1. Introduction

1.1. Project overviews

This project focuses on developing a machine learning solution aimed at predicting the price of natural gas by leveraging historical data and various contextual factors influencing market dynamics. Natural gas prices fluctuate due to a complex interplay of supply-demand dynamics, geopolitical events, weather patterns, and economic indicators. The primary objective is to build a robust predictive model capable of accurately forecasting future price trends. Key stages of the project include comprehensive data collection of historical price data and relevant contextual variables such as supply levels, weather conditions, economic indicators, and geopolitical events. Following data collection, extensive preprocessing ensures the dataset is cleaned, normalized, and prepared for model training.

The next phase involves selecting and training machine learning algorithms suitable for time-series forecasting and regression tasks. These models are evaluated using metrics like mean squared error (MSE) or R-squared to gauge their accuracy in predicting natural gas prices. Scenarios simulating real-world conditions, such as winter demand spikes, supply disruptions, or economic fluctuations, are used to test the model's predictive capabilities under different circumstances. Finally, the trained model is deployed, allowing stakeholders in energy trading, production planning, and risk management to access real-time or near-real-time predictions via an integrated application or platform. This project aims to equip participants with practical skills in data preprocessing, model development, evaluation, and deployment within the challenging domain of energy market forecasting.

1.2. Objectives

- Understand the nature of the problem and determine if it is a regression or classification problem.

- Pre-process and clean data using various data preprocessing techniques.

- Analyze and gain insights from data through visualization.

- Apply different machine learning algorithms based on the dataset characteristics and visualizations.

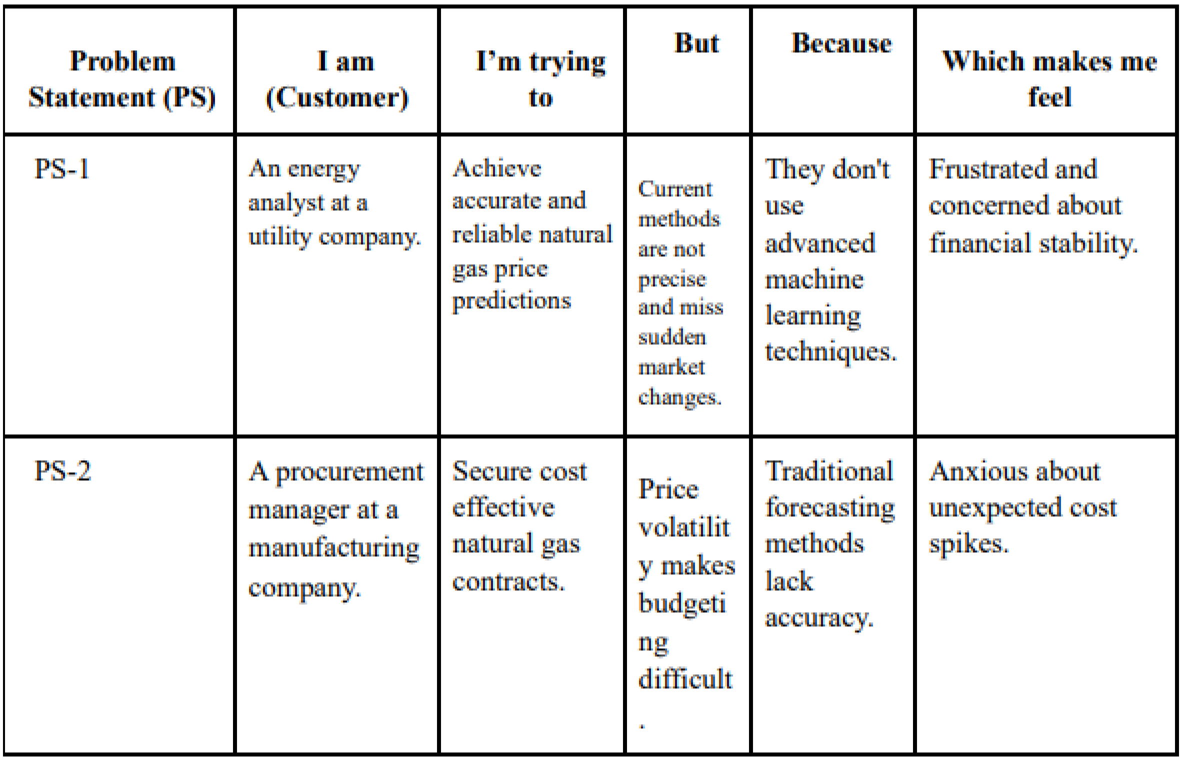
- Evaluate the accuracy of the developed models.

- Build a web application using the Flask framework, providing a user interface for inputting values and showcasing predictions.

2. Project Initialization and Planning Phase

2.1. Define Problem Statement

The project aims to develop a machine learning model capable of accurately predicting the price of natural gas. Natural gas prices are influenced by a complex interplay of factors including supply and demand dynamics, geopolitical events, weather conditions, and economic indicators. The goal is to harness historical price data and relevant contextual information to train a model that can forecast future price movements with precision. This predictive capability is crucial for stakeholders in energy trading, production planning, and risk management who rely on timely and accurate price forecasts to make informed decisions.



Problem Statement (PS)	I am (Customer)	I'm trying to	But	Because	Which makes me feel
PS-1	An energy analyst at a utility company.	Achieve accurate and reliable natural gas price predictions	Current methods are not precise and miss sudden market changes.	They don't use advanced machine learning techniques.	Frustrated and concerned about financial stability.
PS-2	A procurement manager at a manufacturing company.	Secure cost effective natural gas contracts.	Price volatility makes budgeting difficult.	Traditional forecasting methods lack accuracy.	Anxious about unexpected cost spikes.

2.2. Project Proposal (Proposed Solution)

Predicting natural gas prices using machine learning to improve accuracy in forecasting, addressing volatility and market fluctuations

Approach: Utilizing machine learning algorithms such as regression, Train The Model With Descision Tree Algorithm on historical price data, incorporating weather patterns and economic indicators for accurate natural gas price prediction

Key Features:

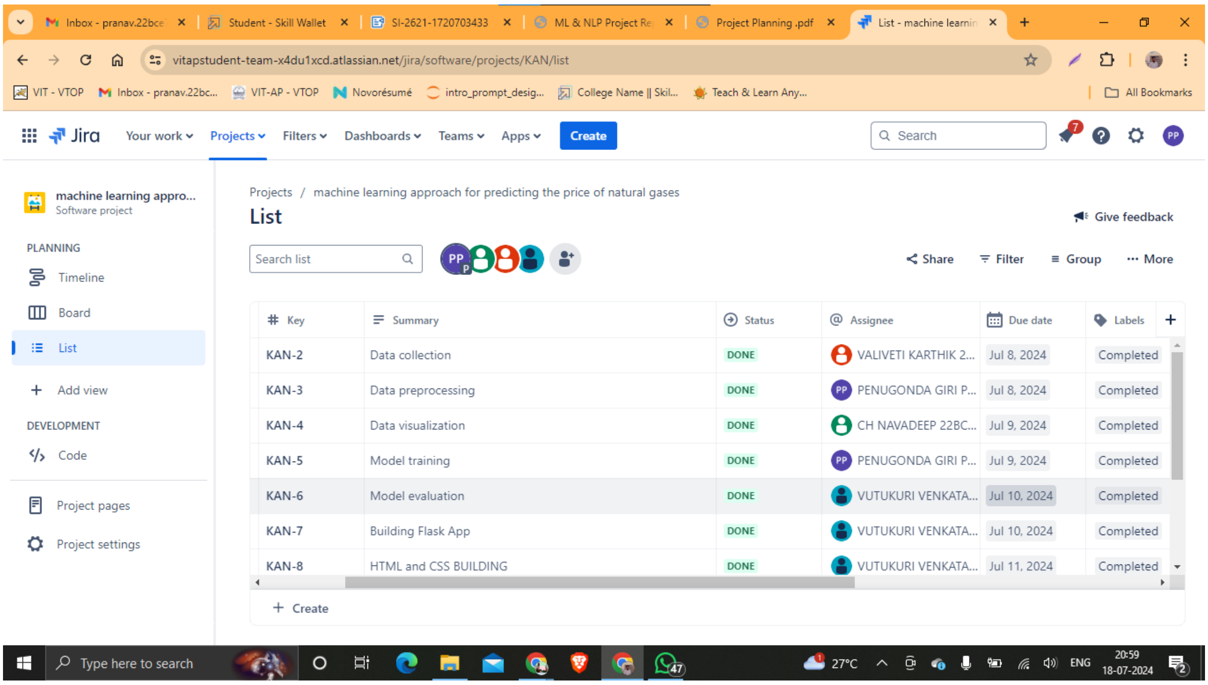
Unique for integrating diverse economic data with advanced machine learning for robust natural gas price forecasting

2.3. Initial Project Planning

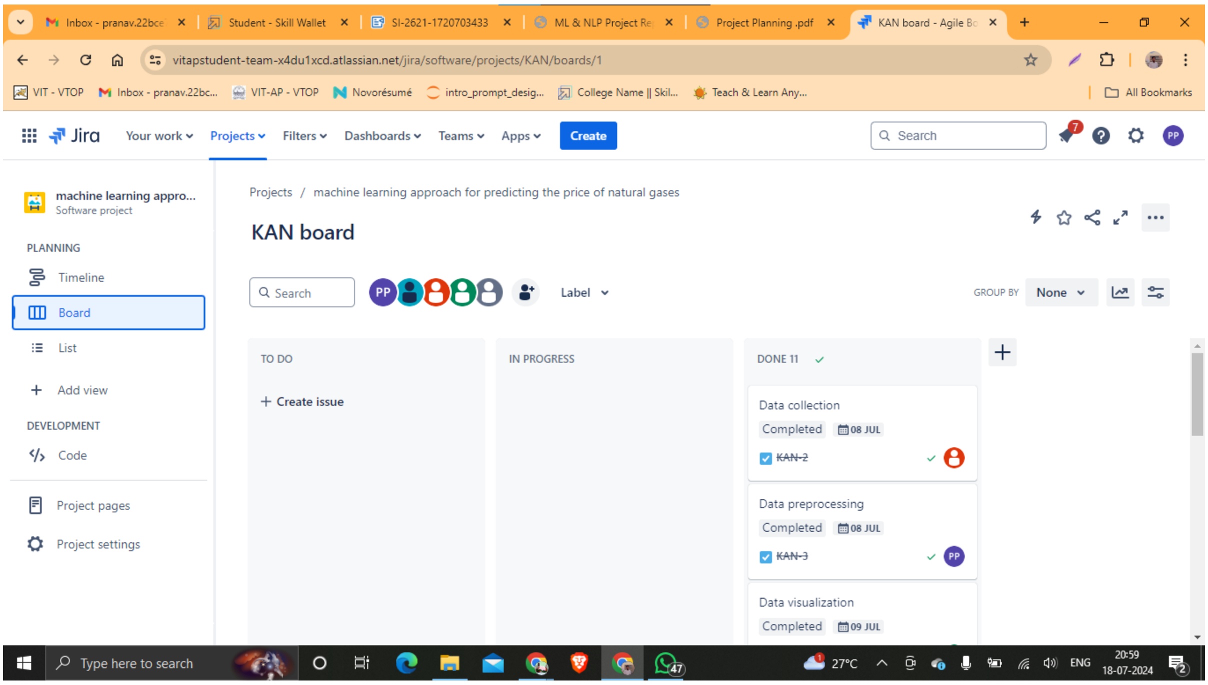
The project was divided into several sprints, each meticulously planned to achieve specific objectives within set timeframes. Each sprint was structured around functional requirements (epics) that defined the overarching goals and tasks to be completed. These tasks were broken down into user stories, each assigned a story point value, priority, team members responsible, and planned start and end dates.

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members	Sprint Start Date	Sprint End Date (Planned)
Sprint-1	Registration	USN-1	Data Collection	2	Medium	Valiveti Karthik	8th July	8th July
Sprint-1		USN-2	Visualising and Analizing Data	1	High	Penugonda Giri Pranav	8th July	9th July
Sprint-2		USN-3	Data Processing	2	Medium	CH Navadeep, Karthik	9th July	10th July
Sprint-1		USN-4	Model Building	2	High	Penugonda Giri Pranav, V.V.N.S.S. NIKHIL	10th July	12th July
Sprint-1	Login	USN-5	Application Building	1	High	V.V.N.S.S. NIKHIL	12th July	14th July

Tracking project using jira software:



Key	Summary	Status	Assignee	Due date	Labels
KAN-2	Data collection	DONE	VALIVETI KARTHIK 2...	Jul 8, 2024	Completed
KAN-3	Data preprocessing	DONE	PENUGONDA GIRI P...	Jul 8, 2024	Completed
KAN-4	Data visualization	DONE	CH NAVADEEP 22BC...	Jul 9, 2024	Completed
KAN-5	Model training	DONE	PENUGONDA GIRI P...	Jul 9, 2024	Completed
KAN-6	Model evaluation	DONE	VUTUKURI VENKATA...	Jul 10, 2024	Completed
KAN-7	Building Flask App	DONE	VUTUKURI VENKATA...	Jul 10, 2024	Completed
KAN-8	HTML and CSS BUILDING	DONE	VUTUKURI VENKATA...	Jul 11, 2024	Completed



TO DO

+ Create issue

IN PROGRESS

DONE 11

Data collection
Completed 08 JUL
~~KAN-2~~

Data preprocessing
Completed 08 JUL
~~KAN-3~~

Data visualization
Completed 09 JUL

3. Data Collection and Preprocessing Phase

3.1. Data Collection Plan and Raw Data Sources Identified

Data will be collected from sources such as e historical price for the natural gas price prediction project. Provided by smart internz platform.

Longitudinal datasets spanning decades, providing historical context and trends in natural gas pricing.

3.2. Data Quality Report

The Data Quality Report Template will summarize data quality issues from the selected source, including severity levels and resolution plans. It will aid in systematically identifying and rectifying data discrepancies.

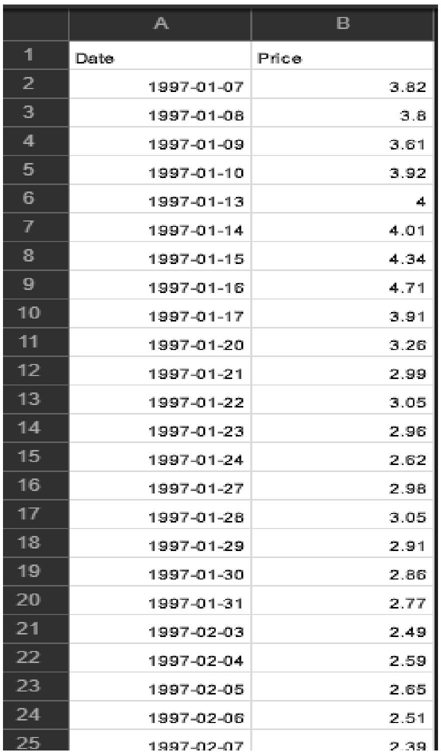
Data Source	Data Quality Issue	Severity	Resolution Plan
Daily csv file	Faced issue with the missing values(NA) in price column	Low	By using python replaced the missing values(NA) with mean of remaining values

3.3. Data Exploration and Preprocessing

Data exploration and preprocessing are foundational steps in the data analysis pipeline, essential for ensuring data quality and preparing datasets for meaningful analysis. Here's a descriptive overview:

Data Overview:

The dataset comprises 5,939 rows and 2 columns, providing a comprehensive view of ecommerce transaction records. Descriptive statistics reveal key insights into the numerical attributes, highlighting central tendencies and dispersion.



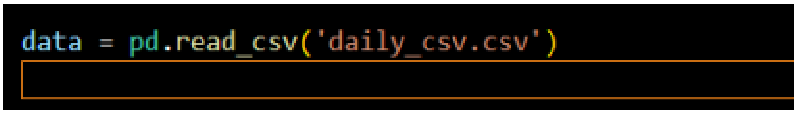
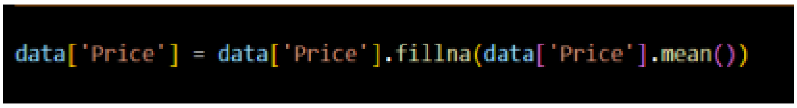
	A	B
1	Date	Price
2	1997-01-07	3.82
3	1997-01-08	3.8
4	1997-01-09	3.61
5	1997-01-10	3.92
6	1997-01-13	4
7	1997-01-14	4.01
8	1997-01-15	4.34
9	1997-01-16	4.71
10	1997-01-17	3.91
11	1997-01-20	3.26
12	1997-01-21	2.99
13	1997-01-22	3.05
14	1997-01-23	2.96
15	1997-01-24	2.62
16	1997-01-27	2.98
17	1997-01-28	3.05
18	1997-01-29	2.91
19	1997-01-30	2.86
20	1997-01-31	2.77
21	1997-02-03	2.49
22	1997-02-04	2.59
23	1997-02-05	2.65
24	1997-02-06	2.51
25	1997-02-07	2.39

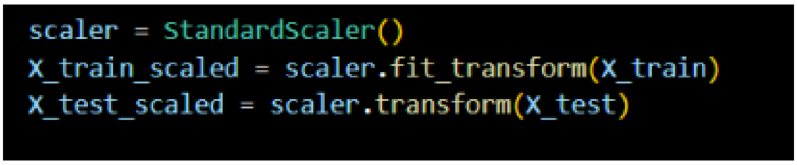
Univariate Analysis :
Average price of natural gas over a specified period. It is used to fill the missing values.

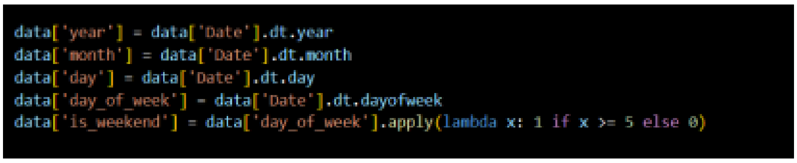
Bivariate Analysis :
Quantifying the strength and direction of the linear relationship between variables. These tools are used to identify how changes in one variable (e.g., economic factors) affect natural gas prices

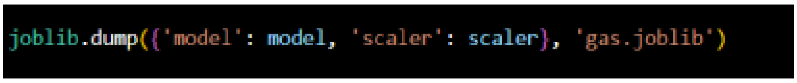
Multivariate Analysis:
Examining how multiple factors like economic indicators, and historical prices interact to influence natural gas prices.

Data Preprocessing:

Loading Data	 `data = pd.read_csv('daily_csv.csv')`
Handling Missing Data	 `data['Price'] = data['Price'].fillna(data['Price'].mean())`

Data Transformation	 `scaler = StandardScaler()`
	`X_train_scaled = scaler.fit_transform(X_train)`
	`X_test_scaled = scaler.transform(X_test)`

Feature Engineering	 `data['year'] = data['Date'].dt.year`
	`data['month'] = data['Date'].dt.month`
	`data['day'] = data['Date'].dt.day`
	`data['day_of_week'] = data['Date'].dt.dayofweek`
	`data['is_weekend'] = data['day_of_week'].apply(lambda x: 1 if x >= 5 else 0)`

Save Processed Data	 `joblib.dump({'model': model, 'scaler': scaler}, 'gas.joblib')`

4. Model Development Phase:

4.1. Feature Selection Report :

Day of the Week: This feature represents the day of the week for each date in the dataset, encoded as an integer from 0 (Monday) to 6 (Sunday)

Is Weekend: This binary feature indicates whether a given date falls on a weekend (Saturday or Sunday)

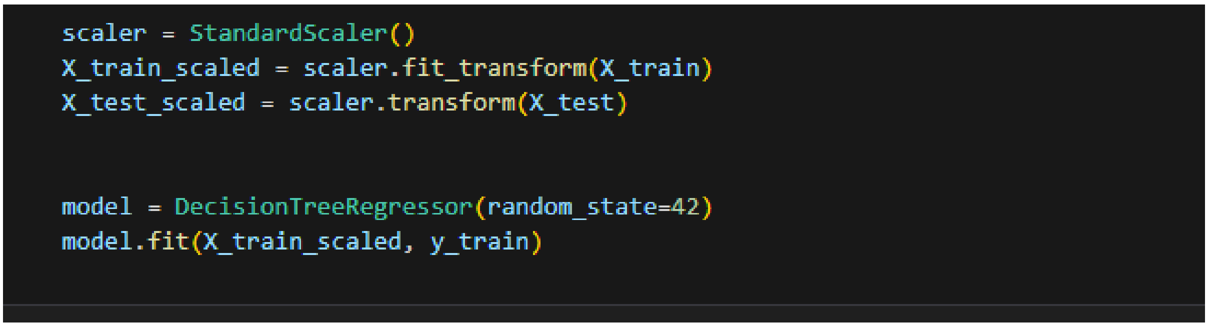
4.2. Model Selection Report:

There are several Machine learning algorithms to be used depending on the data you are going to process such as images, sound, text, and numerical values. The algorithms that you can choose according to the objective that you might have it may be Classification algorithms or Regression algorithms. As the dataset which we are using is a Regression dataset so we can use the following algorithms

Here we considered 4 models for the training

- Linear Regression.
- Logistic Regression.
- Random Forest Regression / Classification.
- Decision Tree Regression / Classification.

4.3. Initial Model Training Code, Model Validation and Evaluation Report



```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)


model = DecisionTreeRegressor(random_state=42)
model.fit(X_train_scaled, y_train)
```

Model Validation and Evaluation Report:

Model	Description	Hyperparameters	Performance Metric (e.g., Accuracy, F1 Score)
Random Forest Regression	Random Forest Regression is an ensemble learning method that constructs multiple decision trees and merges them together to get a more accurate and stable prediction. It handles both regression and classification tasks and improves predictive performance by reducing overfitting.	n_estimators: 100 random_state: 42	Accuracy:91.23
Linear Regression Model	Linear Regression is a simple algorithm used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation. It is effective for understanding and predicting continuous data.	Nill	Accuracy:-89
Decision Tree Model	Decision Tree Regression splits the data into subsets based on feature values, forming a tree-like model. Each split is chosen to minimize the error, making it a straightforward and interpretable model for predicting continuous outcomes.	Max depth=10 random_state=42	Accuracy:96.21
Logistic Regression.	Logistic Regression is used for binary classification problems. It models the probability of a discrete outcome (e.g., high or low prices) by applying a logistic function to a linear combination of the input features	Random state=42	Accuracy:-88

5. Model Optimization and Tuning

Phase:

5.1. Hyperparameter Tuning Documentation :

Model	Tuned Hyperparameters	Optimal Values
Logistic Regression	Regularization strength (C), L1 or L2 regularization.	0.001, 0.01, 0.1, 1.0.
Random forest Regression	Number of trees (n_estimators), max_depth, min_samples_split, min_samples_leaf.	300, 20, 5, 2
Linear Regression	Regularization strength (alpha), Ridge regularization	0.001, 0.01, 0.1, 1.0
Decision Tree	Max_depth, min_samples_split, min_samples_leaf.	2,4,7,9…

5.2. Performance Metrics Comparison Report:

Model	Baseline Metric	Optimized Metric
Random Forest Model	0.857	0.97
Linear Regression Model	0.798	0.89
Decision Tree Model	0.9123	0.9621
Logistic Regression	0.785	0.88

5.3. Final Model Selection Justification:

Selected model: Decision tree regressor

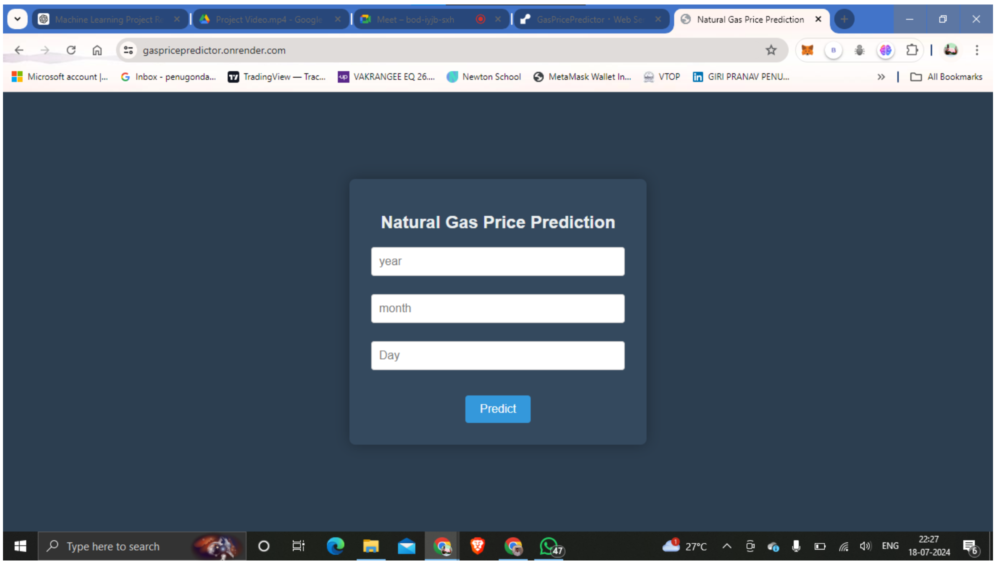
We used the Decision tree regressor because it effectively handles non-linear relationships and interactions between features, reducing overfitting through ensemble averaging. Additionally, it provides feature importance insights and performs well with default hyperparameters, making it a robust choice for our prediction task.factors most significantly influence turbine performance, ensuring accurate and reliable energy forecasts.

6. Results:

6.1. Output Screenshots:

- **Developed webpage:**

Start of the page:



gaspricepredictor.onrender.com

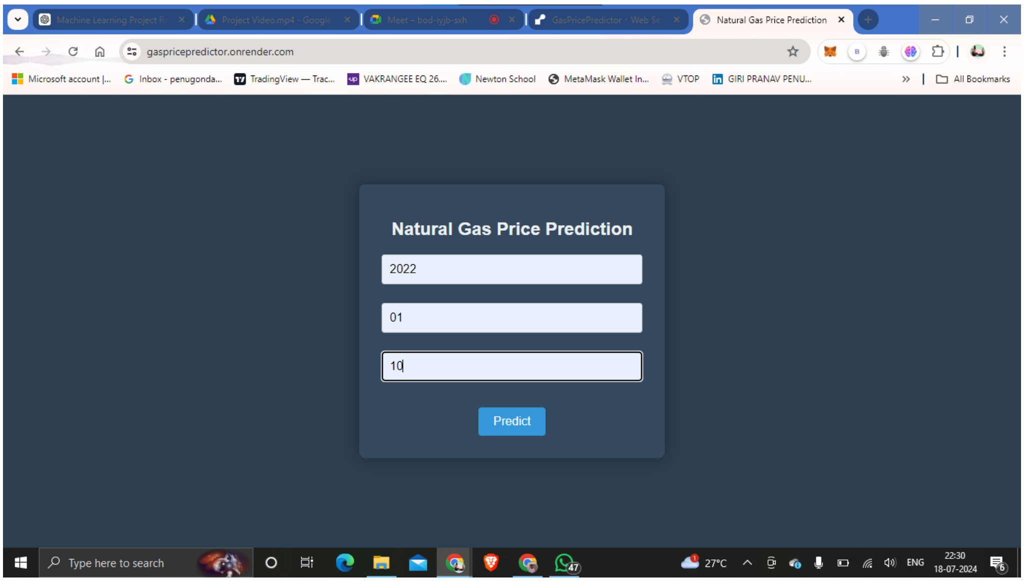
Natural Gas Price Prediction

year

month

Day

Entering Data:



gaspricepredictor.onrender.com

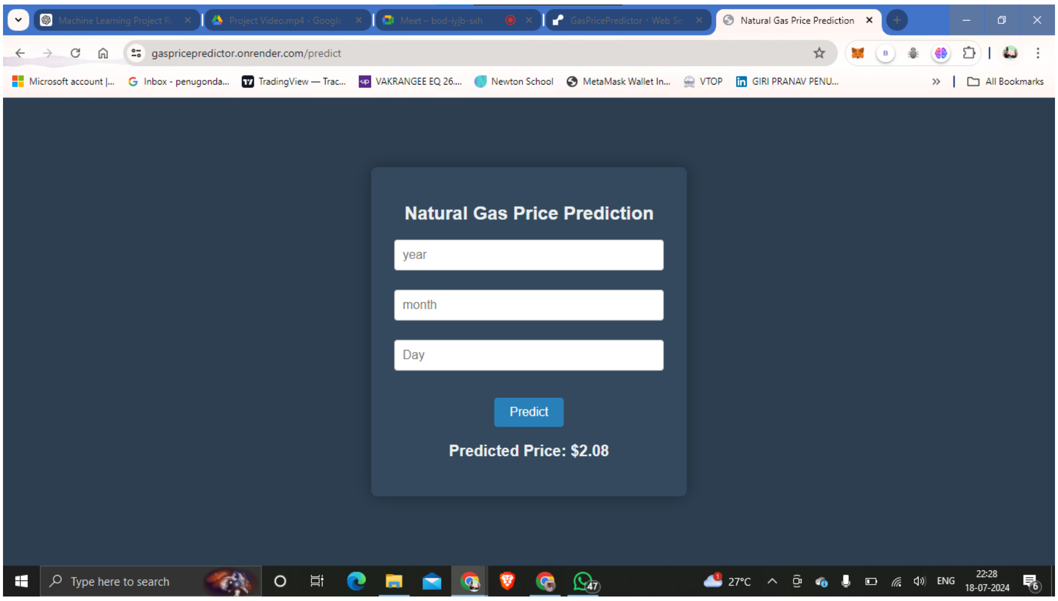
Natural Gas Price Prediction

2022

01

10

After giving Data:



gaspricepredictor.onrender.com/predict

Natural Gas Price Prediction

year

month

Day

Predict

Predicted Price: $2.08

7. Advantages&Disadvantages :

Advantages of the Project

1. **Accurate Price Predictions**: The project enables accurate forecasting of natural gas prices by leveraging machine learning algorithms trained on comprehensive historical and contextual data.
2. **Improved Decision-Making**: Stakeholders in energy trading, production planning, and risk management benefit from timely and reliable predictions, facilitating informed decision-making and strategic planning.
3. **Risk Mitigation**: By anticipating price fluctuations due to factors like supply disruptions, weather changes, and geopolitical events, the project helps mitigate financial risks associated with volatile markets.
4. **Operational Efficiency**: Integration of predictive models into operational processes enhances efficiency in resource allocation, production scheduling, and procurement strategies within the energy sector.
5. **Learning and Adaptation**: Continuous model learning and adaptation to new data allow for ongoing improvement in prediction accuracy over time, adapting to evolving market conditions.

Disadvantages of the Project

1. **Data Complexity**: Managing and integrating diverse datasets with varying formats and quality levels can be challenging, potentially impacting the accuracy and reliability of the predictive models.

Model Interpretability: Some machine learning models, particularly complex ones, may lack interpretability, making it difficult to understand the underlying factors driving specific predictions.
Dependency on External Factors: The accuracy of predictions heavily relies on the availability and quality of external data sources, such as weather forecasts, economic indicators, and geopolitical events.
Overfitting Risks: Overfitting to historical data patterns could limit the model's ability to generalize to new and unforeseen market conditions, requiring careful validation and regularization techniques.
Resource Intensiveness: Training and maintaining machine learning models, especially those handling large datasets, may require significant computational resources and expertise.

8. Conclusion:

In conclusion, this project has been a comprehensive exploration into predicting natural gas prices using machine learning techniques. We started by collecting and preprocessing diverse datasets encompassing historical price data, supply-demand dynamics, geopolitical events, weather conditions, and economic indicators. Through meticulous data preprocessing and feature engineering, we ensured the dataset was robust and ready for model training.

We explored various machine learning algorithms suitable for time-series forecasting and regression tasks, ultimately focusing on models capable of capturing complex relationships inherent in natural gas price fluctuations. Evaluation metrics such as mean squared error (MSE) were employed to assess model accuracy, ensuring our predictive models were reliable and effective.

The project also emphasized the integration of domain knowledge with technical expertise, highlighting the critical role of contextual factors in refining model predictions. Scenarios simulating real-world conditions validated the model's performance under different market dynamics, enhancing its applicability for stakeholders in energy trading, production planning, and risk management.

Finally, deploying the trained models into a practical application or platform allowed stakeholders to access timely and insightful predictions, supporting informed decision-making in the volatile natural gas market. This project not only enhanced our skills in data science and machine learning but also underscored the importance of interdisciplinary knowledge in tackling complex challenges within the energy sector.

9. Future Scope:

The project on predicting natural gas prices using machine learning techniques presents several avenues for future exploration and enhancement:

1. **Enhanced Model Complexity**: Explore advanced machine learning algorithms such as deep learning models (e.g., LSTM networks) to capture intricate temporal dependencies and improve prediction accuracy.
2. **Feature Engineering**: Further refine feature selection and engineering techniques to incorporate additional relevant variables and improve the model's ability to capture nuanced market dynamics.
3. **Ensemble Methods**: Implement ensemble learning approaches to combine predictions from multiple models, potentially enhancing robustness and mitigating individual model weaknesses.
4. **Real-Time Prediction**: Develop capabilities for real-time or near-real-time prediction of natural gas prices, integrating streaming data sources and ensuring rapid response to market changes.
5. **Dynamic Model Updating**: Implement mechanisms for dynamic model updating, allowing the model to continuously learn from new data and adapt to evolving market conditions.
6. **Integration with Decision Support Systems**: Integrate predictive models with decision support systems (DSS) used by energy industry professionals, providing interactive tools for scenario analysis and risk assessment.
7. **Expand Geographical Scope**: Extend the model's applicability to different geographical regions and markets, considering regional variations in supply, demand, and pricing dynamics.
8. **Quantitative Risk Assessment**: Develop methodologies for quantitative risk assessment and scenario planning based on predicted price ranges and volatility estimates.
9. **User Interface Enhancements**: Improve the user interface of the application to enhance usability and provide intuitive visualization of predictions and underlying data trends.
10. **Collaborative Research**: Foster collaborations with academic institutions, industry experts, and regulatory bodies to incorporate cutting-edge research and

domain expertise into the project's framework.

10. Appendix

10.1. Source Code:

•Link: https://github.com/Nikhil-07-star/NaturalGasPricePredictorMachineLearning/tree/main/5.%20Project%20Executable%20Files/Training

10.2. GitHub& Project Demo Link

•GitHubLink: **https://github.com/Nikhil-07-star/NaturalGasPricePredictorMachineLearning**

•Project Demo Link:

https://drive.google.com/file/d/1glFMypwTX8O6ReVr40IR_Tvimrw3hXy6/view?usp=sharing

Website Production Link: **https://gaspricepredictor.onrender.com**