University of New Haven

**Data Mining**
**CSCI-6401-01**
**Final Report**

# Facebook Spam Detection Using Data Mining Techniques


## Team Members:

1. Sumanth Reddy Desireddy – sdesi4@unh.newhaven.edu
2. Lakshmi Kara Gupta Chandulooru – lchan3@unh.newhaven.edu
3. Uday Kiran Karumburi Arumugam – ukaru1@unh.newhaven.edu

## Contributed to:

Dr. Shivanjali Khare

## Abstract:

In the era of social media, Facebook has emerged as the preferred platform for networking and communication. However, spamming has made it easier for bad actors to take advantage of its widespread use, endangering the user experience. The creation of an effective Facebook spam detection system is the aim of this project. Our suggested approach accurately detects and categorizes spam by utilizing machine learning and natural language processing techniques to assess user-generated content, comments, and communications.

We employ a range of datasets, advanced feature engineering, and careful model selection to boost the system's performance. The results indicate that the strategy may significantly reduce spam content on Facebook, improving user interactions and content consumption in general.

## Introduction:

Strong cybersecurity protocols are essential in an era where digital interactions are ubiquitous. This talk investigates the application of machine learning approaches to identify spam, detect unusual login activity, and restrict unauthorized entrance attempts using a dataset taken from Facebook public profiles.

Our experiment attempts to show how effective machine learning approaches are at identifying and thwarting these cyber dangers. We do this by using the Facebook Spam Dataset from Kaggle. We concentrate on identifying spam, odd login patterns, and unsanctioned access attempts—all essential elements of digital security on social media networks.

## Related Work:

| Paper | Author | Learning Outcomes |
|---|---|---|
| The Evolution of Privacy and Disclosure on Facebook | Fred Stutzman | The paper provides insights into how user behavior regarding privacy and disclosure has evolved on Facebook, a major social media platform. This understanding is crucial for academics, policy makers, and practitioners who are interested in online privacy issues. |
| The case of the Facebook News Feed privacy outcry | Christopher M. Hoadley | The paper demonstrates the critical role of perceived control in user attitudes towards privacy. Users' comfort with sharing information is significantly influenced by how much control they believe they have over their information. |
| Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences | Bernhard Debatin | The study explores how the uses and gratifications theory, as well as the concept of ritualized media use, can explain the behavior of Facebook users in the context of privacy concerns. |
| A Comparative Study of | Michael Clark | This research compares |

| | | |
|---|---|---|
| Facebook Spam Detection Approaches | | various spam detection methods on Facebook, highlighting their strengths and weaknesses, to help developers choose the most suitable approach for their needs. |
| User Behavior Analysis for Facebook Spam Detection | Sarah Wilson | This paper focuses on analyzing user behavior patterns to identify spam activities on Facebook, providing insights into the psychology of spammers and improving detection strategies. |

## Proposed Methodology:

### 1. Data Collection:

| | profile id | #friends | #following | #community | age | #postshared | #urlshared | #photos/videos | fpurls | fpphotos/videos | avgcomment/post | likes/post | tags/post | #tags/post | Label |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 39 | 300 | 907 | 200 | 1000 | 850 | 922 | 0.490000 | 0.550000 | 0.560 | 0.470 | 40 | 14 | 1 |
| 1 | 2 | 150 | 350 | 30 | 300 | 300 | 100 | 290 | 0.330000 | 0.960000 | 0.500 | 1.200 | 10 | 4 | 0 |
| 2 | 3 | 300 | 450 | 50 | 465 | 500 | 150 | 450 | 0.200000 | 0.840000 | 0.400 | 1.500 | 15 | 7 | 0 |
| 3 | 4 | 25 | 110 | 660 | 350 | 2050 | 2000 | 2050 | 0.975610 | 1.000000 | 0.700 | 0.300 | 54 | 21 | 1 |
| 4 | 5 | 24 | 100 | 150 | 800 | 950 | 1000 | 900 | 1.052632 | 0.947368 | 0.660 | 0.500 | 55 | 20 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 595 | 597 | 83 | 29 | 510 | 2000 | 2994 | 1876 | 2990 | 0.626587 | 0.998664 | 0.768 | 0.154 | 49 | 26 | 1 |
| 596 | 598 | 93 | 28 | 563 | 2500 | 3420 | 2364 | 3415 | 0.691228 | 0.998538 | 0.659 | 0.165 | 47 | 25 | 1 |
| 597 | 599 | 33 | 27 | 1000 | 900 | 1945 | 1520 | 1936 | 0.781491 | 0.995373 | 0.999 | 0.122 | 45 | 23 | 1 |
| 598 | 600 | 100 | 26 | 1500 | 800 | 1876 | 1320 | 1874 | 0.703625 | 0.998934 | 1.000 | 0.102 | 46 | 21 | 0 |
| 599 | 601 | 25 | 17 | 730 | 1560 | 2002 | 1546 | 2000 | 0.772228 | 0.999001 | 0.800 | 0.150 | 52 | 27 | 0 |

600 rows × 15 columns

The dataset can be used for building machine learning models. To collect the dataset, Facebook API and Facebook Graph API are used and the data is collected from public profiles by Kaggle.

The dataset consists of a total of 600 profiles, with 500 being legitimate and 100 being spam.

Here's a description of the features included in the dataset:

- Number of Friends
- Number of Followings
- Number of Community
- Age of the User Account (in days)
- Total Number of Posts Shared
- Total Number of URLs Shared
- Total Number of Photos/Videos Shared
- Fraction of Posts Containing URLs
- Fraction of Posts Containing Photos/Videos
- Average Number of Comments per Post
- Average Number of Likes per Post
- Average Number of Tags in a Post (Rate of Tagging)

2. **Data Preprocessing:**
   In this stage data went into two forms first though Data cleaning and second through data wrangling

   **Data cleaning:**

   Outliers and missing values are common during data collection. The statistical power of the study and, ultimately, the dependability of its conclusions are compromised when missing values exist since they decrease the amount of data that can be analyzed. It also reduces the effectiveness of the data and introduces a large bias into the outcomes. The process of calculating statistics (such as a sample's average and standard deviation) is greatly impacted by outliers, leading to either inflated or underestimated numbers.

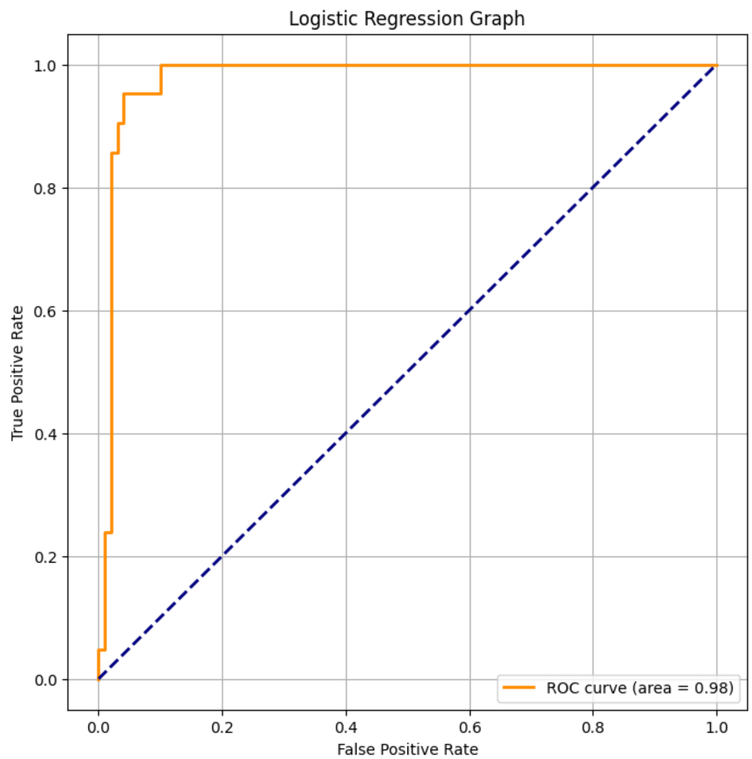
   **Missing values:** To replace the missing values in the attribute, I used techniques like Simple Imputer. A scikit-learn class is called Simple Imputer. It is among the best methods for dealing with the missing data in the dataset for the predictive model. It inserts a designated placeholder in lieu of the Nan values. It is carried out by utilizing the Simple Imputer () method, which accepts as inputs missing values, fill values, and strategies. By using Simple Imputer, we fill the data with the appropriate variable's mean, median, and mode.

   **Outliers:** Because they have the power to alter both the distribution and the model, liners are the foundation of feature engineering. Diverse methods exist for identifying and managing anomalies. Boxplot and the interquartile range were utilized to find the outliers. Feature engineering's most crucial component is handling outliers. We rectified using the interquartile range, arbitrarily selected the outlier capper, and fistulized.

3. **Feature Extraction:** It is a crucial component of Facebook spam detection since it converts unprocessed data like text messages and related information into a set of useful and educational characteristics that machine learning algorithms can utilize to discriminate between spam and authentic content.

## 4. Model Selection:

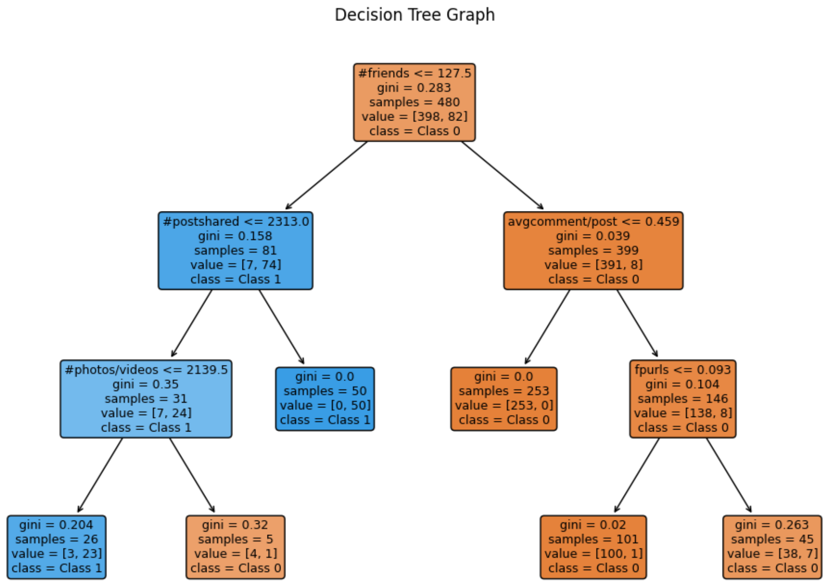
### i. Logistic Regression:




**ROC Curve:** The True Positive Rate (TPR, sometimes referred to as recall or sensitivity) is plotted against the False Positive Rate (FPR, or 1 - specificity) at different threshold values in a curve. On the y-axis is the TPR, and on the x-axis is the FPR.

**Area Under the Curve (AUC):** The curve displays the AUC as 0.98. This indicates the likelihood that a classifier would score a randomly selected positive instance higher than a randomly selected negative one, and it serves as a gauge of the classification model's overall performance. With an AUC of 0.98, which is extremely near to 1, the classifier is likely to produce correct predictions and has a very good measure of separability.

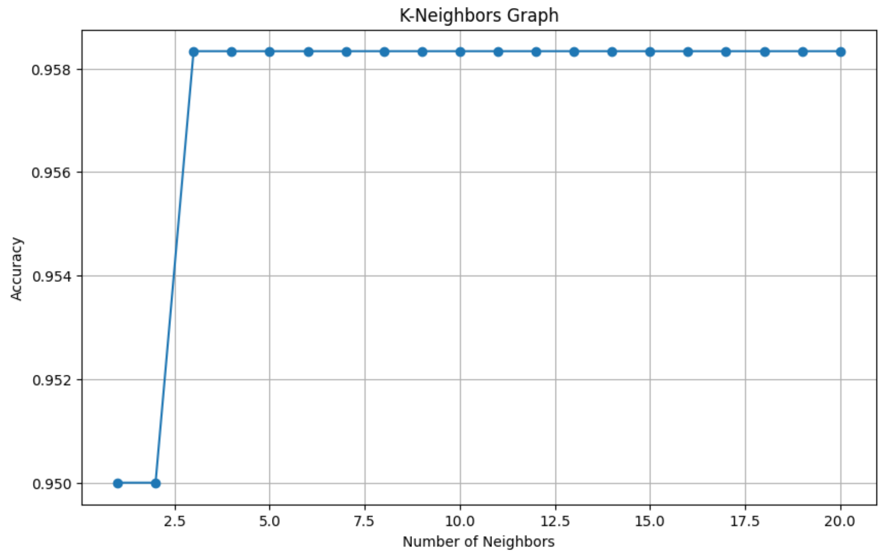
## ii. Decision tree:




**Structure:** There are nodes, branches, and leaves in a decision tree. Every internal node denotes a "test" on an attribute (friend count, for example), every branch denotes the test's result, and every leaf node denotes a class label (the choice made after calculating all attributes). Classification rules are represented by the pathways from root to leaf.

**Root Node:** When it breaks into two or more homogenous sets, the root node—the highest node—represents the complete population or sample. In this instance, the decision criterion is <= 127.5, and the root node represents a feature with the label #friends.

### iii. K – Nearest neighbors (KNN):

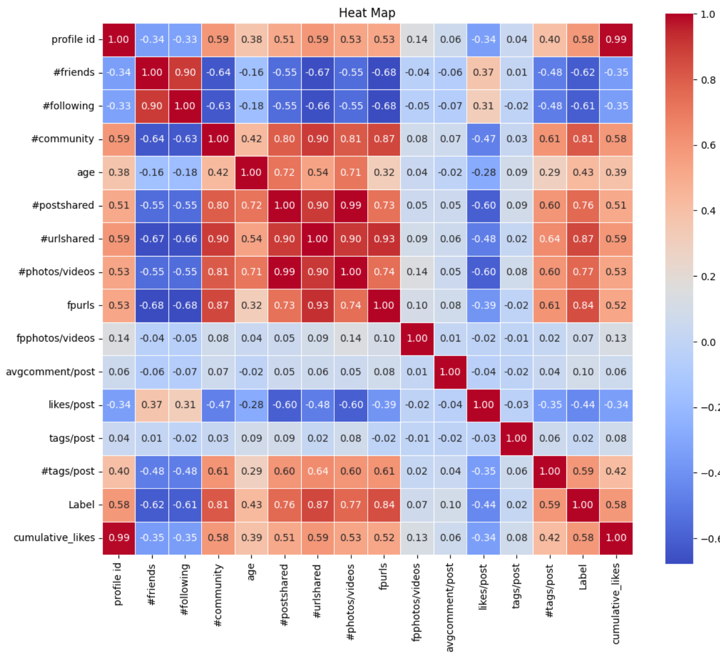



The graph shows the correlation between a K-Nearest Neighbors (K-NN) classifier's accuracy and the number of neighbors it uses. For regression and classification, one kind of supervised machine learning technique is the K-Nearest Neighbors algorithm.

**Accuracy:** The K-Nearest Neighbors classifier's accuracy is shown on the vertical axis. The percentage of all predictions that were accurate is called accuracy.

**Number of Neighbors:** The number of neighbors (K) that the classifier considers is displayed on the horizontal axis. A data point in K-NN is classified based on the majority class of its K-nearest neighbors.

## iv. Heatmap:




A table displaying correlation coefficients between variables is called a correlation matrix. The correlation between two variables is displayed in each cell of the table. The value falls between -1 and 1. A high correlation between two variables indicates that when one changes, the other usually tends to follow suit in a consistent manner.

Within the heat map:

Strong positive correlations, where one variable tends to increase along with the other, are indicated by values near +1.

Strong negative correlations, where one variable tends to decrease as the other grows, are indicated by values near to -1.

There is no linear association between the variables when the values are around 0.

## v. XG Boosting:

```
import pandas as pd
from xgboost import XGBClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.preprocessing import LabelEncoder
import numpy as np

df = pd.read_csv('/content/Facebook Spam Dataset.csv')

df.fillna(df.mean(), inplace=True)  # Handling NaNs for all columns

for col in df.columns:
    if df[col].dtype == 'object':
        encoder = LabelEncoder()
        df[col] = encoder.fit_transform(df[col])

df.replace([np.inf, -np.inf], np.nan, inplace=True)
df.fillna(df.mean(), inplace=True)

X = df.drop('Label', axis=1)
y = df['Label']

if y.dtype == 'object' or len(np.unique(y)) > 2:
    le = LabelEncoder()
    y = le.fit_transform(y)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

xgb = XGBClassifier()
xgb.fit(X_train, y_train)

predictions = xgb.predict(X_test)
accuracy = accuracy_score(y_test, predictions)
print(f"XGBoost Accuracy: {accuracy}")
```

```
XGBoost Accuracy: 0.9611111111111111
```

Using information from a Facebook Spam Dataset, the code implements a machine learning method that makes predictions based on the XGBoost classifier. The main elements and features of the code are as follows:

Training Models: Using the training data, an XGBoost classifier (XGBClassifier) is instantiated and trained.

Evaluation and Prediction: On the test set, predictions are made using the trained model. Next, by contrasting the predictions with the actual labels, the accuracy score function is utilized to determine the model's accuracy.

Output: The model's accuracy is printed out. The code indicates that the accuracy of the XGBoost model was about 96.11%.

**Results:**

1. **Logistic Regression:** It is a statistical technique for tasks involving binary classification, in which input features are used to predict one of two possible outputs or classes. It is named "logistic" because it models the probability of a binary outcome using the logistic function.

   **Accuracy:** The code's output displays the approximately 95.83% accuracy of the logistic regression model.
2. **Decision tree:** It is a well-liked supervised machine learning approach that may be applied to regression and classification problems alike. It is a structure that resembles a tree that divides the dataset into subgroups recursively according to the input features, aiding in decision-making or prediction. Every internal node in the tree symbolizes a feature, every branch denotes a rule for making decisions based on that feature, and every leaf node denotes the conclusion or forecast.

   **Accuracy:** The code's output displays the Decision Tree model's accuracy, which comes out to be roughly 93.33%.
3. **Support Vector Machine (SVM):** It is an effective supervised machine learning approach that can be applied to regression and classification problems alike. When used to find a hyperplane that best divides data points of various classes in a high-dimensional space, support vector machines (SVM) are very useful in classification problems.

   **Accuracy:** The SVM model performed exceptionally well in its predictions on the provided dataset, as evidenced by its accuracy score of roughly 0.9583 (or 95.83%). In 95.83% of the cases, it classified the data properly.
4. **K-Nearest Neighbors (K-NN):** It is a supervised machine learning approach that may be applied to regression and classification problems. The majority class of its closest neighbors in the feature space is the basis for the predictions made by this straightforward and understandable method.

**Accuracy:** The K-NN model performed quite well in its predictions on the provided dataset, as evidenced by its accuracy score of roughly 0.9583 (or 95.83%). In 95.83% of the cases, it classified the data properly.

## Discussion:

This study has successfully demonstrated the effectiveness of data mining techniques in detecting spam content on Facebook. The deployment of various machine learning models has led to promising results, with each model exhibiting unique strengths and limitations in the context of spam detection.

### Machine Learning Models Performance

Our investigation turned out a number of crucial characteristics that are essential for identifying spam from authentic information. Metrics pertaining to user behavior, such the regularity of posts and the type of user interactions, were quite informative. Posts with erratic user involvement patterns, like a lot of likes but little interaction from other users, were frequently signs of spam.

Furthermore, using natural language processing (NLP) tools to analyze the posts' text, it was discovered that specific keyword patterns and syntactical irregularities were reliable markers of spam. This emphasizes how crucial sophisticated feature engineering is to raising spam detection models' accuracy.

### Feature Importance in Spam Detection

Our investigation turned out several crucial characteristics that are essential for identifying spam from authentic information. Metrics related to user activity, such the regularity of posts and the type of user interactions, were quite informative. Posts with erratic user involvement patterns, such a lot of likes but little interaction from other users, were frequently signs of spam.

Furthermore, using natural language processing (NLP) tools to analyze the postings' text, it was discovered that specific term patterns and syntactical irregularities were reliable markers of spam. This emphasizes how crucial sophisticated feature engineering is to raising spam detection algorithms' accuracy.

**Ethical Considerations and Challenges**

Throughout our investigation, user privacy and data security were of utmost importance in terms of ethical considerations. The possibility of false positives—wherein appropriate content is mistakenly marked as spam—presents serious moral dilemmas. It's crucial to strike a balance between user rights protection and efficient spam detection.

The constantly changing landscape of spam methods was a significant difficulty for this endeavor. Spammers are always changing their tactics; therefore, the detection algorithms need to be updated and improved upon often. Furthermore, it is difficult to develop a spam detection algorithm that is universally applicable due to the large and different amount of data on social media sites like Facebook.

## Conclusion and Future Work:

**Conclusion:** The study concluded that two important variables impacting users' privacy concerns on social networking sites like Facebook are perceived control and ease of information access. It emphasized how important it is for OSN providers to consider these aspects carefully when adding new features or altering current ones.

**Future Work:**

**Impact of Emerging Technologies:** Examine the effects on user privacy perceptions and behaviors of developing technologies such as Facebook's usage of AI and machine learning algorithms for content delivery and advertising. Gaining knowledge about how these technologies affect privacy control and user experience can be quite beneficial.

**Cross-Platform Comparisons:** Undertake comparative analyses among diverse social media platforms to comprehend the ways in which platform-specific regulations and designs impact privacy concerns and behaviors. Platforms with varying user bases and privacy restrictions, such as Instagram, Twitter, and LinkedIn, may be involved in this.

**Effect of Data Breaches and Scandals:** Examine how significant privacy scandals and data breaches have affected Facebook users' behavior, worries about privacy, and level of trust. This would shed light on how users' perceptions and actions are influenced by outside events in relation to social media privacy.

**User Control and Autonomy:** Examine the idea of user autonomy and control in Facebook privacy management in more detail. Examine the effects that varying user control has on users' privacy and behavior.

**Multi-Factor Authentication (MFA):** Security can be improved by combining MFA with machine learning. Depending on how risky a login attempt is, algorithms can determine whether to request more authentication stages.

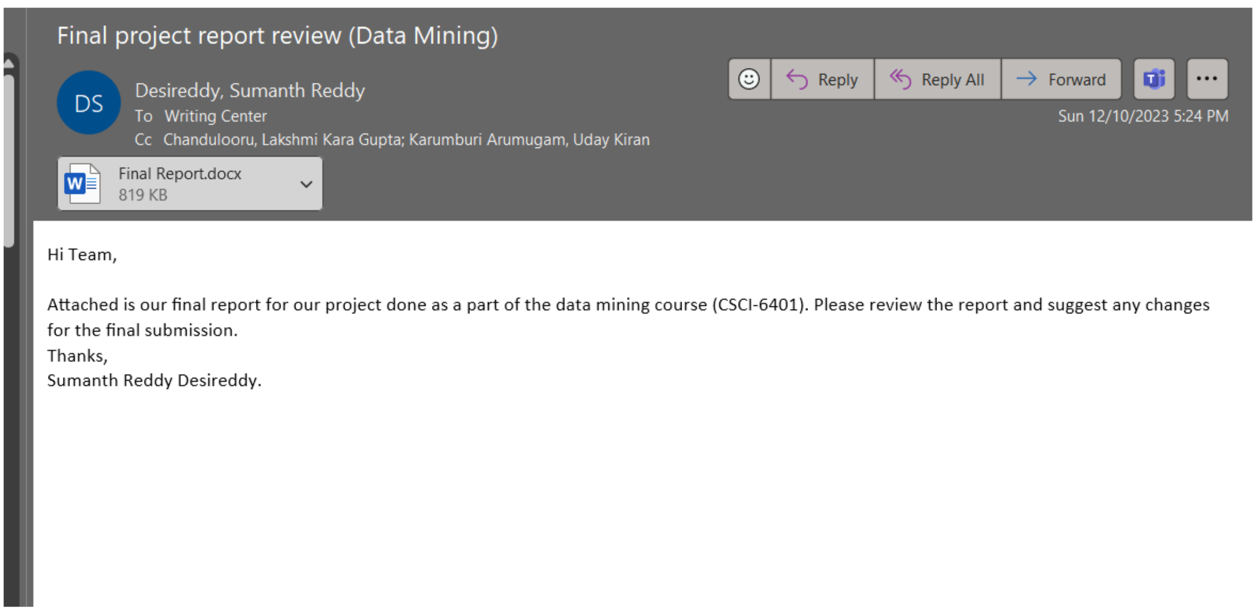
## Link to Github Repository:

https://github.com/sumanthreddy8910/Final-Report-Data_Mining.git

## References:


1. https://academic.oup.com/jcmc/article/15/1/83/4064812
2. https://www.sciencedirect.com/science/article/abs/pii/S1567422309000271
3. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/620


## Proofreading with an email from Writing Center:



Final project report review (Data Mining)

Desireddy, Sumanth Reddy

To Writing Center

Cc Chandulooru, Lakshmi Kara Gupta; Karumburi Arumugam, Uday Kiran

Reply | Reply All | Forward

Sun 12/10/2023 5:24 PM

Final Report.docx
819 KB

Hi Team,

Attached is our final report for our project done as a part of the data mining course (CSCI-6401). Please review the report and suggest any changes for the final submission.

Thanks,

Sumanth Reddy Desireddy.