Predictive Modeling of Student Feedback Using Multi-Output Classification

Submitted by

Nathan Furtado

24251312

MSc. Big Data Analytics

AIMIT, St. Aloysius (Deemed to be University)

Mangalore, Karnataka

Submitted in Partial Fulfilment of the Requirements for the Award of the Degree of

Master of Big Data Analytics

Under the guidance of

Dr. Hemalatha N

Dean, School of IT

AIMIT, St. Aloysius (Deemed to be University),

Mangaluru-575 022.

Submitted to

ALOYSIUS INSTITUTE OF MANAGEMENT AND INFORMATION TECHNOLOGY

(AIMIT)

ST ALOYSIUS COLLEGE (DEEMED TO BE UNIVERSITY)

MANGALORE, KARNATAKA

2025

Abstract:

Project Overview

This project leverages supervised machine learning to analyse and predict multiple aspects of student feedback collected via a structured survey. The primary goal is to identify the factors that influence students' satisfaction, preference for learning modes, recommendation tendencies, and housing choices (hostel or not). A multi-output classification approach was employed to predict four interrelated binary target variables simultaneously using a variety of classification models.

---

1. Introduction

1.1 Background

This report analyses survey data collected from university students to gain insights into their academic preferences, lifestyle choices, stress factors, and satisfaction with university facilities. Understanding these aspects is crucial for university administrators aiming to enhance student engagement, academic success, and overall well-being.

The dataset includes responses from both undergraduate (UG) and postgraduate (PG) students, with variables covering demographics, study modes, extracurricular involvement, stress levels, and opinions on university offerings. The goal is to identify patterns and relationships that can inform policy and program development.

1.2 Problem Statement

Educational institutions seek to understand and improve student experiences. Feedback from students can inform strategic decisions about curriculum design, infrastructure development, and extracurricular offerings. However, student preferences and satisfaction levels are multi-dimensional and interdependent.

Challenge: Predict multiple student outcomes (recommendation, satisfaction, online class preference, and hostel stay) based on various demographic, academic, and behavioural features.

1.3 Objectives

- Perform data cleaning, preprocessing, and exploratory analysis on the student survey dataset.
- Build a multi-output classification model using machine learning algorithms.
- Compare the performance of multiple models including Logistic Regression, Multi-Layer Perceptron, Random Forest, and SVM (Linear & Polynomial).
- Determine key patterns that influence student satisfaction and engagement.
- Recommend the best-performing model based on predictive accuracy.

2. Material and Methods

2.1 Data Description

The dataset consists of responses from approximately 300 UG and PG students, with the following key variables:

- Demographics: Gender, postgraduate status, full-time student status, employment status.
- Living Arrangements: Hostel residence, daily commuting.
- Academic Preferences: Preference for online vs offline classes, manageability of coursework, preference for practical vs theoretical learning.
- Extracurricular Activities: Participation in clubs, sports involvement, leadership roles, opinion on extracurricular offerings.
- Stress and Study Habits: Exam stress, preference for studying alone or in groups, library usage.
- Future Plans: Plans for higher studies, gap semester history.
- Satisfaction and Recommendations: Satisfaction with university facilities, willingness to recommend the university.

The data is categorical, mostly binary (Yes/No), enabling frequency analysis and cross-tabulations to explore associations.

2.2 Data Preprocessing Steps

Null Check

- Checked for missing values (none were handled, assuming dataset is clean).
- #Code: df.isnull().sum()[4]

Encoding

- **Binary encoding** of Yes/No responses → 1/0.
- **One-hot encoding** of categorical variables like gender and degree (with drop_first=True to prevent dummy trap).[4]
- #Code: df.replace({'Yes': 1, 'No': 0}), pd.get_dummies(..., drop_first=True)[3]

Infer Object Types

- Ensures correct data types for modeling (e.g., int instead of object).
- #Code: df.infer_objects()[4]

Feature-Target Split

- #Code: X = df.drop(columns=target), y = df[target][4]

Train-Test Split

- 70% training / 30% testing split with fixed random state for reproducibility.
- #Code: train_test_split(..., test_size=0.3, random_state=42)[4]

Feature Scaling

- Scales features to zero mean and unit variance — essential for SVMs and MLP.
- #Code: StandardScaler()[4]

2.3 Tools & Libraries

Library	**Purpose**
pandas, numpy	Data loading, transformation, and numerical operations
seaborn, matplotlib.pyplot	Data visualization (feature skewness/imbalance)
sklearn.model_selection	train_test_split to split data into training/testing
sklearn.preprocessing	StandardScaler for feature standardization
sklearn.linear_model	LogisticRegression model
sklearn.neural_network	MLPClassifier (ANN)
sklearn.ensemble	RandomForestClassifier
sklearn.svm	SVC for SVM models (linear & polynomial)
sklearn.multioutput	MultiOutputClassifier for handling multi-target outputs
sklearn.metrics	Evaluation: accuracy_score, confusion_matrix, classification_report

Table 2.1 Libraries used in project

2.4 Machine Learning Models

All models are wrapped in MultiOutputClassifier[5] to allow **multi-label classification** (i.e., predicting multiple binary targets simultaneously).

2.4.1 Logistic Regression

LogisticRegression(max_iter=1000)[4]

- Linear model. Good for quick, interpretable baseline.

2.4.2 Multi-Layer Perceptron (ANN)

MLPClassifier(hidden_layer_sizes=[100, 100, 100], max_iter=1000)[3]

- Deep neural network with 3 hidden layers of 100 neurons each.
- Suited for complex patterns; scaling required.

2.4.3 Random Forest

RandomForestClassifier(n_estimators=100)[3]

- Ensemble of decision trees; robust to outliers and non-linearity.
- Handles multi-output natively, but here wrapped for consistency.

2.4.4 Support Vector Machine

SVC(kernel='linear' / 'poly', probability=True)[2]

- Linear kernel: fits straight decision boundaries.
- Polynomial kernel: captures curved/non-linear boundaries.
- Requires standardized input.

2.5 Evaluation Metrics Used

2.5.1 Confusion Matrix

confusion_matrix(y_test, y_pred)

- Shows counts of TP, TN, FP, FN — useful for error type analysis.

2.5.2 Classification Report

classification_report(y_test, y_pred)

- Includes:
    - **Precision**: How many predicted positives were correct.
    - **Recall**: How many actual positives were found.
    - **F1-score**: Balance between precision and recall.
    - **Support**: Actual occurrences in test set.

2.5.3 Accuracy Score

accuracy_score()

- For each target label individually.
- Average accuracy across all targets used to compare model performance.

2.6 Workflow Design (End-to-End Flow)

- Load Data
- Check for Nulls
- Encode Binary & Categorical Features
- Feature-Target Split
- Train-Test Split
- Feature Scaling
- Imbalance/Skewness Visualization[1]
- Train Multiple Models
- Evaluate with Confusion Matrix, Classification Report, Accuracy
- Compare & Rank Models by Average Accuracy

3. Results and Discussion

3.1 Model Performance

Model	Avg. Accuracy (%)	Strengths	Weaknesses
SVM (Poly)	**84.24**	Strong balanced performance across targets. Best at capturing complex decision boundaries.	May be computationally heavy; polynomial kernel can overfit if not tuned well.
Random Forest	**81.25**	Consistent and interpretable; handles feature interactions and non-linearity well.	Not best for extremely imbalanced classes (e.g., online class preference).
MLP Classifier	**79.89**	Excellent on binary, clean splits (e.g., recommendation). Strong learner with depth.	Slight overfitting tendency; performance drops on class-imbalanced outputs.
Logistic Regression	**78.80**	Very interpretable. Solid baseline. Good precision/recall trade-off.	Struggles with non-linear relationships.
SVM (Linear)	**77.99**	Fast and efficient; good generalization.	Suffers with non-linearly separable data (e.g., online preference class).

Table 3.1: Model Performances

3.2 Feature Importance

- Satisfaction and recommendation driven by academic quality, facilities, faculty interaction.
- Hostel stay prediction impacted by distance from home, transport, and cost.
- Online class preference related to tech savviness, course type, past experience.

3.3 Discussion Points

Model Bias vs Variance

- Logistic Regression underfits in complex relationships (e.g., online classes).
- SVM (Poly) is slightly prone to overfitting — check generalizability on unseen data (e.g., using cross-validation).
- MLP performs excellently on structured data but suffers in imbalanced targets (likely high variance).
- Random Forest gives great balance but not always best at minority class recall.

4. Conclusion

4.1 Conclusion

This project successfully implemented a **multi-output classification**[1] approach to predict multiple aspects of student experiences using survey data. Five classification models—Logistic Regression, MLP, Random Forest, SVM (linear), and SVM (polynomial)—were evaluated.

Among the models, the **Polynomial SVM** achieved the highest overall performance (**84.24% average accuracy**), closely followed by **Random Forest** and **MLP**. Each model demonstrated strengths with specific target variables. For example, **MLP** and **Random Forest** performed exceptionally well on predicting student satisfaction and recommendation likelihood, while more complex models like **SVM (poly)** better handled the non-linear nature of online class preference.

The results indicate that student feedback data can be leveraged effectively for predictive modeling, supporting informed decision-making for institutional planning and personalized student support systems.

4.2 Limitations

- Imbalanced Target Classes[1]
- Limited Features

- Subjective Responses
- Generalizability
- Model Interpretability

4.3Future Work

- Address Class Imbalance[1]
- Feature Engineering & Expansion
- Ensemble & Hybrid Models

5. References

- He, H., & Garcia, E. A. (2009). *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

- Huang, S. Y., Lin, C. H., & Cheng, W. T. (2012). *Predicting student academic performance with support vector machine*. International Journal of Information and Education Technology, 2(3), 143–146. https://doi.org/10.7763/IJIET.2012.V2.123

- LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

- Zhang, M. L., & Zhou, Z. H. (2007). *ML-KNN: A lazy learning approach to multi-label learning*. Pattern Recognition, 40(7), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

6. Appendix

6.1 Data sample

Are_you_an_Postgraduate_student	What_is_your_gender	Are_you_a_full-time_student	Do_you_live_in_a_hostel	Do_you_prefer_online_classes_over_offline_classes	Have_you_participated_in_any_university_clubs	Are_you_satisfied_with_the_university's_facilities	Do_you_commute_to_the_university_daily	Are_you_currently_employed	Do_you_find_your_coursework_manageable	Are_you_a_pursuing_a_UG_or_PG_Degree	Are_you_planning_to_pursue_higher_studies_after_graduation	Do_you_feel_stressed_about_exams	Have_you_ever_taken_a_gap_semester	Are_you_involved_in_sports_activities	Do_you_prefer_studying_alone_rather_than_in_groups	Do_you_often_use_the_university_library	Have_you_ever_taken_a_leadership_role_in_a_student_organization	Do_you_prefer_practical_learning_over_theoretical_learning	Do_you_think_the_university_should_offer_more_extracurricular_activities	Would_you_recommend_your_university_to_others
Yes	Male	Yes	Yes	No	No	Yes	Yes	No	Yes	PG	No	Yes	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Yes	Female	Yes	Yes	No	No	Yes	Yes	No	Yes	PG	Yes	No	No	No	Yes	No	Yes	Yes	Yes	Yes
Yes	Male	Yes	Yes	No	Yes	No	No	No	Yes	PG	No	No	No	Yes	No	No	Yes	Yes	Yes	No
Yes	Female	Yes	Yes	No	Yes	No	No	Yes	No	PG	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No

Figure 6.1: Data Sample

6.2 Code snippet

Load data

```
df=pd.read_csv("ML Assignment.csv")
```

Encoding

```
df = df.replace({'Yes': 1, 'No': 0})

df = pd.get_dummies(df, columns=['What_is_your_gender',
'Are_you_a_pursuing_a_UG_or_PG_Degree'], drop_first=True)

df = df.infer_objects(copy=False)
```

Multiple Output Classifier

```
MultiOutputClassifier()
```

Additional figures and tables

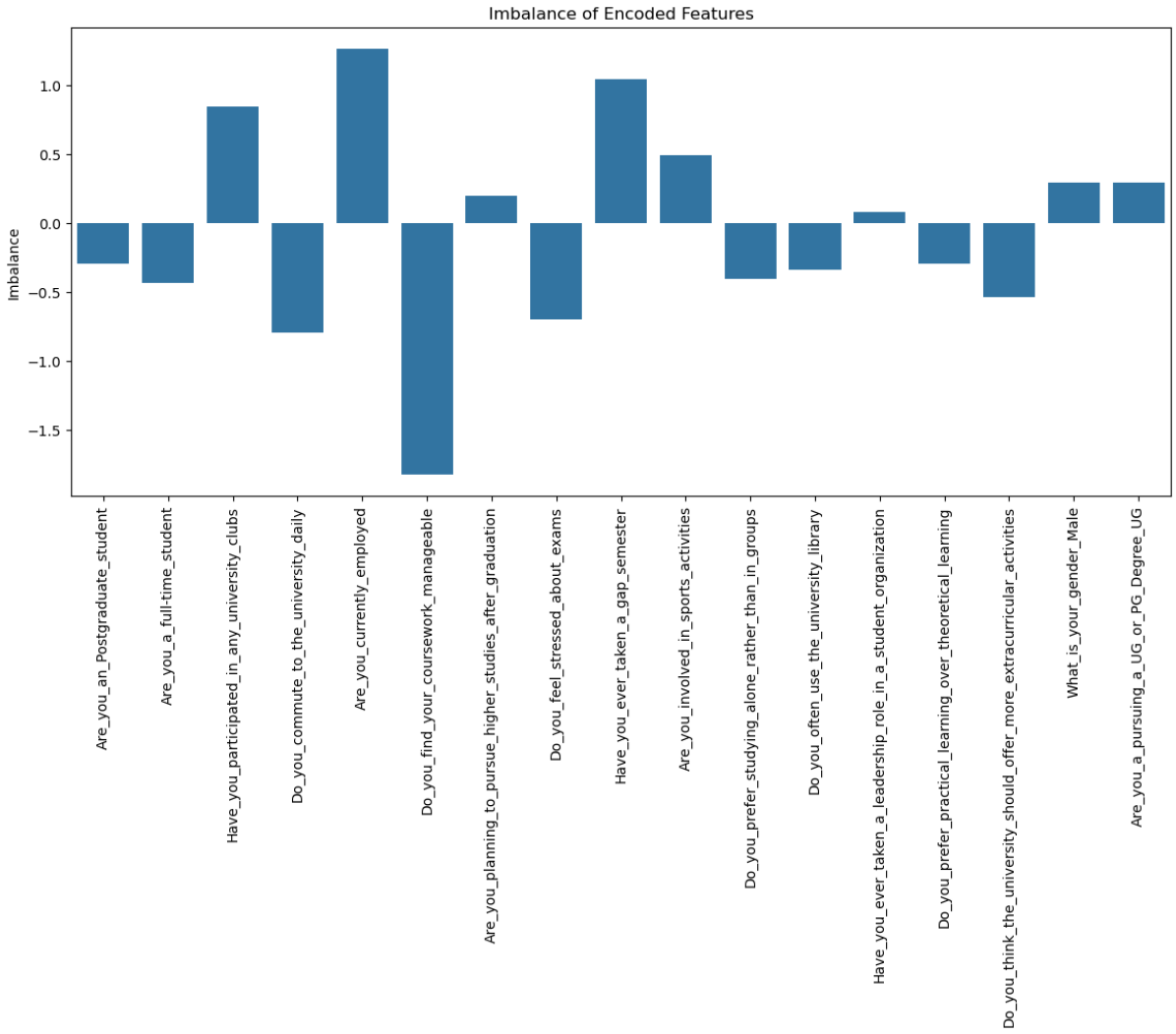

Figure 6.2: Data Imbalance graph