Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

1.About me

Name : NIket Kumar Dheeryan
Email : Nkdheeryan01@gmail.com
Phone : +91(8851019918)
Github : https://github.com/Niketkumardheeryan
Linkedin : **click here**
Country : India
Timezone : **I**ndian Standard Time (UTC +5:30)

2. Project Proposal

2.1 Abstract

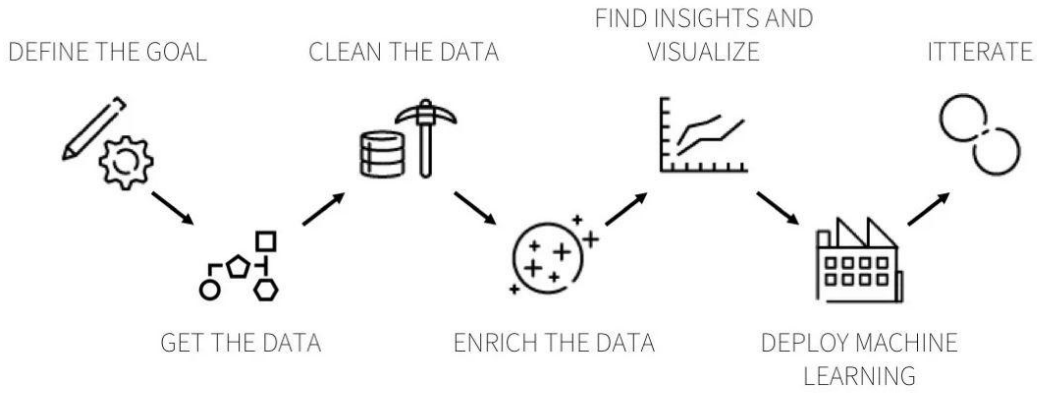
Our recent research dramatically changing the understanding of the secrets of the galaxy. And the whole universe is filled with gas, so the study of gases could be beneficial to understand the galaxies. In recent days, to know the evolution of galaxies, the study of CMG(Circumgalactic Medium) is required. CGM contains useful information about atoms and ions. Or Machine learning is a powerful tool to study anything. We are using machine learning models to understand CMG. But unnecessary Features present in the Dataset reduce the accuracy and increase the time complexity as well. Dimension reduction techniques can remove this problem easily.

2.2 Background

In machine learning, Dimension reduction is a method to achieve useful features or columns for prediction. There are lots of techniques used for dimension reduction ie. PCA, RFE, LDA, etc. After reducing dimensions a machine learning algorithm can perform well.it provides better dependent features to train the models.

2.2 Goals

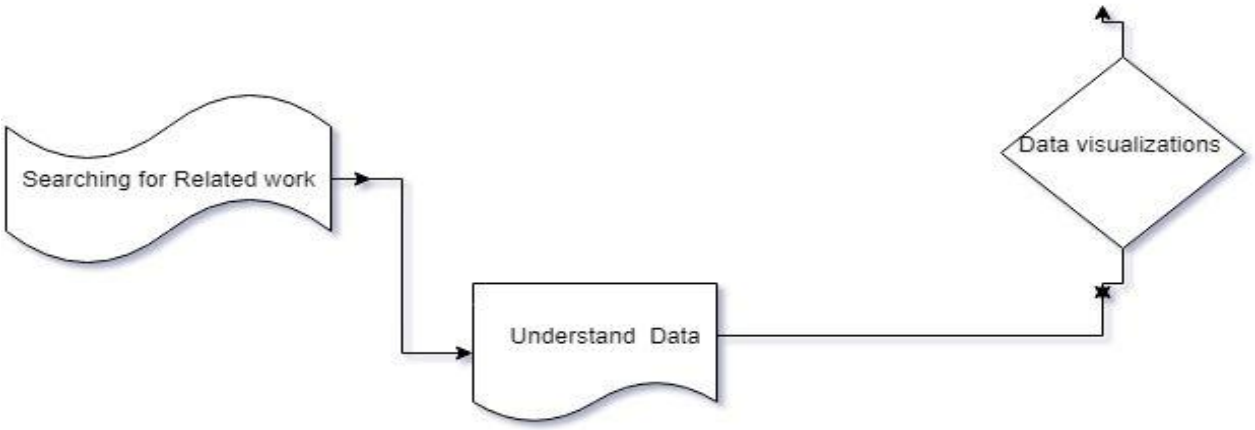
Basic Goals as shown below



3. Timezone and deliverables

Week 1--3 (May 17 - June 7)

- Searching for related work available ie. Research papers
- Understand the Dataset by visualization tools like Power BI, Google Data studio.



Basic Planning for First 3 weeks

. Week 4--5 (June 7 - June 21)

- Data analysis
- - Statistics analysis
- - handling Distribution of features
  - - Hypothesis generation
  - - Chi-squared test
- - Handling outliers
- - NULL values filling
- - Features Encoding

- - Different types of transformations
- - Feature scaling
    - - Standard scaling
    - - Robust scaling
    - - Normalization

Code snippets:

```
# Knowing the Data
def about_data(df):
    col=df.columns
    skewness=df.skew()
    return col, df.shape, df.isnull().sum(), skewness

```

Week 6 (June 21 -June 28)

- Training for Basics Models

Code snippets:

```
# 2. Logistic Regression
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(xtrain1,ytrain1)
```

```
# structure of NN

def build_model():
    model = Sequential()

    model.add(Dense(128, activation="relu", input_shape = (xtrain1.shape[1],)))  # Hidden Layer 1
    model.add(Dense(64, activation="relu"))                                      # Hidden Layer 2
    model.add(Dropout(0.2))
    model.add(Dense(32, activation="relu"))                                      # Hidden Layer 3
    model.add(Dropout(0.2))
    model.add(Dense(16, activation="relu"))                                      # Hidden Layer 4
    model.add(Dropout(0.2))
    model.add(Dense(1, activation="sigmoid"))                                    # Outout Layer
    model.summary()

    return model
```

Week 7--8 (June 28 - July 12)

- Optimization for each Machine learning model
    - Grid search cv
    - Iteration method

- Searching for First label issues
- Writing Blogs on related work

Week 9--10(July 12 - July 26)

- Dimension Reduction techniques
    - PCA
    - LDA
    - RFE
    - etc

- Iteration for all Models to improve accuracy

Code snippets:

```
# PCA
from sklearn.decomposition import PCA
pca = PCA(n_components=20)
x_mod = pca.fit_transform(x)
explained_variance=pca.explained_variance_ratio_
explained_variance.sort()
explained_variance
```

Week 11(July 26 - Aug 2)

- Testing models on External datasets.
    - Different evaluation metrics for accuracy
- Bug Fixing
- Blog writing at the end of the week.

Week 12 (Aug 2- Aug 9)

- Documentation of project

- I Will try to create a Mobile Application
- Web application as well
- Blogs

- Request for merge PRs

Week 13 (Aug 9 - Aug 16)

- I will try to merge more pull requests
- Blogs

Week 14(Aug 16 - Aug 23)

- Blogs
- I will create more interactive UI for this project

4. Possible Extensions

- We could create a custom Dimension reduction technique

- We could find the total amount of atoms present in CMG
    - On the given dataset find the total amount of atoms present in CGM

- Show the reactiveness of an ion

5. Related Projects

My related projects are given below

1. **Covid-19 Tracker**

   It is a machine learning model which predicts the percentage of affected people based on given ages and deployed on Heroku Cloud.

2. **Churn rate identification**

   It is a data analysis or dimension reduction Project. which classifies whether a customer will purchase the product in the future or not.

3. **Buyer's Time prediction**

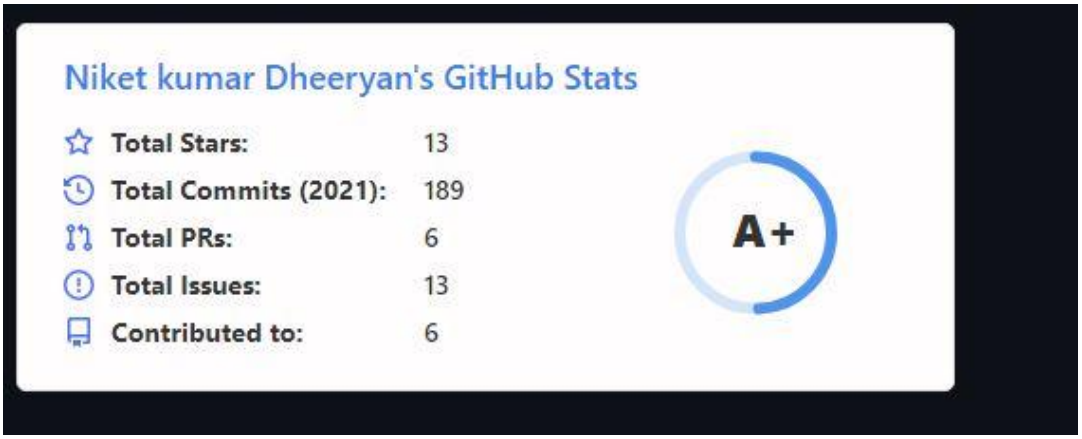
   It is also a machine learning model to predict how much time a buyer spends to buy a product. This project was created during the hackathon at **MachineHack** and Got the rank in the top 3 %.

I have participated in several machine learning Hackathons as well and got a good rank.

6. Open Source work Experience

My Github - click here

- In Progress
- I have contributed to 6 repositories,4 of which were merged in **Hacktoberfest 2020.**



Snip is taken from my Github profile

.

- I commit continuously and try to learn through contribution in open source projects.

- I create issues, PRs, and commits as well. As shown below.

- Currently, I am a mentor of an Open-source organization GSSOC 21 (**Girl script summer of code**). I have created more than 15 issues and merged pull requests as well.

7. Educational and programming Background

I'm an undergraduate student from Bharati Vidyapeeth's College of Engineering New Delhi, specializing in computer science. I have 2 years of experience in Python and 1.5 years of experience in Machine learning. Recently published a research paper in the Machine learning field.

8. Additional Notes

I'm planning to dedicate 42 hours per week (7 Hours /Day, Mon to Sat). Sunday will be spent restructuring the code created throughout the week and submitting it in a pull request for evaluation**.**

The project will utilize the git-flow methodology
:

- Features will be developed in separate branches and merged into develop at the end of each phase.

- The develop branch will be merged into master on the midterm and final evaluation releases, and corresponding GitHub releases (provisional SemVer versions: 0.5.0 and 1.0.0 respectively) will be created.

9.References

1. https://www.britannica.com/science/interstellar-medium
2. http://astronomy.nmsu.edu/cwc/Group/QALsims/
3. http://rongmonbordoloi.com/CGM.html