

**INSY 5378: Data Science: A Programming Approach**

**Group Project – Group 1 (Blue)**

|  |  |
| --- | --- |
| Submitted By: |  |
| Kajal Jain | 1001815317 |
| Tanvi J. Bapna | 1001719315 |
| Rohith Ramesh | 1001518031 |

**Submitted to:**

Mahyar S Vaghefi

**PROJECT:**

To develop a predictive model that can predict whether or not a movie is a Comedy

**PROJECT DESCRIPTION:**

For our project, we are using four different methods for text cleaning and text-processing namely:

• Removing Stop words

• TF-IDF

• Latent Dirichlet Allocation (LDA)

• Stemming

During each method, we applied multiple Supervised Learning models for calculating the accuracy of the prediction power of the models. Models used are:

• Logistic Regression

• Random Forest

• Multinomial Naïve Bayes

• Stochastic Gradient Descent

• Support Vector Machine

For each method, we compare the different model's Accuracy, F1, Recall, and Precision score. We used data visualization for showing the comparison.

After comparing different methods and models, for our project we got **Logistic Regression** as the best model with an **Accuracy Score of 78.3**when used along with **CounterVectorization**. Considering the Recall and F-1 value for the given model, which is best for the above-specified model: are 0.95, 0.87.

Classification report of the used model:

![Table

Description automatically generated]()

Thus, we are using Logistic Regression for training the main\_dataset and using it for Prediction purposes to calculate the accuracy of our final prediction on evaluation\_dataset. Our final model gives us an accuracy score of 72.0.

Classification report for the final\_model:![A picture containing text, receipt

Description automatically generated]()

Out of available 23498 values, our model can predict 16919 values correctly, which justifies the defined accuracy.