iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37291

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Natural Gas Price Prediction Using Machine Learning

Sanjana G P[1], K S Meghana[2], Lanchana Gudami[3], Meghana M[4], Spandana J[5], Anu C S[6]
[1, 2, 3, 4, 5]*Final year student, Department of CSE, BIET College, Davangere*
[6]*Assistant Professor, Department of CSE, BIET College, Davangere*

Abstract: ***Natural gas varies with season. In addition, natural gas supply, demand, storage, and imports are important indicators related to natural gas price. There are plenty of methods for analyzing and forecasting natural gas prices and machine learning is increasingly used. Machine learning algorithms can learn from historical relationships and trends in the data and make data-driven predictions or decisions. Here a new model for predicting price for natural gas by using Machine Learning concepts. Here some algorithms have been used to build the proposed model: Random Forest Regression, Linear Regression, Decision Tree, Multilinear Regression. By using the algorithm, a Flask model has been implemented and tested. The results have been discussed and a full comparison between algorithms was conducted. Random forest Regression was selected as best algorithm based on accuracy.***

I. INTRODUCTION

A. Machine Learning

The name Machine Learning coined in 1959 by Arthur Samuel. & Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in P, as measured by P, improves with experience E." Machine Learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

B. Machine Learning Methods

Overview of various Machine Learning models are as follows:

1) Supervised learning Supervised learning algorithms are trained using labelled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points labelled either “F” (failed) or “R” (runs). The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabelled data. Supervised learning is commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.

2) Unsupervised learning Unsupervised learning is used against data that has no historical labels. The system is not told the "right answer." The algorithm must figure out what is being shown. The goal is to explore the data and find some structure within. Unsupervised learning works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the main attributes that separate customer segments from each other. Popular techniques include self-organizing maps, nearest-neighbour mapping, k-means clustering and singular value decomposition. These algorithms are used to segment text topics, recommend items and identify data outliers.

3) Semi-supervised learning Semi-supervised learning is used for the same applications as supervised learning. But it uses both labelled and unlabelled data for training – typically a small amount of labelled data with a large amount of unlabelled data (because unlabelled data is less expensive and takes less effort to acquire). This type of learning can be used with methods such as classification, regression and prediction. Semi-supervised learning is useful when the cost associated with labelling is too high to allow for a fully labelled training process. Early examples of this include identifying a person's face on a web cam. 4. Reinforcement learning Reinforcement learning is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with) and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. So, the goal in reinforcement learning is to learn the best policy.

C. *Applications of Machine Learning*

1) *Classification:* Classification or categorization is the process of classifying the objects or instances into a set of predefined classes. The use of machine learning approach makes a classifier system more dynamic. The goal of the ML approach is to build a concise model. This approach is to help to improve the efficiency of a classifier system. Every instance in a data set used by the machine learning and artificial intelligence algorithm is represented using the same set of features. These instances may have a known label; this is called the supervised machine learning algorithm. In contrast, if the labels are known, then its called the unsupervised. These two variations of the machine learning approaches are used for classification problems.
2) *Prediction:* Prediction is the process of saying something based on previous history. It can be weather prediction, traffic prediction, and may more. All sort of forecasts can be done using a machine learning approach. There are several methods like Hidden Markov model can be used for prediction.
3) *Regression:* Regression is another application of machine learning. There are several techniques for regression is available. Suppose, X1, X2, X3…. Xn are the input variables, and Y is the output. During this case, using machine learning technology to provide the output (y) on the idea of the input variables (x). A model is used to precise the connection between numerous parameters as: Y=g(x) Using machine learning approach in regression, the parameters can be optimized.
4) *Image Recognition:* Image Recognition is one of the most significant Machine Learning and artificial intelligence examples. Basically, it is an approach for identifying and detecting a feature or an object in the digital image. Moreover, this technique can be used for further analysis, such as pattern recognition, face detection, face recognition, optical character recognition, and many more. Though several techniques are available, using a machine learning approach for image recognition is preferable. In a machine learning approach for image-recognition is involved extracting the key features from the image and therefore input these features to a machine learning model.
5) *Sentiment Analysis:* Sentiment analysis is another real-time machine learning application. It also refers to opinion mining, sentiment classification, etc. It's a process of determining the attitude or opinion of the speaker or the writer. In other words, it's the process of finding out the emotion from the text. The main concern of sentiment analysis is "what other people think?". Assume that someone writes 'the movie is not so good.' To find out the actual thought or opinion from the text (is it good or bad) is the task of sentiment analysis. This sentiment analysis application can also apply to the further application such as in review-based website, decision-making application. The machine learning approach is a discipline that constructs a system by extracting the knowledge from data. Additionally, this approach can use big data to develop a system. In the machine learning approach, there are two types of learning algorithm supervised and unsupervised. Both of these can be used to sentiment analysis.
6) *Speech Recognition:* Speech recognition is the process of transforming spoken words into text. It is additionally called automatic speech recognition, computer speech recognition, or speech to text. This field is benefited from the advancement of machine learning approach and big data. At present, all commercial purpose speech recognition system uses a machine learning approach to recognize the speech. The speech recognition system using machine learning approach outperforms better than the speech recognition system using a traditional method. Because, in a machine learning approach, the system is trained before it goes for the validation. Basically, the machine learning software of speech recognition works two learning phases:

a) Before the software purchase (train the software in an independent speaker domain)
b) After the user purchases the software (train the software in a speaker dependent domain). This application can also be used for further analysis, i.e., health care domain, educational, and military.

7) *Recommendation for Products and Services:* A Recommender System refers to a system that is capable of predicting the future preference of a set of items for a user, and recommend the top items. One key reason why we need a recommender system in modern society is that people have too much options to use from due to the prevalence of Internet. We can find large scale recommender systems in retail, video on demand, or music streaming. In order to develop and maintain such systems, a company typically needs a group of expensive data scientist and engineers. Machine learning algorithms in recommender systems are typically classified into two categories - content based and collaborative filtering methods although modern recommenders combine both approaches. Content based methods are based on similarity of item attributes and collaborative methods calculate similarity from interactions. Below we discuss mostly collaborative methods enabling users to discover new content dissimilar to items viewed in the past.
8) *Information Retrieval:* The most significant machine learning and AI approach is information retrieval. It is the process of extracting the knowledge or structured data from the unstructured data. Since, now, the availability of information has been grown tremendously for web blogs, website, and social media. Information retrieval plays a vital role in the big data sector. In a machine learning approach, a set of unstructured data is taken for input and therefore extracts the knowledge from the data.
9) *Robot Control:* A machine learning algorithm is used in a variety of robot control system. For instance, recently, several types of research have been working to gain control over stable helicopter flight and helicopter aerobatics. In Darpa-sponsored competition, a robot driving for over one hundred miles within the desert was won by a robot that used machine learning to refine its ability to notice distant objects.
10) *Virtual Personal Assistant:* A virtual personal assistant is the advanced application of machine learning and artificial intelligence. In the machine learning technique, this system acts as follows: machine-learning based system takes input, and processes the input and gives the resultant output. The machine learning approach is important as they act based on the experience. Different virtual personal assistants are smart speakers of Amazon Echo and Google Home, Mobile Apps of Google Allo.

II. SYSTEM DESIGN

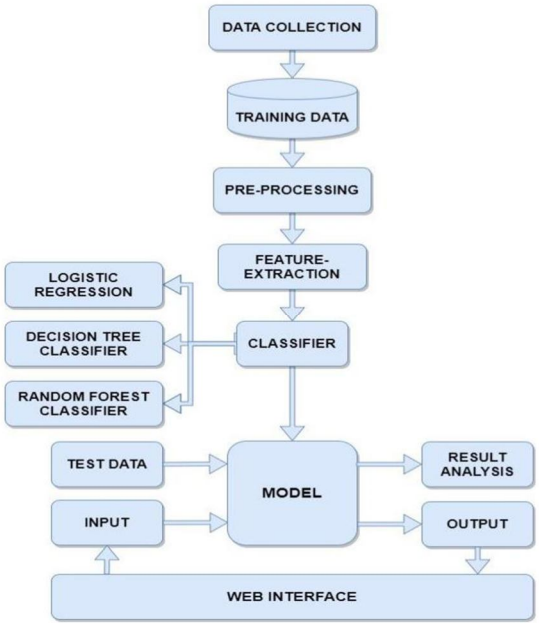


Fig 1. Flow Diagram

A. *Data Collection*

The process of gathering data depends on the type of project, for an ML project, real-time data is used. The data set can be collected from various sources such as a file, database, sensor and other sources and some free data sets from internet can be used. Kaggle and UCI Machine learning Repository are the repositories that are used the most for data collection for Machine learning models.

B. *Pre-processing*

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. There are certain steps executed to convert the data into a small clean data set and make it feasible for analysis, this part of the process is called as data pre-processing. Most of the real-world data is messy, like:

1) Missing Data
2) Noisy Data
3) Inconsistent Data

Some of the basic pre-processing techniques that can be used to convert raw data are:

a) Conversion of Data
b) Ignoring the missing values
c) Filling the missing values
d) Detection of outliers

C. *Feature Extraction*

When the input data to an algorithm is too large to be processed and it is suspected to be redundant then it can be transformed into a reduced set of features. Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data. Feature extraction involves reducing the number of resources required to describe a large set of data. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power, also it may cause a classification algorithm to overfit to training samples and generalize poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy. Many machine learning practitioners believe that properly optimized feature extraction is the key to effective model construction.

D. *Model Selection*

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset. Model selection is a process that can be applied both across different types of models and across models of the same type configured with different model hyper parameters.

The types of classification models are:

1) K-Nearest Neighbor
2) Naive Bayes
3) Decision Trees/Random Forest
4) Support Vector Machine
5) Logistic Regression

E. *Train and Test Data*

For training a model we initially split the model into 2 sections which are ‘Training data’ and ‘Testing data’. The classifier is trained using ‘training data set’, and then tests the performance of classifier on unseen ‘test data set’. Training set: The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. Training data set is used for learning and to fit the parameters of the classifier. Test set: A set of unseen data used only to assess the performance of a fully-specified classifier.

F. *Evaluation*

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents the data and how well the chosen model will work in the future. To improve the model hyper-parameters of the model can be tuned and the accuracy can be improved. Confusion matrix can be used to improve by increasing the number of true positives and true negatives. The output is predicted by analysing the test data as input along with test data output and then the output is displayed.

G. *Interface*

A web interface is built to take input and display an output. Flask language is used to build a web interface and pickle library is used to integrate both model and web page.

H. Dataset Description

We used the dataset that is available from the Kaggle repository

Steps involved in predicting the Natural Gas Price

1) Step 1: Install and import required libraries

```
In [1]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

In [2]:
#importing dataset
dataset=pd.read_csv(r'C:\Users\user\intership\natural-gas-master\natural-gas-master\dat
a\daily.csv')
dataset
```

2) Step 2: Check the information of the data set. It is an object of the class 'pandas.core.frame.DataFrame'. It also shows the datatype of the variables.

```
In [5]:
dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5480 entries, 0 to 5479
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    5480 non-null   object
 1   Price   5479 non-null   float64
dtypes: float64(1), object(1)
memory usage: 85.8+ KB
```

3) Step 3: Converting the columns as required for preprocessing

```
In [6]:
dataset['year'] = pd.DatetimeIndex(dataset['Date']).year
dataset['month'] = pd.DatetimeIndex(dataset['Date']).month
dataset['day'] = pd.DatetimeIndex(dataset['Date']).day
```

4) Step 4: Data Munging Check for missing values in the dataset for each column.

```
In [8]:
#droping the column
dataset.drop('Date', axis=1, inplace=True)

In [9]:
#checking the null values
dataset.isnull().any()
```

5) Step 5: Using matplotlib package for plotting the graph.

```
In [11]:
#handling of missing values
dataset['Price'].fillna(dataset['Price'].mean(),inplace=True)

In [12]:
#data visualization
#import the matplotlib library
import matplotlib.pyplot as plt
#plot size
fig=plt.figure(figsize=(5,5))
plt.scatter(dataset['day'],dataset['Price'],color='blue')
#Set the label for the x-axis.
plt.xlabel('Day')
#Set the label for the y-axis.
plt.ylabel('Price')
#Set a title for the axes.
plt.title('PRICE OF NATURAL GAS ON THE BASIS OF DAYS OF A MONTH')
#Place a legend on the axes.
plt.legend()

In [13]:
import matplotlib.pyplot as plt
plt.bar(dataset['month'],dataset['Price'],color='green')
plt.xlabel('Month')
plt.ylabel('Price')
plt.title('PRICE OF NATURAL GAS ON THE BASIS OF MONTHS OF A YEAR')
plt.legend()
```

6) *Step 6:* Using seaborn to plot the graph on the plane.

```
In [14]:
import seaborn as sns
sns.lineplot(x='year',y='Price',data=dataset,color='red')
```

7) *Step 7:* Label encoding Using sklearn.preprocessing to import Label Encoder and encode the data.

```
In [19]:
from sklearn.preprocessing import LabelEncoder
```

```
In [20]:
lb= LabelEncoder()
```

```
In [21]:
dataset.iloc[:,0]=lb.fit_transform(dataset.iloc[:,0])
dataset
```

8) *Step 8:* Using stats to predict the z-score and to check the threshold .

```
In [22]:
from scipy import stats
z= np.abs(stats.zscore(dataset))
z
```

```
In [23]:
threshold=3
np.where(z>threshold)
Out[23]:
```

9) *Step 9:* Data pre-processing This step fills the missing values of categorical variables with the mode of its respective variable.

```
In [24]:
x=dataset.iloc[:,1:4].values #inputs
y=dataset.iloc[:,-4].values #output price only
```

```
In [25]:
```

10) *Step 10:* This step the model is fit and transformed using Standard Scaler .

```
In [27]:
from sklearn.preprocessing import StandardScaler
```

```
In [28]:
sc= StandardScaler()
x=sc.fit_transform(x)
x
```

11) *Step 11:* The model is dumped into the joblib library and tested.

```
In [29]:
import joblib
joblib.dump(sc,'n1')
Out[29]:
```

12) *Step 12:* The model is trained using the train_test_split library and hence can predict the output.

```
In [30]:
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

13) *Step 13:* Decision Tree Regressor: Apply Decision Tree Regression to the train dataset. Predict the output of test dataset from trained model. Calculate accuracy, precision and confusion matrix.

```
In [33]:
#decision tree regressor
from sklearn.tree import DecisionTreeRegressor
dtr=DecisionTreeRegressor()
#fitting the model or training the model
dtr.fit(x train,y train)

In [34]:
y_pred=dtr.predict(x_test)
y_pred
Out[34]:

In [35]:
from sklearn.metrics import r2_score
accuracy=r2_score(y_test,y_pred)
accuracy

In [36]:
y_test
Out[36]:

In [37]:
y_train
Out[37]:
array([516, 236, 184, ..., 367, 554, 568], dtype=int64)

In [38]:
x_test
Out[38]:
```

14) *Step 14:* Random Forest Classifier: Apply Random Forest Classifier to the train dataset. Predict the output of test dataset from trained model. Calculate accuracy and precision .

```
In [39]:
from sklearn.ensemble import RandomForestRegressor

In [40]:
regressor= RandomForestRegressor()

In [41]:
from sklearn.model_selection import RandomizedSearchCV

In [42]:
#randomized search cv parameters
n_estimators=[int(x) for x in np.linspace(100,1200,12)]
max_depth=[int(x) for x in np.linspace(5,30,6)]
min_samples_leaf=[1,2,5,10]
criterion=['entropy','gini']

In [43]:
#crate a random grid
random_grid={'n_estimators':n_estimators,
             'max_depth':max_depth,
             'min_samples_leaf':min_samples_leaf
             }
random_grid

In [44]:
rf_random= RandomizedSearchCV(estimator=regressor,param_distributions=random_grid,scori
ng='neg_mean_squared_error', n_iter = 10, cv = 5,random_state=0)

In [45]:
rf_random.fit(x_train,y_train)

In [46]:
rf_random.best_params_
Out[46]:

In [47]:
predictions=rf_random.predict(x_test)
predictions
Out[47]:
```

```
In [48]:
from sklearn.metrics import r2_score
r2_score(y_test,predictions)
Out[48]:

In [49]:
import joblib
joblib.dump(rf_random,'n11.save')
Out[49]:
```

I. Algorithm: Decision Tree

Decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and an…

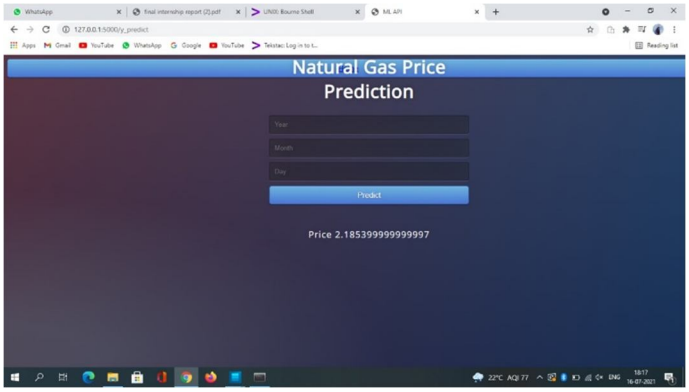
1) *Decision Nodes* – typically represented by squares
2) *Chance Nodes* – typically represented by circles
3) *End Nodes* – typically represented by triangles

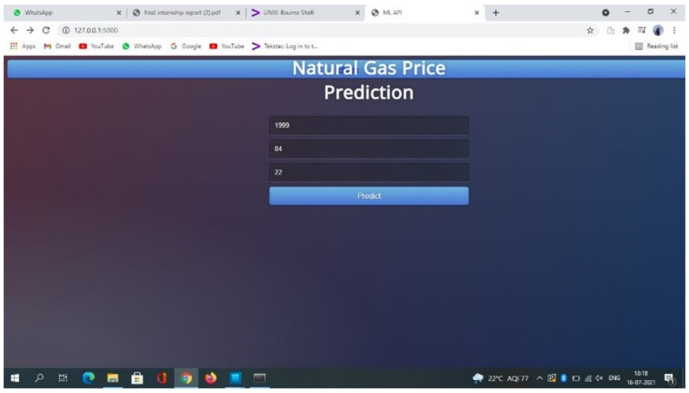
Decision trees are commonly used in operations research and operations management. If, in practice, decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities.

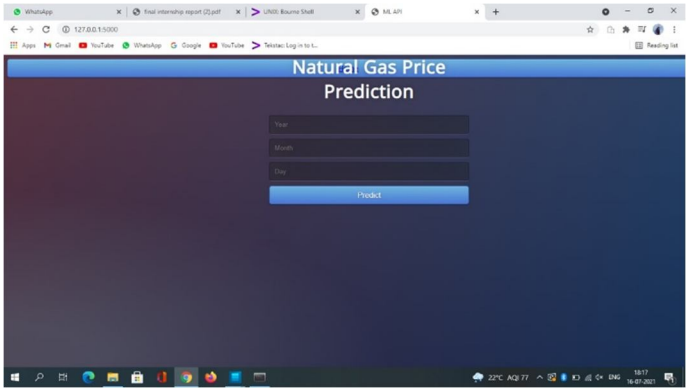
J. Random Forest Regression

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance. The first algorithm for random decision forests was created by Tin Kam Housing the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman[8] and Adele Cutler, who registered "Random Forests" as a trademark (as of 2019, owned by Minitab, Inc.).The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration in packages such as scikit-learn.

III. RESULT





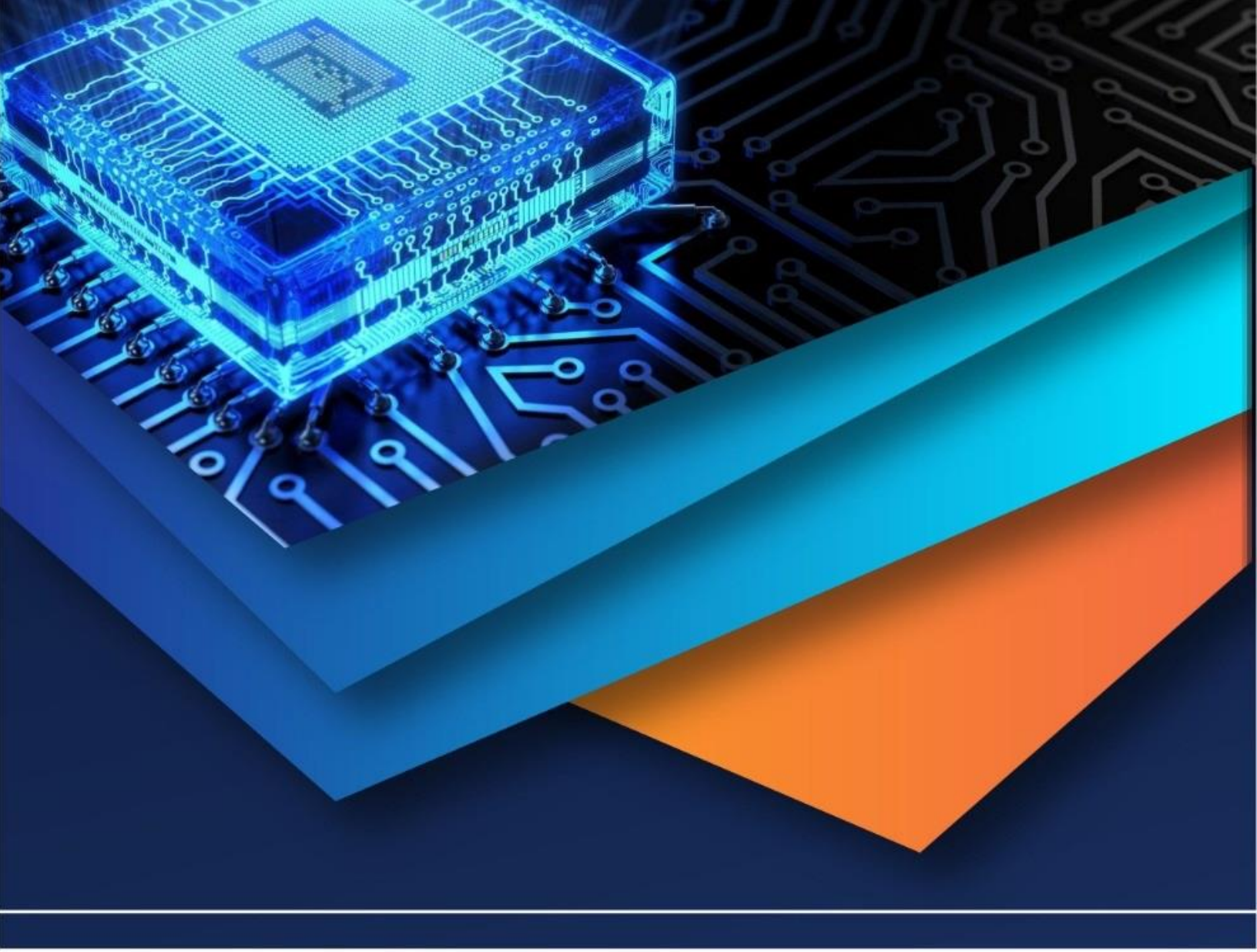


IV. CONCLUSION

The present model has presented the current state in the field of natural gas forecasting. The empirical results demonstrate the prediction methods have decent performance inforecasting natural gas price. It has always been a difficult task to predict the exact daily price of the natural gas. But our model would be convenient and can predict the price of next 12 months also it is cost-effective. The presented algorithms highlighted that although models constitute a wide and efficient choice for addressing price detection also provide a significant boost in increasing the forecasting performance. The main contribution of this work is the development of a new forecasting model for the short-term prediction of natural gas price.

REFERNECES

[1] Natural gas price link stooil market frustrate regulator sefrtsto develop competition". New York Times. 29 October 2008.
[2] Roben Farzad(19April2012)."High Oil Prices Cut the Cost of Natural Gas". Business week.com. Retrieved 15 May 2015.
[3] "Archivedcopy".Archivedfromtheoriginalon1April2017.Retrieved13May2013.
[4] Disavino,Scott; Krishnan,Barani(25September2014)."HenryHub, kingofU.S.natural gas trade, losing crown to Marcellus". Reuters. Retrieved 21 October2014.
[5] Natural Gas Exports". The World Factbook. Central Intelligence Agency. Retrieved 11 June 2015.
[6] ”Background".Naturalgas.org. Archived from the original on 9July2014. Retrieved 14 July 2012.
[7] "Electricity from Natural Gas". Archived from the original on 6 June 2014. Retrieved 10 November 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY