidden layers, and an output layer. Each neuron in the network is connected to every neuron in the adjacent layers, and each connection is associated with a weight that determines its strength. The neuron itself contains its own “activation function”, which can be treated as a weight applied to the numbers it receives. The input variables are put into the input layer and processed through all the activation functions and weights. During training, the regressor learns to approximate complex nonlinear relationships between input features and target values by adjusting the weights of its connections through a process called backpropagation. This involves iteratively propagating errors backward from the output layer to the hidden layers, simultaneously updating the weights to minimize the loss function. Over the course of many iterations, or epochs, the neural network optimizes all the weights and activation functions to predict accurate output values based on the specific input parameters.

Model Evaluation

Once the model is trained, its ability to make predictions on new data is evaluated through various metrics, specifically a loss function. A loss function is used by machine learning models to determine the total amount of value that is “lost” by the predictor in order to measure its accuracy. When these models “optimize” loss functions, they aim to minimize the amount of loss that was measured. The chosen metrics for these models are a mean squared error (MSE) loss function and a R-squared analysis on a validation dataset.

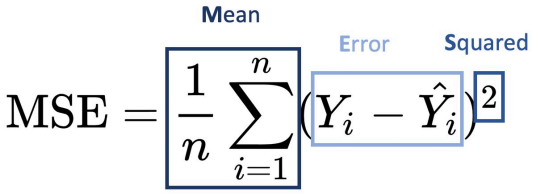

Figure 1.1

As shown in Figure 1.1, MSE used to evaluate the performance of regression models through analyzing the error. In the context of our study, Y_i represents the actual value of a data point in our test set while $\hat{Y}_i$ represents our models' predicted value. The difference between them is squared, to ensure positivity, and the mean of all these values is taken, denoted by the summation notation. The closer the MSE is to 0, the more accurate the model is and more effective it is at making predictions on new data.

R-squared is a value that is very valuable for understanding the overall fit of a certain model. It will take the independent variable, our real value, and the dependent variable, our predicted value, and measure the variance in relationship from the line of best fit. Although it doesn't necessarily indicate whether the model's predictions are unbiased or reliable, it does provide valuable information as to the variance of the data.

Using these two metrics on both the training and test data, we will be able to gauge machine learning's ability to predict the Happiness Index of T20 University students.

Model Performance

After both models were trained and evaluated, their performance was compared in the following chart.

Method	Training MSE	Training R^2	Test MSE	Test R^2
Linear Regression	1.813028	0.429099	1.342804	0.642001
Random forest	1.554617	0.51047	1.257606	0.664716
MLP	3.183484	0.002441	1.966309	0.475772

Figure 1.2

All algorithms, besides MLP, produced relatively similar Mean Squared Errors and variance. One key observation is that the linear regression performed better on the training data, as seen with the lower MSE and R^2, while performing worse on the Test. Although this may seem counterintuitive, this can be a possible sign of overfitting. Overfitting refers to when a machine learning model tailors itself too well to the training data and becomes unable to generalize to new data sets because it becomes hyper focused on the unique nuances of its training data. As a result, it will perform well on the training data but when given new data, with new characteristics, it will perform noticeably worse, as seen with our random forest algorithm. An analogous situation would be an athlete who develops the specific skills needed to perform well in their desired sport without developing the necessary athleticism to have transferable skills to other sports or movements. Our linear regression algorithm may be overfitting as a result of the model being overly complex, relative to the primitive data set.

While the linear regression model and the random forest algorithm both had relatively accurate models with variance above 0.6, the MLP had a significantly worse variance around 0.47 and a training MSE that was significantly lower. The abnormally low accuracy on the MSE and variance of the training data tells us that the model was not well trained and simply got the .47 variance accuracy due to chance and the smaller size of the testing set. MLPs are far more efficient when training on larger data sets. The smaller sample size of this study may have created inefficient conditions for the use of an MLP regression model. Additionally, better results

could have potentially been produced by the model if the number of epochs, or iterations of training, was optimized.

Regardless, the R^2 value of the linear regression model on the test set being greater than 0.5 indicates that the model is able to predict the happiness index of T20 University students to some degree. While not fully effective, it indicates that there is some opportunity for machine learning to be utilized in evaluating elite university students and their mental health.

Conclusion

The purpose of this study was to determine the feasibility of machine learning in predicting the happiness, as defined by the Happiness Alliance's Happiness Index Survey, of elite university students, as defined as undergraduate students in the top 20 ranked national universities. These students often experience unique work environments, lifestyles, and factors that make typical measures of happiness inefficient. By training a machine learning model to learn the nuances behind such students, this study hoped to create a more reliable tool that could be utilized by institutions, governments, and students, alike, in fostering healthier educational environments that create growth without harming a student's mental health.

The linear regression algorithm, which proved slightly more effective than the random forest algorithm due to the random forest algorithm experiencing overfitting with the training data set, produced a R^2 value that was greater than 0.5, indicating marginal levels of success. Although this number may not seem that high, it is crucial to contextualize it with the limitations of this study.

The first major limitation was the dataset. It is impossible to get a perfectly balanced dataset of all T20 university students as they must volunteer to fill out a form. Without adequate incentive, outreach, and resources, the only dataset that could be produced was an incredibly

small, uneven dataset. Although the size is technically large enough to perform rudimentary machine learning analysis, it is not nearly as large as most machine learning datasets as thousands and thousands of data points would be required. Additionally, the dataset unevenly representing some of the universities creates a representative bias that makes it difficult to generalize the results to the larger population.

With the limitations taken into account, the results of this study lay critical groundwork and set major precedents for future studies with more resources that can produce larger datasets and implement more complicated models. This study hopes to incentivize more research on the mental health of T20 university students by bringing awareness to not only the issue at hand, but the capability of modern technology to address it. Primarily, this study highlights the potential of machine learning in understanding nuanced populations that are not fully understood. Other nuanced populations than elite university students, such as special education students, marginalized populations, immigrants, international students, and more, in education aren't fully understood by researchers, administrators, and parents, alike. Using machine learning to understand these unique populations and the factors that can contribute to their success when typical, standards metrics fail can create a new word of educational diversity and opportunity.

References

US News (2024). 2024 Best National Universities | US news rankings. https://www.usnews.com/best-colleges/rankings/national-universities

Azure machine learning - ML as a service: Microsoft Azure. ML as a Service | Microsoft Azure. (2024). https://azure.microsoft.com/en-us/products/machine-learning

Billings, K. R. (2020). Stigma in class: Mental illness, social status, and tokenism in Elite College Culture. *Sociological Perspectives*, *64*(2), 238–257. https://doi.org/10.1177/0731121420921878

Bravata, D. M., Watts, S. A., Keefer, A. L., Madhusudhan, D. K., Taylor, K. T., Clark, D. M., Nelson, R. S., Cokley, K. O., & Hagg, H. K. (2019). Prevalence, predictors, and treatment of Impostor Syndrome: A systematic review. Journal of General Internal Medicine, 35(4), 1252–1275. https://doi.org/10.1007/s11606-019-05364-1

Burroughs, J. (2019). Imposter Syndrome at an Elite University. Students Union UCL. https://studentsunionucl.org/articles/imposter-syndrome-at-elite-univeristy

Durón-Ramos, M. F., & Vázquez, F. G. (2018). Orientation to happiness as a predictor of university students' engagement. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(4), 294. https://doi.org/10.11591/ijere.v7i4.15446

A guide to impostor syndrome-and overcoming it. Understanding and Overcoming Impostor Syndrome | McLean Hospital. (2023). https://www.mcleanhospital.org/essential/impostor-syndrome

Lin, Z., Lai, J., Chen, X., Cao, L., & Wang, J. (2022). Curriculum reinforcement learning based on K-fold cross validation. *Entropy*, *24*(12), 1787. https://doi.org/10.3390/e24121787

Musikanski, L., Cloutier, S., Bejarano, E., Briggs, D., Colbert, J., Strasser, G., & Russell, S. (2017). Happiness index methodology. *Journal of Social Change*, *9*(1). https://doi.org/10.5590/josc.2017.09.1.02

Nam, M. H., & Kwon, Y. C. (2013). Factors influencing happiness index of hospital nurses. *Journal of Korean Academy of Nursing Administration*, *19*(3), 329. https://doi.org/10.11111/jkana.2013.19.3.329

Sailaja, N. V., Reddy, K. L., Aditya, G., Shashank, B., & Sai, V. H. (2023). Happiness index prediction of students using machine learning. *Atlantis Highlights in Computer Sciences*, 85–96. https://doi.org/10.2991/978-94-6463-314-6_9

Singh, A. (2023, N). A comprehensive guide to ensemble learning (with python codes). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models

Song, H. (2017). The cost of attending an Elite College. *Economics Letters*, *159*, 173–176. https://doi.org/10.1016/j.econlet.2017.07.029

Tirmizi, A. M. (2023). *Machine learning vs. predictive analytics*. Dataversity. https://www.dataversity.net/machine-learning-vs-predictive-analytics/

What is overfitting?. IBM. (2024). https://www.ibm.com/topics/overfitting

What is Random Forest?. IBM. (2024). https://www.ibm.com/topics/random-forest

What is supervised learning?. IBM. (2024). https://www.ibm.com/topics/supervised-learning

Yağcı, M. (2022). Educational Data Mining: Prediction of Students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1). https://doi.org/10.1186/s40561-022-00192-z

You, L. (2021). Utilizing machine learning to predict happiness index. 2021 2nd International Conference on E-Commerce and Internet Technology (ECIT). https://doi.org/10.1109/ecit52743.2021.00058