

Introduction

In this project, we will try to predict the University undergraduate admission for students in Bangladesh, based on various academic, social, and familial factors we will take advantage of the use of machine learning algorithms in our dataset, thanks to identifying influential factors in the acceptance process. The motivation for this study is to address the competitive nature of university admissions nowadays.

**Classification Vs Regression**

We will use classification instead of regression because the goal is to predict a discrete outcome. We will study whether or not a student will be admitted to the university. So, Classification fits more in these types of problems where our target variable (admission) is categorical or binary (1 for "admitted", 0 for "not admitted").

**Problem Statement**

The problem we are trying to solve is a binary classification task because we need to predict whether a student will be admitted to a university or not (1 for admitted, 0 for not admitted) based on various features.

**Objective**

The objective is to find and apply some classification models, for example, Decision Trees, Random Forests, and others to compare their performance.

Bearing in mind that, reliably determines admission probability, and we will do our best to pick the right algorithm to do it, justifying the use of machine learning in educational forecasting.

Data Preparation

I took this dataset from Kaggle containing multiple features relevant to this admission prediction, including:

* SSC (Secondary School Certificate) GPA.
* HSC (Higher Secondary Certificate) GPA.
* Family\_Economy, Social\_Media\_Engagement, Family\_Education, Residence along with other factors.

Data Characterization:

The data also includes categorical variables representing various binary options, actually mostly all of them, (e.g., 1 for "Yes," 0 for "No") for features like Relationship and Residence\_with\_Family as an example

After an initial examination of data.describe(), features with significant spread, like SSC\_GPA and HSC\_GPA, were selected for analysis due to their apparent impact on the target variable (University admission).

Data pre-processing:

I split this data into 70% training and 30% testing following best practices given in all classes.

After that, we examined shapes.

5 Fold Cross-validation was utilized in this study, in simple terms a fold represents a set of data, 5 folds are 5 sets of data which then we run the ML algorithms on top of. In this case, I decided to split it in 5 instead of 3 because; First of all our dataset is not small, 5 folds create a balance of bias-variance. Second, compared to the 3 folds, 5 folds reduces the event of overfitting.

Modeling Approach and Choice Justification

Given the binary classification goal nature, classification models were chosen, Let’s see them:

1. **Decision Tree**: A simple model useful for handling non-linear data, sometimes has the problem of overfitting.
2. **Random Forest**: Average predictions across multiple decision trees, reducing overfitting.
3. **Support Vector Machine (SVM)**: A robust classifier effective in high-dimensional spaces.
4. **Tuned SVM**: We apply a GridSearchCV tuning to this high dimensional classifier to boost its performance, we will see the results after.

Random Forest was expected to perform well due to its nature. Why? it provides a balance between bias-variance. SVM, known for its accuracy in binary classification, I included to examine its generalization ability and, later on, modified as well with some parameter tuning.

Hyperparameter Tuning

Hyperparameters control model structure, function, and performance. Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success.

GridSearchCV or RandomSearchCV are used to find the optimal combination of hyperparameters. As mentioned, for this project, GridSearchCV was used, it allowed me to change between some hyperparameter grids in order to find the best combinations for the SVM model.

GridSearchCV techniques like optimized C and kernel type, was used to try to improve classification and reduce misclassifications.

Cross Validation

All models were evaluated using cross-validation. Then contrasted them one against the other, to see which one performed better.

The 5-fold cross-validation applied in this study was used to check if model's performance is consistent across different splits of the data.

This helps us in the generalization; Showing all model's performance on unseen data.

Results Interpretation and Model Evaluation

After training and evaluating each model, the results were:

* **Decision Tree**: Initial accuracy was 76.11%. This model demonstrated lower performance due to higher variance and overfitting, especially on smaller training sets.
* **Random Forest**: Initial accuracy was 80%. This shows the importance of solving the problem of overfitting in ML.
* **SVM**: Initial accuracy of 80%.
* **Tuned SVM**: Stayed the same as the regular SVM, 80%.

In the end, we will find out that Random Forest is the most reliable for this task, due to its high accuracy and the way of handling bias-variance trade-off more effectively. Visualizations (e.g.: ROC curves, and Confusion Matrices) were generated to show the model's performance. Furthermore, showing a strong predictive capability, the Random Forest ROC curve showed the AUC number closest to 1.

Discussion on Overfitting, Underfitting, and Generalization

High bias, Underfitting, seem to be a problem in this dataset, let me explain. All models performed better in the train set than in the test set, which is a sign of Underfitting.

Fixing overfitting means preventing the model from learning associations that are specific to the training set. There are two common ways to fix overfitting: modifying the training set or regularizing the model.

To address potential overfitting, techniques such as cross-validation and hyperparameter tuning were applied in this study.

Conclusion

The study successfully ran and compared machine learning models to predict undergraduate admissions in Bangladesh. Random Forest emerged as the optimal model due to its high accuracy, defeating any other ML in cross-validation as well.

At the very beginning, it was evident that SSC\_GPA, HSC\_GPA, and Family\_Economy were key predictors of admission outcomes.

Hyperparameter tuning with GridSearchCV was used to try a more fancy model, prevent overfitting, have good cross-validations, etc… in the end, it ended up being in third place as a model performer. This confirms the saying that sometimes a “Linear Regression can do the work better”

Cross-validation was the key to picking the right model because three out of four models performed the same.

Additionally, visualizations as ROC curves and Confusion Matrices, supported our findings by showing model performance across different metrics.

This project demonstrates that reliable prediction models can assist in decision-making.

References

Anwar Hossain EFAT (2024). Undergraduate Admission Test Survey in Bangladesh. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/ahefatresearch/undergraduate-admission-test-survey-in-bangladesh [Accessed 10 Oct. 2024].

AWS (n.d.). What is Hyperparameter Tuning? - Hyperparameter Tuning Methods Explained - AWS. [online] Amazon Web Services, Inc. Available at: <https://aws.amazon.com/what-is/hyperparameter-tuning/>.

Iqbal, M. (n.d.). Machine Learning HDip in Science for DAB/ AI CCT College Dublin.

Iqbal, M. (n.d.). Machine Learning HDip in Science for DAB/ AI CCT College Dublin Classification and Regression Week 2.

Linkedin Community, A. (2024). How can you use support vector machine algorithms to improve customer retention? [online] Linkedin.com. Available at: https://www.linkedin.com/advice/1/how-can-you-use-support-vector-machine-id7lc [Accessed 10 Oct. 2024].

Lui, D. and Angus, G. (2023). How to Guide: Overcoming overfitting in your ML models - Predibase. [online] Predibase.com. Available at: <https://predibase.com/blog/how-to-guide-overcoming-overfitting-in-your-ml-models>.

Mujtaba, H. (2020a). An Introduction to Grid Search CV | What is Grid Search. [online] GreatLearning. Available at: <https://www.mygreatlearning.com/blog/gridsearchcv/>.

Mujtaba, H. (2020b). What is Cross Validation in Machine learning? Types of Cross Validation. [online] GreatLearning Blog: Free Resources what Matters to shape your Career! Available at: <https://www.mygreatlearning.com/blog/cross-validation/>.

Scikit Learn (2012). 3.2. Tuning the hyper-parameters of an estimator — scikit-learn 0.22 documentation. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/grid_search.html>.

Simplilearn (2021). What is Dimensionality Reduction? Overview, and Popular Techniques. [online] Simplilearn.com. Available at: <https://www.simplilearn.com/what-is-dimensionality-reduction-article>.

Singh, V. (2023). Evaluation Metrics in Machine Learning - Shiksha Online. [online] Shiksha.com. Available at: <https://www.shiksha.com/online-courses/articles/evaluating-a-machine-learning-algorithm/>.

Terra, J. (2022). What is a ROC Curve, and How Do You Use It in Performance Modeling? | Simplilearn. [online] Simplilearn.com. Available at: <https://www.simplilearn.com/what-is-a-roc-curve-and-how-to-use-it-in-performance-modeling-article>.

Learn, S. (2024). RocCurveDisplay. [online] scikit-learn. Available at: https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.RocCurveDisplay.html [Accessed 5 Nov. 2024].

Bobbitt, Z. (2021). How to interpret a confusion matrix for a machine learning model. [online] www.evidentlyai.com. Available at: <https://www.evidentlyai.com/classification-metrics/confusion-matrix>.

Turing.com. (2022). Different Types of Cross-Validations in Machine Learning and Their Explanations. [online] Available at: https://www.turing.com/kb/different-types-of-cross-validations-in-machine-learning-and-their-explanations# [Accessed 7 Nov. 2024].

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An introduction to statistical learning : with applications in R. [online] Springer, pp.181, 183. Available at: <https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf>.

RudderStack. (n.d.). What is Generalization in Machine Learning? [online] Available at: https://www.rudderstack.com/learn/machine-learning/generalization-in-machine-learning/.