|  |
| --- |
|  |
| Final Project Plan  Data Wizards |
| |  |  |  | | --- | --- | --- | | Shivam Choudhary (100949298) Ishan Sevak (100951570) Sai Kurmapu (100951915) Bhavika Lathigara (100917684) Thejanu Meze (100943100) Monisha Senthil Velu (100933456);Choudharyshivam09@outlook.com | 6/6/25 |  | |

Table of Contents

[Problem/Analytical Statement 2](#_Toc200135571)

[Project Plan 2](#_Toc200135572)

[Project Significance 2](#_Toc200135573)

[Anticipated Outcomes 3](#_Toc200135574)

[Course Relevance 3](#_Toc200135575)

[Assumptions 3](#_Toc200135576)

[Constraints 4](#_Toc200135577)

[Timeline and Deliverables 5](#_Toc200135578)

[References 6](#_Toc200135579)

# Problem/Analytical Statement

In Modern Digital Age it is challenging to determine and differentiate between fake and real news in the media articles. Based on which, our clients want to develop a machine learning model to predict if the news is "Fake" or "Real" based on the content of the news. This model will not only help us in identifying fake news so that it can be fact-checked but using this will also help to mitigate the impact caused by fake news.

We are using “fake-and-real-news-dataset” from Kaggle and about to start implementing classification algorithms which will help us to accurately classify news articles as real or fake. We will use metrices like Accuracy, Precision, Recall, and F1-Score to evaluate model effectiveness as default. Our model’s success will be defined by F1-score of 85% or above.

# Project Plan

## Project Significance

The problem of fake news reduces people’s trust, makes it hard to make wise decisions and threatens democracy. The fact that digital news is consumed more and more makes it crucial to have fast and trustworthy fact-checking tools. The aim of this project is to design a machine learning model able to group false news from genuine information.

Besides constructing a model that forecasts outcomes, our goal is to make an API that journalists, editors and the public use to verify news easily.

**Social and Practical Impact**

* Helps maintain the integrity of the media by lessening the distribution of false information.
* Helps people believe in journalism again by highlighting suspected articles.
* It keeps its relevance with quarterly training to meet new trends in false news.
* These results will clearly improve information quality and digital literacy as media becomes noisier.

## Anticipated Outcomes

* Create a model that can determine if news is real or fake, with a goal of reaching at least an 85% F1-score.
* Make an API that allows users to verify the accuracy of their articles with a simple assessment.
* Look for common features in fake news, including common words which can improve the way content is screened.
* Repeat retraining the model every quarter to avoid any performance issues as trends change.
* Prevent the spreading of false news, support readers’ trust in the media and help good journalism.

## Course Relevance

This project aligns with the following courses:

* DATA 1202: Introduction to Data Analytics (for model development and evaluation).
* DATA 2204: Statistics II (for advanced data analysis techniques and model optimization).
* DATA 1202: Data Analysis Tools Analytics (for data cleaning, EDA)

We will utilize tools such as Python (for model implementation using Scikit-learn), along with machine learning algorithms such as Random Forests and Logistic Regression.

## Assumptions

1. The Dataset Reflects Real-World News Content – We assume that our model understands the difference between real and fake news by memorizing examples contained in the data. We think this dataset has the same tone, topics and style as news that’s found online nowadays. If we only use outdated info, the model is likely to have problems in real predictions.

2. Assumes News is Clearly Written – Assumption that news articles in dataset is in formal English, as you see from official news sources, work best with the model. When writing contains a lot of slang, has poor grammar or is printed in a language the model cannot read, the model might misunderstand and give out incorrect predictions. Our project assumes that the content has good writing and is easy to handle, as was the case for the articles in its training data.

3. Spotting Patterns in How News is Written – We assume that fake news tends to be written differently from genuine news. Sometimes, fake articles use sensational titles, emotional words and unrealistic facts. When the news is accurate, facts are clear, and it is not biased. It begins to spot the distinctions to effectively classify one from another. However, if fake news becomes like real news, the model may get confused.

4. News Writing Won’t Change Too Quickly – We must make assumption that how fake news and real news are written will probably stay the same. The model was created using a collection of past articles which means it might miss some new tricks used in fake news today. However, if fake news is written in a new way, more training may be required to maintain the model’s accuracy.

## Constraints

1. Limit to one Language – We are going to train our model based on newspaper articles from English newspapers written in standard, formal English. This Can affect the performance of the model in other languages. It restrict our model to English language only.
   * Type – Hard constrain.
   * Assumption – Assumes News is Clearly Written.
2. Depending on Textual Patterns – The success of the model lies in assuming fake and true news can be told in different ways. However, if people who create fake news start writing in a style like real news, the model could have a harder time telling them apart which might reduce the model’s overall accuracy.
   * Type – Soft constrains.
   * Assumption - Spotting Patterns in How News is Written.
3. The Model May Get Outdated - Information from old news stories is used to train the model. However, the tricks and writing approaches of fake news can alter with time. A heavy shift in the content may cause the model to struggle. Updating the data in the model regularly is necessary to preserve its accuracy.
   * Type – Soft constrains.
   * Assumption – News Writing Won’t Change Too Quickly.
4. Dataset Quality - Our model will be trained on a certain dataset, and it believes that this dataset truly reflects real news writing. If what’s in the data does not match what is being reported currently such as when it is old or has biases, it could affect the model’s performance. This may produce miscalculations and lower the level of trust in what the system provides.
   * Type – Hard constrains.
   * Assumption - The Dataset Reflects Real-World News Content.

# Timeline and Deliverables

|  |  |  |
| --- | --- | --- |
| Deliverable | Due Date | Description |
| Project Proposal | Completed | Outlined the project scope, objectives, dataset, stakeholders, and methods. |
| Data Collection & Cleaning | June 7, 2025 | Download and inspect the dataset from Kaggle. Perform text preprocessing. |
| Preliminary Model Development | June 14, 2025 | Implement baseline models. Evaluate and compare with metrics like Accuracy, Precision, Recall, and F1-score. |
| Model Optimization & Finalization | June 17, 2025 | Tune hyperparameters, test advanced models (Random Forest, Gradient Boosting), and finalize the best-performing one. |
| API Integration | July 5, 2025 | Develop an API to allow editors/moderators to use the model for article classification, with confidence score output. |
| Final Report | July 19, 2025 | Compile technical documentation, evaluation results, model explanation, feature analysis, and recommendations for retraining and usage. |
| Final Presentation | Week of August 3, 2025 | Present the project to classmates and faculty: cover the problem, approach, models used, performance, and business value. |

(Dutta, 2025)

# References

Dutta, R. (2025). *Capstone Group Project Project Plan-Example*. Retrieved June 06, 2025, from DC COnnect Durham College: https://durhamcollege.desire2learn.com/d2l/le/content/620744/viewContent/8654265/View