**IST-707-MACHINE LEARNING**

Dear Professor,

Please follow the steps below to review the folder contents in the GIVEN order:

**1. Read the Project Overview**

* **File:** README.md
* **Action:**  
  Start by reading this file first.  
  It explains the overall goal: detecting misinformation in online health blogs using Machine Learning (ML), PageRank, and Explainable AI (SHAP values).

**2. Review the Notebooks (in sequence)**

|  |  |  |
| --- | --- | --- |
| **Order** | **File** | **Description** |
| 2.1 | Extraction\_of\_weblinks.ipynb | Web scraping of health blogs (extraction of URLs). |
| 2.2 | Content\_Analysis (2).ipynb | Analyzing blog content: text extraction, feature creation. |
| 2.3 | customer\_ratings1.ipynb | Classifying customer ratings using ML models (KNN, RF, XGBoost). |
| 2.4 | Page\_Rank\_score (1).ipynb | Calculating PageRank scores of websites for credibility analysis. |
| 2.5 | Regression\_score (1).ipynb | Using Regression models (Linear, SVM, Decision Tree) for scoring credibility. |
| 2.6 | XAI\_Shapley\_final.ipynb | Applying Explainable AI (SHAP) to interpret feature impacts on predictions. |

**Recommended:**  
Please follow this notebook order step-by-step, as the work is interconnected (data collection → feature building → model training → explainability).

**3. Review the Supporting Python Scripts**

|  |  |
| --- | --- |
| **File** | **Purpose** |
| NibblerScraper.py | Scrapes technical website scores (like SEO, mobile-friendliness) from Nibbler tool. |
| web\_scraper.py | Scrapes paragraph text from blog pages. |

**4. Review the Dataset File**

* **File:** healthblog credibility detection - customer ratings dataset (1).csv
* **Action:**  
  This CSV file contains scraped or preprocessed customer rating data used in the customer ratings analysis notebook.

**Software Requirements**

|  |  |
| --- | --- |
| **Purpose** | **Software/Library Needed** |
| Reading README and .csv | Text editor (VSCode, Notepad++) or Excel |
| Running Jupyter Notebooks | **Python 3.8 or higher** + **Jupyter Notebook** environment (Anaconda recommended) |

**Python Packages Needed:**

* pandas
* numpy
* sklearn
* matplotlib
* seaborn
* nltk
* shap
* networkx
* beautifulsoup4
* requests
* csv
* re

**Final Access Summary**

1. **Start** with README.md.
2. **Follow** the Jupyter Notebooks in the order listed above.
3. **Refer** to supporting .py scripts as needed.
4. **Open** the dataset if needed for deeper exploration.

**THANKYOU!**

**GITHUB:-** <https://github.com/MIHIR0501/Credibility-Detection-of-Health-Web-Blogs-Using-Explainable-AI>