Machine Learning for Science (ML4SCI) Umbrella Organization

Proposal-GSoc 2021

Personal Details

- **Name:** Deepthi M (weeks)
- **Email:** deepthireddymavillapally@gmail.com
- **Country:** India
- **Timezone:** IST (India)
- **School Name & Study:** Lovely Professional University, Computer Science and Engineering (B.Tech(Hons.))
- **Typical Working Hours:** 10.30-12:30, 15:30-17:00, 20:30-22:30 (IST)
- **GitHub profile:** https://github.com/deepthi1107

Project Proposal

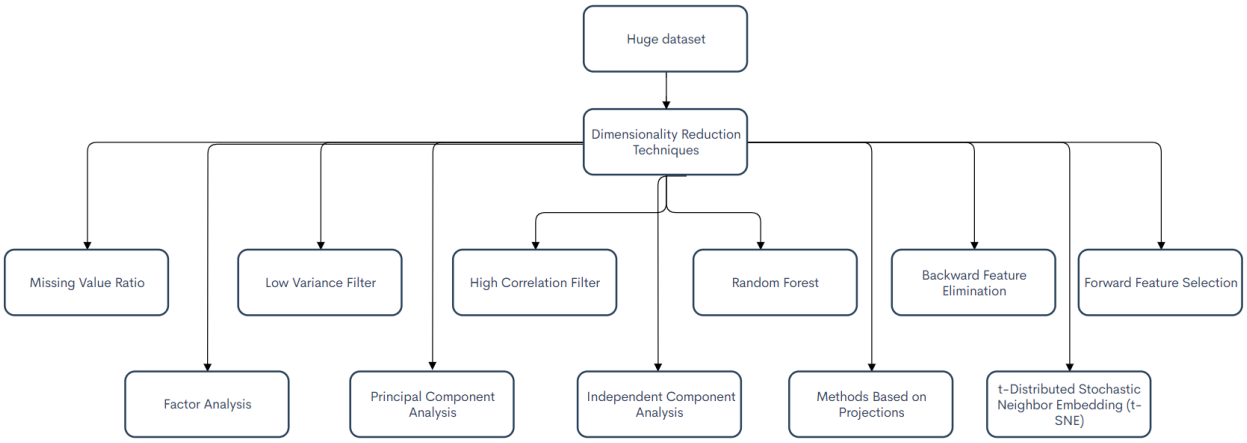
Project Title

Dimensionality Reduction for Studying Diffuse Circumgalactic Medium.

Overview

Having a high number of variables is both a boon and a curse. It's great that we have loads of data for analysis, but it is challenging due to size.It's not feasible to analyze each and every variable at a microscopic level. It might take us days or months to perform any meaningful analysis and we'll lose a ton of time and money for our business. We need a better way to deal with high dimensional data so that we can quickly extract patterns and insights from it.Using dimensionality reduction techniques we can reduce the number of features in your

dataset without having to lose much information and keep (or improve) the model's performance.
As the history of a galaxy is having the huge columns and its harder-to-detect diffuse circumgalactic medium (CGM), using machine learning tools to use quasar absorption lines to determine the gas composition, temperature and density make tasks easier.
Through this project i will work on the data to reduce its dimensions using various techniques(Missing Value Ratio, Low Variance Filter, High Correlation Filter, Random Forest, Backward Feature Elimination, Forward Feature Selection, Factor Analysis, Principal Component Analysis, Independent Component Analysis, Methods Based on Projections, t-Distributed Stochastic Neighbor Embedding (t-SNE), UMAP) and find the accurate technique to improve the model's performance.



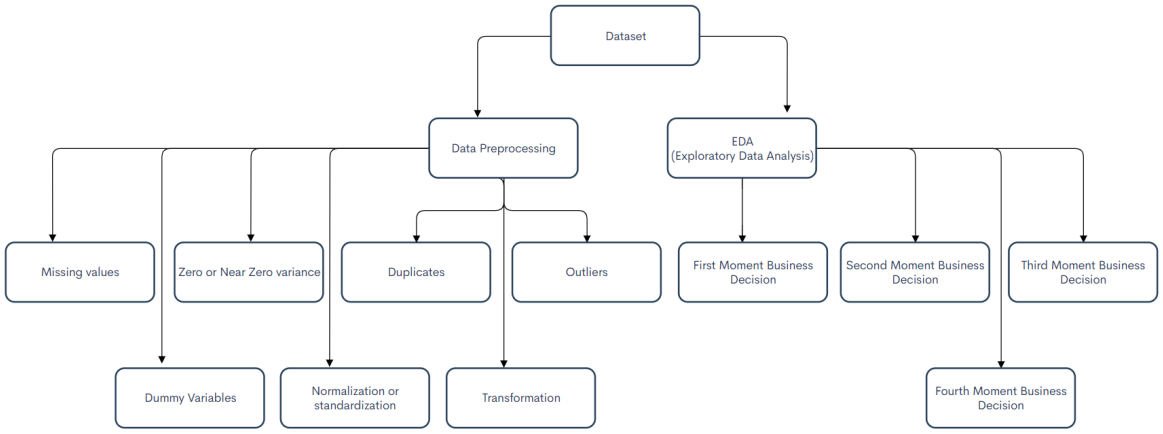
Timeline

Week 1 (May 17 - May 23)

- Understanding the
  - ❖ dataset
  - ❖ Relationship between the columns
  - ❖ Importance of each column

Week 2 (May 24 - May 30)

● Exploratory data analysis of the dataset so that there is no data loss.
- ❖ First Moment Business Decision
- ❖ Second Moment Business Decision
- ❖ Third Moment Business Decision
- ❖ Fourth Moment Business Decision



Week 3 (May 31 - June 06)

● Data Preprocessing (every column and row is important there should be no data loss)
- ❖ Missing values
- ❖ Zero or Near Zero variance
- ❖ Duplicates
- ❖ Outliers
- ❖ Dummy Variables
- ❖ Normalization or standardization
- ❖ Transformation

Week 4 (June 07 - June 13)

● Apply various Machine Learning technique (dimension reduction)
- ❖ Missing Value Ratio
- ❖ Low Variance Filter
- ❖ High Correlation Filter
- ❖ Random Forest
- ❖ Backward Feature Elimination

- ❖ Forward Feature Selection
- ❖ Factor Analysis
- ❖ Principal Component Analysis
- ❖ Independent Component Analysis
- ❖ Methods Based on Projections
- ❖ t-Distributed Stochastic Neighbor Embedding (t-SNE)
- ❖ UMAP

Week 5 (June 14 - June 20)

● Analysis of the accurate technique for the model accuracy.

Week 6 (June 21 - June 27)

● Discuss with mentors and follow their guidelines.

Week 7 (June 28 - July 04)

● Debug the implementation and improve the performance.

Week 8 (July 05 - July 11)

● View the accurate machine learning technique for dimension reduction and if any changes require upgrading them.

Week 9 (July 12 - July 18)

● Clean up the code and discuss the final techniques with the mentors.

Week 10 (July 19 - July 25)

● Create a pull request.

Week 11 (July 26 - August 01)

● Change according to the discussion in the pull request.

Week 12 (August 02 - August 08)

● Merge and write documentation for evaluation.

Availability:

- As I am a student pursuing btech in india its totally fine and comfortable for me to work for 4-6 hours/day and if required I can even work for more hours in the weekends .I might be having exams for 12 days in the month of May but will be able to work for 2 hours on those days.

Technical requirement

- For doing the project one must have the knowledge of data science ,machine learning and python. I have hands-on experience in data science and machine learning techniques by using python programming language.

Benefits of dimension reduction

- It reduces the time and storage space required.
- The removal of multicollinearity improves the interpretation of the parameters of the machine learning model.
- It becomes easier to visualize the data when reduced to very low dimensions such as 2D or 3D.

Contributions

- As a part of GSSoC'21, I have started doing contributions in the field of data science to the project (Stack Overflow Tag Predictions web application).

Previous Projects

1. Worked with the University dataset.

- I have followed CRISP-DM (cross industry standard process for data mining)steps.
- As the first step I took some time to understand the data and what data type it is (like continuous, discrete, qualitative , quantitative, semi structured or structured) .
- Then I have done data preprocessing like outlier analysis, zero and non zero variance, dummy variable creation, normalization standardization, transformation and string manipulation .
- Later I have done Exploratory Data Analysis / Descriptive Statistics which included First Moment Business Decision(Measures of Central Tendency), Second Moment Business Decision (Measures of Dispersion), Third Moment Business Decision(Skewness) and Fourth Moment Business Decision(kurtosis).
- Then i did the graphical representations like for Univariate Bar Plot,Index Plot,Also called as Univariate Scatter Diagram,Dot Plot,Strip Plot,Violin Plot,Stem & Leaf Plot,Candle Plot,Pie Chart,Time Series .
- Then followed by supervised learning technique and then trained the model and did model evaluation and then deployed the project.

2. scientific-calculator-
    - Developed the project using python programming language.

3. Emotion analysis of Reviews (Amazon)
    - Extracted the data using library amazon-review-scraper.
    - Used NLP techniques (BOW, ngram, cleanse the data).
    - Emotion Analysis of the data.