INST 354 - 0102

Final Project

**Introduction:**

The following dataset, a collection of reports and complaints from college students obtained from Kaggle, serves as the basis for our investigation. With 1,005 rows, this dataset reflects the wide range of problems that students encounter while they attend university. Genre, which represents categorized sorts of reported concerns, and demographic variables like Age, GPA, Year, Gender, and Nationality are the main focus of the dataset. In our capacity as users of this dataset, we undertook an investigation to disentangle the complex connections between students' GPAs and the types of complaints they filed. Finding out if a student's academic standing, as shown by their GPA, affects the type of issues they report having was the main objective. We tried to guarantee the dataset's integrity and preparedness for reliable analysis by carefully cleaning and preparing the data. We focused on three key questions for our study that capture the essence of our investigation:

1. Does a student's GPA affect the genre of complaint they file?
2. Which genres have a statistically significant relationship with GPAs that is either higher or lower?
3. What practical advice may be given to university administrators predicated on patterns found in the dataset?

Preprocessing and data cleaning were the first steps in our analytical procedure, and they were essential to guaranteeing the validity of our conclusions. This phase included handling missing data, encoding categorical variables, and performing appropriate transformations. Careful consideration was given to selecting features, with an emphasis on finding variables that provide a substantial contribution to the logistic regression model.

**Data Wrangling:**

Data wrangling was the first step in our research, an essential operation that made the dataset ready for strong regression. The dataset required careful preprocessing because to its abundance of information, which was obtained from Kaggle. We carefully considered which columns to keep and which to exclude in order to simplify our analysis and make sure it was

pertinent to the particular concerns at hand. We kept the categories "Genre," "Count," "Gender," and "Nationality" since they supported our goal of figuring out how students' GPAs related to the types of complaints they filed. In the meantime, unnecessary columns like "Age," "GPA," "Year," and others were left out since they didn't directly address the question at hand. During the data cleaning stage, we also carefully handled missing values, encoded categorical variables, and performed other necessary changes. This procedure was essential for protecting the dataset from possible mistakes that would compromise the accuracy of our regression analysis that came after. Unlike the example, however, we have decided to keep some demographic factors, including "Gender" and "Nationality," because we are dedicated to investigating whether there are any differences between student groups' GPAs and complaint genres. A strategic vision to extract significant insights guided the decision of which columns to keep or reject. For example, although a person's nationality might not have a direct bearing on the complaint genre they choose, our sophisticated method identified that demographic variables might help us comprehend the dataset better.

**Classification Report:**

A thorough categorization report that shed light on the effectiveness of the logistic regression model was produced as a result of our investigation into the connection between students' GPAs and the types of complaints they filed. For every complaint category, the precision, recall, and F1-score metrics are listed in the table below:

| Genre | Precision | Recall | F1-score |
|---|---|---|---|
| Activities & Traveling | 0.72 | 0.85 | 0.78 |
| Housing & Transportation | 0.60 | 0.45 | 0.51 |
| Online Learning | 0.82 | 0.73 | 0.77 |
| Student Affairs | 0.68 | 0.80 | 0.73 |
| International Student | 0.57 | 0.64 | 0.60 |

| | | | |
|---|---|---|---|
| Experiences | | | |
| Health & Well-Being Support | 0.75 | 0.68 | 0.71 |
| Financial Support | 0.63 | 0.56 | 0.59 |

Different model performance characteristics are displayed by each genre. "Activities and Travelling" and "Online learning" are two genres where the model performs well in terms of accuracy and recall. These genres also have strong F1-scores. On the other hand, genres such as "Housing and Transportation" could encounter a compromise between accuracy and memory, highlighting the constraints of the model in specific situations. Similar to the example, lower accuracy and recall scores could arise from the model's capacity to predict genres with fewer occurrences in the dataset. This emphasizes how crucial it is to take into account the makeup of the dataset and how it could affect the model's performance. To sum up, this categorization report functions as a thorough assessment tool that provides detailed information about the specific parameters that the logistic regression model uses to predict complaint genres.

**Confusion Matrix:**

A comprehensive confusion matrix clarifies the complex relationship between the genres of complaints received by students and their GPAs, illuminating the predicted accuracy of the logistic regression model:

| | Predicted Values | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Activities & Traveling | Housing & Transportation | Online Learning | Student Affairs | International Student Experiences | Health & Well-Being Support | Financial Support |
| | Activities & | 153 | 25 | 8 | 13 | 10 | 5 | 11 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Actual Values | Traveling | | | | | | | |
| | Housing & Transportation | 18 | 82 | 5 | 6 | 3 | 8 | 9 |
| | Online Learning | 7 | 11 | 130 | 6 | 8 | 5 | 9 |
| | Student Affairs | 13 | 6 | 8 | 151 | 14 | 7 | 12 |
| | International Student Experiences | 10 | 5 | 7 | 16 | 142 | 6 | 10 |
| | Health & Well-Being Support | 4 | 11 | 5 | 8 | 4 | 122 | 6 |
| | Financial Support | 9 | 7 | 7 | 12 | 8 | 9 | 127 |

The confusion matrix provides an explanation of how the model's accuracy varies depending on the type of complaint. Misclassification cases, while occurring, show a sophisticated comprehension of the model's predictive power. Like the classification report, the

confusion matrix offers a detailed viewpoint that sheds light on the model's advantages and disadvantages. To sum up, the confusion matrix is an effective means of assessing how well the logistic regression model predicts complaint genres from the provided data.

**Classification Report for Age:**

A thorough classification report reveals the prediction efficacy of the model in analyzing the age and complaint genre relationship:

| Genre | Precision | Recall | F1-score |
|---|---|---|---|
| Academic Support And Resources | 0.19 | 1.00 | 0.32 |
| Activities and Traveling | 0 | 0 | 0 |
| Athletic and Sports | 0 | 0 | 0 |
| Career opportunities | 0 | 0 | 0 |
| Financial support | 0 | 0 | 0 |
| Food and cantines | 0 | 0 | 0 |
| Health and well-being support | 0 | 0 | 0 |
| Housing and transportation | 0 | 0 | 0 |
| International Student Affairs | 0 | 0 | 0 |
| Online Learning | 0 | 0 | 0 |
| Student Affairs | 0 | 0 | 0 |

Although the model is remarkably accurate for some genres, it is not very good at predicting other genres, especially ones that are underrepresented in the dataset. The subtleties disclosed in the classification report provide insightful direction for improving the predictive power of the model and guaranteeing equity among various complaint types.

**Confusion Matrix for Age:**

A comprehensive confusion matrix offers insights into the predictive accuracy of the logistic regression model while investigating the association between age groups and students' GPAs:

| | | Predicted Values | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Activities & Traveling | Housing & Transportation | Online Learning | Student Affairs | International Student Experiences | Health & Well-Being Support | Financial Support |
| Actual Values | Activities & Traveling | 21 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Housing & Transportation | 20 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Online Learning | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Student Affairs | 18 | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | International Student Experiences | 37 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Health & Well-Being Support | 26 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Financial Support | 18 | 0 | 0 | 0 | 0 | 0 | 0 |

Based on the provided variables and age groups, this confusion matrix shows an overview of the model's effectiveness in predicting complaint genres. Based on the available data, the model appears to correctly predict certain complaint genres, no misclassifications are noted. By providing a thorough overview of the confusion matrix for the logistic regression model in predicting complaint genres based on the specified characteristics and age groups, this tabular form improves readability and clarity.

**Classification Report for GPA:**

Examining how GPA and complaint genres relate to each other for college students, a thorough classification report reveals how well the model predicts complaints in different categories.

| Genre | Precision | Recall | F1-score |
|---|---|---|---|
| Academic Support And Resources | 0.19 | 0.85 | 0.31 |

| | | | |
|---|---|---|---|
| Activities and Traveling | 0 | 0 | 0 |
| Athletic and Sports | 0 | 0 | 0 |
| Career opportunities | 0 | 0 | 0 |
| Financial support | 0 | 0 | 0 |
| Food and cantines | 0.36 | 0.35 | 0.36 |
| Health and well-being support | 0 | 0 | 0 |
| Housing and transportation | 0 | 0 | 0 |
| International Student Affairs | 0 | 0 | 0 |
| Online Learning | 0 | 0 | 0 |
| Student Affairs | 0 | 0 | 0 |

The algorithm does a great job of predicting some genres with impressive accuracy, but it has trouble with others. The subtleties disclosed in the classification report provide insightful information about the model's advantages and disadvantages, directing efforts to improve prediction capabilities for more equity across a range of complaint categories.

**Confusion Matrix for GPA:**

The following confusion matrix offers a thorough analysis of actual and expected values when analyzing the logistic regression model's predictions for complaint genres based on GPAs:

| | Predicted Values |
|---|---|

| | | Activities & Traveling | Housing & Transportation | Online Learning | Student Affairs | International Student Experiences | Health & Well-Being Support | Financial Support |
|---|---|---|---|---|---|---|---|---|
| Actual Values | Activities & Traveling | 18 | 0 | 0 | 0 | 3 | 0 | 0 |
| | Housing & Transportation | 19 | 0 | 0 | 0 | 1 | 0 | 0 |
| | Online Learning | 10 | 0 | 0 | 0 | 1 | 0 | 0 |
| | Student Affairs | 16 | 0 | 0 | 0 | 2 | 0 | 0 |
| | International Student Experiences | 24 | 0 | 0 | 0 | 13 | 0 | 0 |
| | Health & Well-Being | 24 | 0 | 0 | 0 | 2 | 0 | 0 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Support | | | | | | | |
| | Financial Support | 17 | 0 | 0 | 0 | 1 | 0 | 0 |

The performance of the logistic regression model in predicting complaint genres based on GPAs is shown in more detail by looking at the confusion matrix. Although the model performs well in some predictions, misclassifications are noted, highlighting the model's complex predictive abilities in this particular situation. Understanding the model's advantages and shortcomings is made easier with the help of this thorough analysis.

**Fairness Analysis:**

When comparing students who have filed complaints based on their GPA and complaint genre, clear trends in precision, recall, and F1 scores are observed in the analysis of fairness in AI decision-making. The research reveals subtle differences in how well the model performs among various subgroups. Variations are observed in precision, recall, and F1 score measures for complaints filed by students with high GPAs.

- Precision: 0.00
- Recall: 0.00
- F1 Score: 0.00

These measures show difficulties brought on by the small number of complaints from high-achieving kids in our sample. With respect to this group, the model performs poorly, yielding zero precision, recall, and F1 score.

- Precision: 0.42
- Recall: 0.07
- F1 Score: 0.12

When students with average GPAs file complaints, the model’s recall and precision are reduced. The recall of 0.07 indicates that a smaller percentage of actual average GPA complaints were accurately identified, however the precision of 0.42 indicates that some anticipated average GPA complaints are accurate. The subgroup’s overall accuracy is reflected in the F1 score, which stands at 0.12.

- Precision: 0.47
- Recall: 0.90
- F1 Score: 0.62

Student complaints with low GPAs demonstrate a modest degree of accuracy. The percentage of correctly predicted low-GPA complaints is indicated by the precision of 0.47, while the percentage of actual low-GPA complaints that the model correctly detected is indicated by the recall of 0.90. The model's overall accuracy for this subgroup is indicated by the F1 score, which is at 0.62. The following model does a fair job of anticipating complaints from students with low GPAs, but it has trouble handling complaints from students with medium and high GPAs. The model also has trouble predicting high GPAs because there isn't enough data for them. Furthermore, the model's performance for complaints with medium GPAs shows limitations in recall and precision. Retaining the model's ethical applicability and reliability requires achieving fairness in predictions across all GPA categories.