**Capstone 3: Project Ideas**

**A. Sentiment Analysis for Mental Health**

This comprehensive dataset is a meticulously curated collection of mental health statuses tagged from various statements. The dataset amalgamates raw data from multiple sources, cleaned and compiled to create a robust resource for developing chatbots and performing sentiment analysis. The dataset contains more than 50,000 statements and their respective status labeled as either “Normal”, “Depression”, “Suicidal”, “Anxiety”, “Bipolar”, “Stress”, and “Personality disorder”.

The dataset is built by **Suchintika Sarkar** by integrating information from the following Kaggle datasets:

1. 3k Conversations Dataset for Chatbot

2. Depression Reddit Cleaned

3. Human Stress Prediction

4. Predicting Anxiety in Mental Health Data

5. Mental Health Dataset Bipolar

6. Reddit Mental Health Data

7. Students Anxiety and Depression Dataset

8. Suicidal Mental Health Dataset

9. Suicidal Tweet Detection Dataset

The dataset contains following features:

1. unique\_id: A unique identifier for each entry.

2. Statement: The textual data or post.

3. Mental Health Status: The tagged mental health status of the statement.

The data is available for download on <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>. I plan to develop my own Sentiment analysis to gauge mental health trends from this data. This project falls under the Natural Language Processing (NLP) project.

**B. Medical Recommendation dataset**

The Medicine Recommendation System is designed to assist medical professionals in suggesting appropriate medications for patients based on their reported symptoms and diagnosed diseases.

The dataset contain records of patient cases, each comprising,

1. patient's DOB,

2. Gender,

3. reported Causes,

4.reported symptoms,

4. identified disease,

5. recommended medicine for treatment.

For the development and evaluation of the Medicine Recommendation System, the dataset was created by **Joy Mathew**, and is available on <https://www.kaggle.com/datasets/joymarhew/medical-reccomadation-dataset>. This project falls under the category of Recommendation systems which predicts the best possible medicine based on the features/information provided.

**C. Residential power usage 3 years data - Timeseries**

The dataset contains hourly power usage of 2 storied house located in Houston, Texas, USA. The data set contains hourly power usage in kwh starting from 01-06-2016 to August 2020. The dataset has marked notes for weekdays, weekends, COVID lockdown & vacation days in notes category column.

The power usage data, extracted from “TRIEAGLE ENERGY LP, The Woodlands, Texas 77393” and the historical weather data for Houston, Texas extracted from “[www.wunderground.com”](http://www.wunderground.xn--com-9o0a/" \t "_blank) by **Sri Polu** and the data is available on <https://www.kaggle.com/datasets/srinuti/residential-power-usage-3years-data-timeseries/data>.

Power usage during daytime is different from nighttime. The electrical devices that are inside the house are security DVR and POI cameras, 2 x refrigerators, 2 x 50 gallon water heater that are on during daytime. At night several electrical bulbs, TV's, washing machine, dryer and AC run from evening 6 pm to morning 8 am. Another data contains historical weather report of Houston, Texas starting from 01-06-2016 to August 2020.

Power Usage 2016 to 2020 dataset has:

1. StartDate

2. Value (kWh)

3. days\_of\_week

4. notes

[column:- Vacation setting:- '"AC and electric bulbs turned off"

COVID- Lockdown:- 'AC is turned on during daytime, laptops, monitors etc., are on"

Weekday:- 'Morning 7am to 5pm "AC is 84 F temperature during summer and heating set at 60 F during winter".

Weekends:- 'Room Temperature is set at 78 F during summer and 68 F for heating during winter"]

Weather 2016 to 2020 daily dataset has:

|  |  |
| --- | --- |
| 1. Date  2. Day  3. Temp\_max (F deg)  4. Temp\_avg (F deg)  5. Temp\_min (F deg)  6. Dew\_max (F deg)  7. Dew \_avg (F deg)  8. Dew \_min (F deg)  9. Hum\_max (in %)  10. Hum\_avg (in %) | 11. Hum\_min (in %)  12. Wind\_max (mph)  13. Wind\_avg (mph)  14. Wind\_min (mph)  15. Press\_max (mmHg)  16. Press\_avg (mmHg)  17. Press\_min (mmHg)  18. Precipit (inch)  19. days\_of \_week |

My goal is to predict the power usage for the day for a given day and temperature details. This project falls under the category of Time Series analysis.