International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

SOLAR POWER GENERATION PREDICTION USING MACHINE LEARNING TECHNIQUES

D.Jeevitha
Department of Computer Science
Jeppiaar Engineering College
Chennai, India

Midumithilash s
Department of Computer Science
Jeppiaar Engineering College
Chennai, India

Vikas Yadav
Department of Computer Science
Jeppiaar Engineering College
Chennai, India

Yathaluru Kireeti
Department of Computer Science
Jeppiaar Engineering College
Chennai, India

Abstract - The solar power plant is based on the conversion of sunlight into electricity. As the use of solar energy has been increased nowadays. Not only we save the electricity with the help of a solar power plant but it also contributes towards the environment. It converts solar energy into electricity either directly using photovoltaics. Nowdays we are using machine learning model.The main necessity of Artificial intelligence is data. The past dataset is collected and that dataset is used to build a machine learning model. The necessary pre-processing techniques are applied like univariate analysis and bivariate analysis are implemented. The data is visualised for better understanding of the features and based on that a classification model is built by using machine learning algorithm and comparison of algorithms are done based on their performance metrics like accuracy MAE,MSE,R2 etc.

***KeyWords*:** *photovoltaics, bivariate analysis,* MAE,MSE,R2,*pre-processing techniques.*

INTRODUCTION

1.1 Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.

The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux. The term “data science” was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.

1.2 Data Scientist:

Data scientists examine which questions need answering and where to find the related data. They have business acumen and analytical skills as well as the ability to mine, clean, and present data. Businesses use data scientists to source, manage, and analyze large amounts of unstructured data.

1.3Natural Language Processing (NLP)

Natural language processing (NLP) allows machines to read and understand human language. A sufficiently powerful natural language processing system would enable natural-language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of natural language processing include information retrieval, text mining, question answering and machine translation. Many current approaches use word co-occurrence frequencies to construct syntactic representations of text. "Keyword spotting" strategies for search are popular and scalable but dumb; a search query for "dog" might only match documents with the literal word "dog" and miss a document with the word "poodle". "Lexical affinity" strategies use the occurrence of words such as "accident" to assess the sentiment of a document.

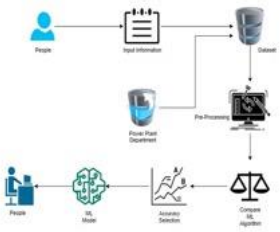

Figure 1 NPL process

1.3 Random Forest Classifier:

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 prediction (regression) of the individual trees. Random decision forests

correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision *trees*, resulting in a *forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.
The following are the basic steps involved in performing the random forest algorithm:

1. HARDWARE REQUIREMENT

- Processor – Intel Xeon E2630 v4 – 10 core processor, 2.2 GHz with Turboboost upto 3.1 GHz. 25 MB Cache
- Motherboard – ASRock EPC612D8A
- RAM – 128 GB DDR4 2133 MHz
- 2 TB Hard Disk (7200 RPM) + 512 GB SSD
- GPU – NVidia TitanX Pascal (12 GB VRAM)
- Intel Heatsink to keep temperature under control
- Storm Trooper Cabinet

2. SOFTWARE REQUIREMENT

- **Python**
  Android Studio is the official integrated development environment for Google's Android operating system, built on JetBrains' Android Studio is the official integrated development environment for Google's Android operating system, built on JetBrains' IntelliJ IDEA software and designed specifically for Android development

- **Anaconda Navigator**.

  The Jupyter Notebook application allows you to create and edit documents that display the input and output of a Python or R language script. Once saved, you can share these files with others. NOTE: Python and R language are included by default, but with customization, Notebook can run several other kernel environments.

4.EXISTING SYSTEM

Anomaly detection relies on individuals' behavior profiling and works by detecting any deviation from the norm. When used for online banking fraud detection, however, it mainly suffers from three disadvantages. First, for an individual, the historical behavior data are often too limited to profile his/her behavior pattern. Second, due to the heterogeneous nature of transaction data, there lacks a uniform treatment of different kinds of attribute values, which becomes a potential barrier for model development and further usage. Third, the transaction data are highly skewed, and it becomes a challenge to utilize the label information effectively. Anomaly detection often suffers from poor generalization ability and a high false alarm rate. We argue that individuals' limited historical data for behavior profiling and the highly skewed nature of fraud data could account for this defect. Since it is straightforward to use information from other similar individuals, measuring similarity itself becomes a great challenge due to heterogeneous attribute values.

5.PROPOSED SYSTEM

Exploratory Data Analysis of Solar power generation Prediction Solar power generations datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

6. LIST OF MODULES

Our proposed system is made up of these following.

Module 1: Data Wrangling

Module 2: Data Collection

Module 3: Data Prediction

MODULE 1-: DATA WRANGLING

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

MODULE 2: DATA COLLECTION

The data set collected for predicting given data is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using machine learning algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done.

MODULE 3: DATA PREDECTION

Anaconda distribution is a free and open-source platform for Python/R programming languages. It can be easily installed on any OS such as Windows, Linux, and MAC OS. It provides more than 1500 Python/R data science packages which are suitable for developing machine learning and deep learning models.

7 .SYSTEM DESIGN:

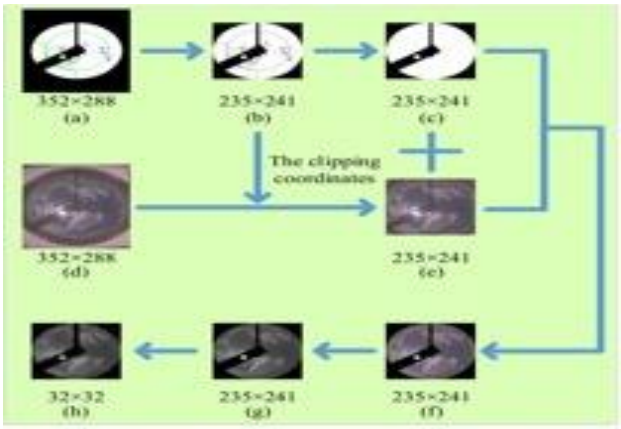

Figure 2 System design

We suggest a deduplication-based scheme for managing heterogeneous data storage. It can be applied in several situations where cloud data deduplication is done by 1) the data owner or 2) a trusted third party. 3) by the cloud's owner or a reliable third party. We use the hash code of data to search for data duplication during cloud storage. The data holder signs the hash code of the data for it to pass CSP's originality verification

Figure 3 output

8.CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out. This application can help to find the Solar power generation.

9.FUTURE SCOPE

- Solar power generation prediction to connect with cloud.
- To optimize the work to implement in Artificial Intelligence environment.

10.REFERENCES

[1] R. Chow et al., "Controlling data in the cloud: outsourcing computation without outsourcing control," in Proc. ACM Workshop Cloud Comput. Secure., 2009, pp.85-90.

[2] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in Proc. 13th ACM Comput. Commun. Secure., 2006, pp. 89-98.

[3] S. Muller, S. Katzenbeisser, and C. Eckert, “Distributed attributebased encryption,” in Proc. 11th Annu. lnt. Conf. lnf. Secure. Crypto., 2008, pp. 20-36.

[4] S. C. Yu, C. Wang, K. Ren, and W. J. Lou, "Attribute-based data sharing with attribute revocation," in Proc. ACM Asia Conf. Comput. Commun. Secure., 2010, pp.261-270.

[5] Dropbox," A file storage and sharing service." [Online]. Available: http://www.dropbox.com/,retrieved March 2017.

[6] Google Drive. [Online]. Available: http://drive.google.com, retrieved May 2017.

[7] Mozy, "Mozy: A file storage and sharing service." [Online]. Available: http://mozy.com/, retrieved May 2017.

[8] J.R. Douceur, A.Adya, W.J.Bolosky, P.Simon, and M.Teimer," Reclaiming space from duplicate files on a serverless distributed file system," in a serverless distributed file system," in Proc.22nd lnt. Conf. Distributed Comput. Syst., 2002, pp. 617-624.

[9] C. Yang, J. Ren, and J. F. Ma, "Provable ownership of the file in deduplication cloud storage," in Proc. IEEE Global Commun. Conf., 2013, pp. 695-700.

[10] C.-l. Fan, S.-Y. Huang, and W.-C. Hsu, "Hybrid data deduplication in a cloud environment," in Proc. Int. Conf. lnf. Secure. Lntell. Control, 2012, pp.174-177.

[11] N. Kaaniche, and M. Laurent, "A secure client-side deduplication scheme in cloud storage environments," in Proc. 6th lnt. Conf. New Technol., Mobility Secure., 2014, pp.1-7.

[12] Z. Yan, M.J. Wang, Y.X. Li, and A. V. Vasilakos," Encrypted data management with deduplication in cloud computing," IEEE Cloud Comput. Mag., Vol.3, no.2, pp.28-35, Mar.-Apr.2016.

[13] Z. Yan, X. Y. Li, M. J. Wang, and A.V. Vasilakos, "Flexible data access control based on trust and reputation in cloud computing," IEEE Trans. Cloud Compt., 2015. doi: 10.1109/TCC.2015.2469662.

[14] J. Hur, D. Koo, Y. Shin, and K. Kang, "Secure data deduplication with a dynamic ownership management in cloud storage," IEEE Trans. Knowl. Data Eng., Vol.28, no.11, pp. 3113-3125, Nov.2016