Republic of Iraq
Ministry of Higher Education and
Scientific Research
University of Babylon
College of Science for Women
Department of Computer Science

Intelligent Prediction System for Maximum Renewable Energy Based on Developed Gradient Boosting Machine (DGBM)

A Thesis

Submitted to the Council of the College of Science for Women at University of Babylon in Partial Fulfillment of the Requirements for the Degree of Master in Science \ Computer Science

By

Zainab Khairallah Al- Janabi

Supervised By

Prof. Dr. Samaher Al-Janabi

2022 A.D. **1444 A.H**

بسم الله الرحمن الرحيم

اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ ۞ خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ ۞ اقْرَأْ وَرَبُّكَ الْأَكْرَمُ ۞ الَّذِي عَلَّمَ بِالْقَلَمِ ۞ عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ ۞

صدق الله العلي العظيم

سورة العلق: الاية [١- ٥]

Dedication

To God Almighty, my Creator, who gave me the will and strength to complete my studies.

To the Prophet of Mercy, Muhammad, and his honorable family, for whom I sought intercession to God, and I found nothing but good.

To the soul of my brother, the martyr "Raad Khairallah Al-Janabi ", who sacrificed himself for the sake of the homeland, and I was accepted into higher studies thanks to his pure and honorable blood.

To my father and son who bore my shortcomings in their service and to all my sisters and brothers whose prayers have not left me.

To all the teachers of the College of Science for Women who gave me science and knowledge and were my second family, especially "Dr. Samaher Hussein Ali", who was a sister who advised me and taught me with kindness and humility.

To the director and employees of the Babylon Statistics Directorate who endured my absence to help them during my study period.

Zainab Khairallah Al- Janabi
2022

Acknowledgments

Praise be to God, who enabled me to complete this message successfully and with great respect for the Great Messenger Muhammad and his pure family.

I would like to extend my thanks and appreciation to my supervisor "**Prof. Dr. Samaher Al-Janabi**" who gave me knowledge, assistance, advice, and encouragement to complete this task.

How much I would also like to thank all my family who gave me support and strength to achieve and progress in my life.

I would also like to thank the Babylon Statistics Directorate for giving me this opportunity to complete my studies, and thanks to all the doctors in the Computer Department, College of Science for Women, for giving me all the requirements for advancement and progress.

Zainab Khairallah Al- Janabi
2022

Abstract

In recent decades, the world has witnessed a great expansion in the world of technology and electronics, in addition to the tremendous development in various industries, which has led to an increase in the need for electrical energy significantly. Renewable energy generated from environmentally friendly sources such as energy (solar, water, windmills, etc.) is the solution. The alternative is to provide that energy, especially as it is clean energy that does not cause the emission of carbon dioxide, which pollutes the air and the environment in general.

This thesis presents a software model for producing the largest amount of energy by developing one of the best prediction techniques and using multi-parameter objective functions. Where the proposed model called ZME-DGBM consists of several stages and addresses many basic challenges during that. The first challenge is software. Although GBM is the best prediction technique, the decision-making part of it is the DT, which makes it somewhat slow and needs to be determined. Number of trees, tree depth, and number of closing nodes in each tree. The proposed model bypasses this problem by replacing the GBM kernel with multi-parameter target functions. As for the second challenge, it is related to the application itself and to include how to produce the largest amount of electrical energy from solar cells with the highest accuracy and the least implementation time.

The proposed model consists of four stages: the stage of data acquisition and preliminary processing, which included: checking it if it contains missing values, calculating the correlation between the characteristics and the target, dividing the data into different periods and deleting the recurring periods. The stage of constructing the predictor, which relied on replacing the GBM kernel with four different multi-parameter functions, as these were the parameters with the highest correlation with the goal, and a threshold value of 0.95 was adopted as determining the importance of the characteristic (parameters). It was reached and both "MSE, RMSE, R^2, " were used.

The proposed model was characterized by giving the best results by using a three-parameter function MPF4 for the GBM Kernel and those

parameters were (AC, TEM, and IRR), where the scale was (R^2= 0.9742), while (MSE=0.0099) and (RMSE= 0.0522), also the system is taken only 80 Ms. to implement on computer Core i5 using the python language 3.8.13.

Supervisors Certification

We certify that this thesis entitled **["Intelligent Prediction System for Maximum Renewable Energy based on Developed Gradient Boosting Machine (DGBM)"]** was done by (Zainab Khairallah Al- Janabi) under my supervision.

Signature:

Name: Prof. Dr Samaher Al-Janabi

Date: / / 2022

Address: University of Babylon/College of Science for Women

The Head of the Department Certification

In view of the available recommendations, I forward the thesis entitled **["Intelligent Prediction System for Maximum Renewable Energy based on Developed Gradient Boosting Machine (DGBM)"]** for debate by the examination committee.

Signature:

Name: Assist. Prof. Dr. Saif Mahmood

Date: / / 2022

Address: University of Babylon/College of Science for Women

Certification of the Examining Committee

We are the members of the examination committee, certify that we have read this thesis entitled (**Intelligent Prediction System for Maximum Renewable Energy based on Developed Gradient Boosting Machine (DGBM)**) and after examining the master student (**Zainab Khairallah Al-Janabi)** in its contents in 2/11/ 2022, and that in our opinion and it is accepted as a thesis for the degree of Master in Science/computer science with degree (Excellent).

Committee Chairman:
Signature:
Name: **Dr.Majid Jabbar Jawad**
Scientific order: Professor
Date: / / 2022

Committee Member:
Signature:
Name: **Dr. Ali Yakoob Yousif Al-Sultan**
Scientific order: Lecturer
Date: / / 2022

Committee Member:
Signature:
Name: **Dr. Karim Hashim Al-Saedi**
Scientific order: Assistant Professor
Date: / / 2022.

Committee Member (Supervisor):
Signature:
Name: **Dr. Samaher Hussein Ali Al-Janabi**
Scientific order: Professor
Date: / / 2022.

Date of examination: 2 / 11 / 2022
Deanship authentication of College of Science for Women
Approved for the College Committee of graduate studies.

Signature:
Name: Dr. **Abeer Fauzi Murad Al-Rubaye**
Scientific order: Professor
Address: Dean of College of Science for Women.
Date: / / 2022.

List of Tables

List of Abbreviation	
Ac	Alternating Current
AC	AC_POWER
AHP	Analytical Hierarchy Process
AM_TEM	AMBIENT_TEMPERATURE
BRT	Binary Regression Trees
CART	Classification and Regression Tree
CI	Consistency Indicator
CNN	Convolutional Neural Networks
CPU	Central Processing Unit
CR	Consistency Ratio
D_YIELD	DAILY_YIELD
DATE_T	DATE_TIME
DC	Direct Current
DMP-DGBM	Deterministic Multi-Parameters based on Developing Gradient Boosting Machine
DNI	Direct Normal Irradiation
DRL	Deep Reinforcement Learning
DT	Decision Tree
EL	Ensemble Learning
ETC	Extra Tree Classifier
FANGBM	Fractional Nonlinear Grey Bernoulli
FPGA	Filed Programming Get Array
GA	Genetic Algorithm
GBM	Gradient Boosting Machine
GHI	Global Horizontal Irradiance
GPU	Graphics Processing Unit
GSA-PS	Hybrid Gravitational Search and Pattern Search algorithm
HM	Hybrid Model
HOMER	Hybrid Optimization of Multiple Electric Renewables
IRR	IRRADIATION
LSTM	Long Short-Term Memory
MAPE	Mean Absolute Percentage Error
MSE	Mean Square Error
MNB	Multinomial Naive Bayes
MO	Meteorological Organization
MOF	Multi-Objective Function

MOPSO	Multi-Objective Particle Swarm Optimization
MPOFs	Multi Parameters Objective Functions
NLP	Natural Language Processing
NN	Neural Networks
P_ID	PLANT_ID
Per	Period
PV	Photovoltaic
R^2	Coefficient of Determination
RER	Renewable Energy Resource
RF	Random Forest
RI	Random Indicator
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
S_KEY	SOURCE_KEY
SVC	Support Vector classifier
T_YIELD	TOTAL_YIELD
TEM	MODULE_TEMPERATURE
TR-HEIC	Total Renewables and Hydro Energy Installed Capacity
WAN	Wide-Area Network
XGBoost	eXtreme Gradient Boosting
ZME-DGBM	Zero to Maximum Energy predictor based on Developing Gradient Boosting Machine

Chapter One:

General Introduction

Chapter One: General Introduction

1.1 Introduction

Due to the technological development that occurred in different areas of life, the world and the environment became more vulnerable to different types of pollution, such as the increase in the emission of carbon dioxide and gases that contain different toxicity rates, especially in industrial areas and dense residential areas, therefore Renewable Energy is the best solution to this problem.

Renewable Energy is a kind of inexhaustible energy that is not finished, it comes from natural resources such as bioenergy, geothermal, hydropower, ocean energy, wind energy, and solar energy[Diezmartínez,2021]. The utilization of renewable energy plays an important role in human life because it satisfies the daily requirements and the multiple of its use in various areas like military, domestic, industrial, and agricultural fields. And can take care of human life since it does not cause air pollution[Vinoth & Pinky,2020]. As a result, in most countries, renewable energy technologies are effective and efficient alternatives for clean and sustainable energy improvement, considering the geographical position of those countries where exhaustive use of most renewable energy resources is essential.

Prediction is a technique that is used to find the values of events or actions that will occur in the future based on the facts or features. The results of the prediction model become true if the predictor is built from facts otherwise the results consider virtual. The techniques of predictions are split into three types: prediction algorithm based on Data Mining, Deep Learning, and Neurocomputing. Data Mining prediction techniques are commonly utilized to aid future decision-making optimization in a variety of sectors[Al-Janabi,2015].

Optimization functions are finding a set of features/ parameters that satisfy the objective function based on all features. There are two types of optimization function, single and multiple objective functions(maximization/ minimization, single parameter/ multi-parameter, with/ without constraints)[Al-Janabi & Mahdi ,2019]. This work will use an objective function that has multi-parameters that represent the parameters most effective in generated Dc-power.

As a result, this thesis will build a predictor to generate maximum Direct current power(Dc-power) through the development of Gradient Boosting Machine(GBM) by replacing its kernel using a multi-parameter objective function to reduce the implementation time and increase the accuracy of prediction results.

1.2 Problem Statement

The problem of this work is divided into parts: The first part is related to programming challenges while; the second part is related to application challenges; as we know the prediction techniques are split based on the scientific field into two fields; prediction techniques related to data mining and predictions related to deep learning techniques; this work deals with the first type of prediction technique.

- One of the data mining prediction techniques is the gradient boosting machine characterized by many features that make it the best. These features (i.e., GBM gives high accuracy results and works with huge data/ stream of data), but on other hand, the core of that algorithm is a decision tree (DT) that has many limitations, it requires choosing the root of the tree, determined the max number of levels of the tree, also it has high computation and long time. Therefore; the first challenge of this thesis is how to avoid these limitations (i.e., high computation and implementation time) of this algorithm and befit from their features.

- The problem of generating electrical energy from environmentally friendly sources with high efficiency is one of the most important challenges in this field; therefore, the second challenge of this thesis is how to avoid these limitations by building an efficient technique to predict max energy from solar energy.

1.2.1 Research Questions

- How is the proposed optimization model Zero to Maximum Energy based on Develop Gradient Boosting Machine (ZME-DGBM) suitable for the increase in the production of electrical power compared to other techniques?
- Can the polynomial optimization function improve the performance of the predictor?
- What are the parameters affecting the generation of the Dc-power?

1.3 Thesis Objectives

The main objectives of this thesis can be summarized in the following points:

- This work attempts to solve one of the limitations related to GBM by suggesting a new kernel of it based on Multi Parameters Objective Functions (MPOFs).
- Build predictor to generate the max Dc-Power from the source of energy-friendly with the environment called solar energy.

1. 4 Literature Survey

The issue of the generation of electrical energy from natural resources is one of the key issues related directly to people's lives and the continued healthful environment. This part of the thesis will try to review the works of previous researchers in the same area and compare our work with it from several points.

Guozhou et. al.,2021[Guozhou et. al.,2021] presented a method to build a dynamic system to optimize energy management in real-time with

considering the uncertain environment, for this hybrid-wind-solar model(HM) was built based on multi targets based on analyzing historical data of daily wind and solar energy through a novel approach of Deep Reinforcement Learning algorithm(DRL). The well-trained agent could be used to reduce the costs by up to 14.17% as compared with other methods. This work is different from our work by using uncertainty states and techniques while similar in considering the same goals.

A.Razmjoo et. al.,2021[A. Razmjoo et. al.,2021] presented a model of hybrid based on gathering data from the meteorological organization(MO) and handled using a software called Hybrid Optimization of Multiple Electric Renewables (HOMER) makes evaluating power system designs for several applications easier. The hybrid architecture of photovoltaic(PV), wind turbine, diesel generator, and battery yielded the greatest results, with an energy cost of 0.151$/kWh and a rate of return of 15.6 percent. Furthermore, the study found that with a greater renewable component of more than 72 percent, this hybrid system may minimize CO2 emissions by more than 2000 kg per family per year. This work is different in relying on multiple components while similar in an attempt to achieve the same goals.

Bahareh et. al.,2021[Bahareh et. al.,2021] presented a method to find the best ordering of renewable energy resources (RER) with few constraints, based on a massive analysis of five constraints ("economic and financial, social, cultural, and behavioral, political and regulatory, technical, and institutional") using analytical hierarchy process(AHP). The result is evaluated by the consistency ratio (CR) which found that solar PV has priority with little constraints than wind and biomass. This work is similar to our work by using the predictions while it is different in the measures and techniques.

Ning et. al.,2021 [Ning et. al.,2021] presented a method to minimize the total cost of setting a micro-grid including the supply and demand technique and

the energy wasted, a model based on analysis of unrealistic data through hybrid gravitational search and pattern search (GSA-PS) algorithm. The result finds that using such technologies has high performance than others and the production cost of the GSA-PS algorithm is greatly less. This work is similar to our work in using hybrid RER, while it is different in using unrealistic data, and different techniques.

Shin'ya et. al.,2021 [Shin'ya et. al.,2021] proposed a model to minimize cost generation by arranging the RER around places connecting and exchanging power through the wide-area network(WAN) to handle current changing, analyzing 7 areas including 42 cities based on the genetic algorithm (GA), using heat storage tank as energy storage instead of battery that is expensive, the result is minimum cost generation and maximum system efficiency evaluated through total cost in addition to economic efficiency. This work is different from our work in the type of dataset used that related to wind while it is similar in using optimization techniques.

Utkucan,2020 [Utkucan,2020] suggested a method for predicting total renewables and hydro energy installed capacity(TR-HEIC) and electricity generation in Turkey from 2019 to 2030, based on a model for analysing data set from 2009 to 2018 through the("fractional nonlinear grey Bernoulli") model(FANGBM). The result predicts that the total renewable and hydro energy installed capacity and electricity generation will minimize from 2019 to 2030, this is evaluated by the mean absolute percentage error (MAPE) and the accuracy(A). This work is similar to our work by using the accuracy measure and prediction model while it is different in using hydro energy and techniques.

Ahmed et. al.,2020 [Ahmed et. al.,2020] presents a mathematical model to evaluate the performance of the many hybrid microgrid forms in Sarawak for the Long San Village based on a Multi-Objective Particle Swarm Optimization (MOPSO)with a regression method. The result is evaluated by the cost and

economic feasibility that has been studied and found that the optimal configuration with the lowest cost of energy and the total cost can be obtained if the installed solar PV is less than 61 kW with 85 kW h of energy storage and 11 kW of hydro generation also points that the dynamic energy cost increases to 0.71 $/kWh when the power generation from RER drops to zero. This work is different from our work in using deterministic and stochastic methods and technologies used, while similar in using a resource of the solar and objective function.

Table 1: compare among the previous works from the seven points including the name of the authors (s), types of renewable energy used in that article, preprocessing archive of the dataset, methodology suggested by the author(s) to solve that problem, main measures used to evaluate results, the main advantages and disadvantages of that methodology.

Table 1: Compare among the previous works

Name of authors	Type of Renewable Energy	Preprocessing	Methodology	Evaluation measures	AdvantageS	Disadvantages
Guozhou et. al.,2021[Guozhou et. al.,2021]	▪ Solar and wind plant.	▪ Analysis of historical data of daily wind and solar energy.	▪ Dynamic system based on a novel Deep Reinforcement Learning algorithm resulted from a combination of deep neural network and the capability of decision making of reinforcement learning.	▪ Cost.	▪ Reduce the costs by up to 14.17% as compared with other methods.	▪ An unrealistic state is considered in the system, therefore any miscommunications between agent and system will fail the work.
A.Ramjoo et. al.,2021[A.Ramjoo et. al.,2021]	▪ Hybrid RE from solar PV plant and wind for electricity production.	▪ Analysis of meteorological organization data using statistical tools.	▪ Hybrid model based on optimization and sensitivity analysis techniques in HOMER program.	▪ Cost.	▪ The hybrid configuration produced the highest results, with an energy cost of 0.151$/kWh and a return on investment of 15.6 percent ▪ Predication that chooses the best pointers can satisfy minimum emission and maximum energy ▪ Provide good predication to invest in the RE field.	▪ High complexity and implementation time.
Bahareh et. al.,2021[Bahareh et. al.,2021]	▪ Solar PV, Wind, Biomass datasets.	▪ Massive analysis of five constraints.	▪ System based on Analytical Hierarchy Process(AHP).	▪ Consistency ratio (CR) which depends on consistency indicator(CI) and random indicator(RI).	▪ Found that the fewer limitations to the development of solar PV.	▪ High complexity needs to determine a huge number of parameters.

Ahmed et. al.,2020 [Ahmed et. al.,2020]	Utkucan,2020 [Utkucan,2020]	Shin'ya et. al.,2021 [Shin'ya et. al.,2021]	Ning et. al.,2021 [Ning et. al.,2021]
▪ Hybrid solar PV and hydro energy in the rural areas of Sarawak.	▪ Hydropower energy.	▪ Solar PV. ▪ Wind.	▪ Hybrid solar PV, and wind.
▪ Sensitively analyzing the different microgrid models, according to electricity price, initial capital cost, emission reduction, and the total cost.	▪ Analysis of historical data from 2009 to 2018 in Turkey.	▪ Analyzing 7 areas using statistical tools.	▪ Probability density functions (PDF) are used to analyze unrealistic data.
▪ A hybrid system based on deterministic and stochastic methods to simulate micro-grid models MOPSOand regression method.	▪ Fractional Nonlinear Grey Bernoulli Model(FANGBM).	▪ Genetic Algorithm (GA).	▪ A hybrid system based on hybrid gravitational search and pattern search (GSA-PS) algorithm.
▪ Total cost and the economic feasibility.	▪ Mean Absolute Percentage Error (MAPE). ▪ Accuracy (A).	▪ Cost.	▪ Total cost and time implementation.
▪ Reduce the cost and found less cost occurs when solar PV is in the range of 61 kW to 85 kW h.	▪ The result shows that the FANGBM produces the highest prediction results.	▪ The RE rate expected by the system is high. ▪ The total cost of the RE is accepted.	▪ It gives high performance. ▪ The production cost of the GSA-PS algorithm is less compared with other algorithms.
▪ An unrealistic state in RE generation is considered in the system. ▪ High complexity.	▪ All grey models suggest reducing the use of hydropower in total renewable electricity generation. ▪ Take high complexity.	▪ High complexity. ▪ Very sensitive to choose fitness function.	▪ Depending on the unrealistic dataset. ▪ Take high complexity.

1.5 Thesis Layout

The remained chapters of this thesis are organized as the following:

- ***Chapter two:*** gives an overview of the main theoretical background in this field.
- ***Chapter three:*** explains the building of Zero to Maximum Energy based on Developing Gradient Boosting Machine (ZME-DGBM) model.
- ***Chapter four:*** illustrates the implementation of the ZME-DGBM model.
- ***Chapter five:*** shows the conclusions of this work together with some recommendations for future work in this field.

Chapter Two:

Theoretical Background

Chapter Two: Theoretical Background

2.1 Introduction

This chapter will present the main theoretical background related to the problem statement that is displayed in chapter one. Then, it will present an analysis of predication algorithms related to Data Mining(DM) and Neurocomputing. Finally, this chapter presents the evaluation measures that will be used in this work including; coefficient of determination (R^2), Mean Square Error(MSE), and Root Mean Square Error(RMSE).

2.2 Prediction Techniques from Side Data Mining

This section shows the main prediction techniques from side data mining and their characteristics as shown in table (2.4).

Recently, machine learning has seen substantial advances in the development of new methods which are considered the most promising approaches in terms of increasing prediction accuracy. Ensemble strategies build a model by training multiple simple models and making a combination between them to perform a more predictive model[Touzani et.al.,2018].

Bagging and boosting are two examples of ensemble techniques used for classification and regression problems. Bagging provides replicate training tuples by sampling with replacements from the training sets while boosting uses all tuples at each iteration but assigns a weight for each tuple in the training set, these weights point to the vector's importance[Luger, 2005]. In bagging, where each classifier was assigned an equal vote, boosting attended to assigns a weight to each classifier's vote, based on the performance of the classifier. Because boosting focuses on the misclassified instances, the resulting composite model suffers from overfitting to such data, on the other hand bagging is less susceptible to such problems, not forget that both can significantly enhance the accuracy

as compared to a single model, boosting tends to achieve maximum accuracy[Han & Kamber, 2006].

2.2.1 Decision Tree

Decision Tree (DT) is a series of logical and mathematical processes based on probabilistic in determining the class label of test records. DT principal works by looking at and comparing the consistency of specific attributes and the threshold node. It makes classification by assigning the class that occurs more frequently to the test records[Hossny et. al.,2020]. DTs are simple to implement and computationally efficient with a large amount of data. They may not effectively generalize the relation between the input parameters and the output which lead to over-fitting problem and as a result lead to weak prediction[Touzani et.al.,2018] [Rustam et. al.,2021]. Since they can feature high depth (i.e., high complexity) [Touzani et.al.,2018].

2.2.2 Random Forest

Random Forest (RF) is used for classification and regression problems. RF is an ensemble model that utilizes bagging techniques which creates many trees and makes voting. Creating a large number of trees will fit prediction accuracy. RF can handle over-fitting. RF is described as the type of the various prediction trees[Rustam et. al.,2021]. RF is more robust in the selection of the training set as compared to the decision tree classifier, RF is hard to interpret, however, its hyper-parameters can be turned with simple[Cotfas et. al.,2021].

2.2.3 eXtreme Gradient Boosting

eXtreme Gradient Boosting (XGBoost) is an advanced version of gradient boosting tree growth to concentrate on computational speed and model efficiency. To fit training data XGBoost assume an initial value always 0.5. It builds its regression tree using the individual tree that fits residuals. A leaf node is used to begin the tree, all residuals are put on a

leaf then computing similarity score, and then compute the gain to specify how to split. Based on these scores and the gain, the XGBoost chooses the largest gain assumed threshold for pruning[Hao, J.,2020]. XGBoost has much popularity as it considered a tree-based model. XGBoost provides a speed boost since trains the number of poor students (i.e, DT) parallelly, different from gradient that does this sequentially. XGBoost can handle over-fittings, which is unfeasible with Gradient Boosting also Adaboost classifiers. XGBoost can be implemented on a distributed system and also can process larger datasets, as a result, it has scalability features. It helps to reduce the loss and enhance accuracy by utilizing a Log Loss function[Rustam et. al.,2021].

2.2.4 Extra Tree Classifier

Extra Tree Classifier (ETC) is an ensemble learning model same as RF. It increases prediction accuracy by using the meta-estimator, which is trains on different samples of the dataset a large number of weak classifiers (i.e. DT). The way of building a tree is different between ETC and RF. ETC creates DT using the original training sample, whereas RF uses samples from the entire dataset. At each iteration, the tree is given a random sample of attributes from the dataset on every test node. Based upon Gini Index criteria and the optimal feature must be determined by DT to partition data. Several de-correlated DTs are created because of providing random sample[Rustam et. al.,2021].

2.2.5 Gradient Boosting Machine

Gradient Boosting Machine (GBM) is a strong excellent prediction technique that reduces error predictions since it uses the concept of boosting. The machine learning field, containing deep learning, and Ensemble Learning (EL). Ensemble Learning is a method to find a good predictor or classifier based on combining weak predictors or classifiers. Bagging and boosting are types of ensemble models and general strategies

to increase the final accuracy. At each iteration, bagging reselects some of the training set more than once to build a model on it while some are not selected[Dogan & Birant, 2020]. While the boosting algorithms are built upon a constructive iterative strategy. The GBM aims at finding an additive model that minimizes the loss function. Thus, the GBM algorithm iteratively adds at each step a new DT that best reduces the loss function. The algorithm starts by initializing the model which is usually a decision tree that maximally reduces the loss function (which is for regression the mean squared error), then at each step a new decision tree is fitted to the current residual and added to the previous model to update the residual. The algorithm continues until a maximum number of iterations[Touzani et.al.,2018].

GBM requires four parameters that have the major effect on its performance and behavior these parameters include; maximum number of trees, learning rate (with a value of fewer than one it is called a Shrinkage rate), maximum number of terminal nodes in the Binary Regression Trees (BRT), and minimum number of data records in a terminal node [Das et. al.,2019].

The final model of GBM is represented by the linear combination of BRT. The performance of the GBM process will be the best if the constructing model performs in stepwise style and that decreases the contribution of each BRT[Al-Janabi,2015].

This work attempts to develop the algorithm for generating the maximum Dc-power. The main steps of GBM are shown in algorithm (2.1).

Algorithm #2.1:GBM *[Al-Janabi,2015]*

Input: *D: dataset, Tr: training data, Tmax: maximum # trees, Sk: Learning rate, Tnmax: # terminal nodes, Smin: # data records in terminal nodes, N: # data records in D, y: index of target.*

Output: *Prediction maximum energy DC*

Initialization: *Fx: Array for predicted values of Tr, rc: rows counter, Tc: trees counter. Org_target: array for the original target of Tr.*

// Find the initial prediction for all data records in Tr by:

Calculate mean of target values Mean(Y).

1: ***While*** *rc < N*
2: *Fx[0,rc] = Mean(Y)*
3: *Org_target[rc] = Tr[y,rc]*
4: *Increase row counter: rc = rc + 1*
5: ***End While***

6: ***While*** *Tc < = Tmax, build boosted T model by:*
7: *rc = 0*
8: ***While*** *rc < N, Update target values of Tr by:*
9: *Residual [rc] = Org_target [rc] – Fx[Tc-1, rc]*
10: *Tr[y,rc] = Residual [rc]*
11: *Increase rows counter: rc = rc + 1*
12: ***End While***
13: ***Call Improved Regression Tree Building (Tr, Tnmax, Smin) and retrieve T.***
14: ***For*** *each terminal node Tn in T:*
15: ***While*** *rc_tn < number of data rows in Tn, update prediction values by:*
16: *Fx [Tc,rc_tn] = Fx [Tc-1, rc_tn] + (Sk * Tn.predictaed_valus)*
17: *Increase counter of data rows in Tn: rc_tn = rc_tn + 1*
18: ***End While***
19: *Increase trees counter: Tc = Tc + 1*
20: ***End For***
21: ***End While***
22: ***Return boosted tree model T model with an array of prediction values Fx.***
23: ***End GBM***

In general; the following example shows the way to apply the GBM.

Table (2.1): Dataset for Persons that have Different Weights, Heights, and Eye Colors

Height	Eye Color	Gender	Target / Original y_i/Weight
1.6	Blue	M	88
1.6	Green	F	76
1.5	Blue	F	56
1.8	Red	M	53
1.5	Green	M	73
1.4	Blue	F	57

Step # 1:

- Create **a Base/Average Model(Fx_0)** for the above dataset by computing the mean for the original values/target.

$$Fx_0 = \frac{1}{n}\sum_{i=1}^{n} y(i)$$

- Set the **Base Model** to 71.2(Fx_0 =71.2 kg) as a **predicted value** for all records.
- Compute the **Residuals** (loss function) values from the difference between the Original $y(i)$ and the **Base Model(Fx_0)**.

$$Residual_1 = y(i) - Fx_0$$

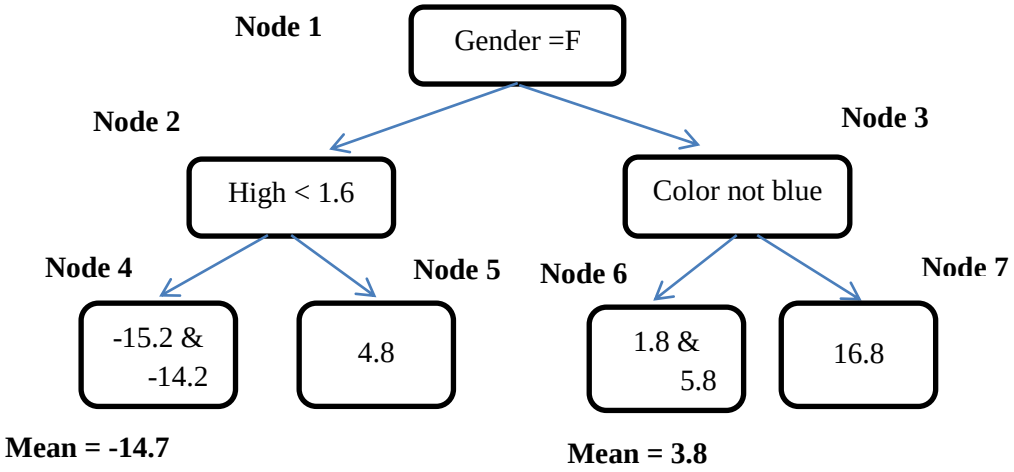
Residual$_1$ = 88 – 71.2= 16.8, 76 – 71.2=4.8,……,-14.2.

Table (2.2): Dataset after applying the GBM

Height	Eye Color	Gender	Weight	Fx_0	Target Residual
1.6	Blue	M	88	71.2	16.8
1.6	Green	F	76	71.2	4.8
1.5	Blue	F	56	71.2	-15.2
1.8	Red	M	53	71.2	1.8
1.5	Green	M	73	71.2	5.8
1.4	blue	F	57	71.2	-14.2

Step # 2:

- Create a second DT model **FX_1** to fit on **Residual$_1$.**
- Compute the mean for the leaf ended with more than one **Residual.**



- Compute the last prediction $\mathbf{Fx_1}$ based on the base model and **Residuals** multiplied by the **learning rate value(α = 0.1).**

$$\mathbf{Fx1} = \mathbf{Fx_0} + (\boldsymbol{\alpha} * \mathbf{Residual_1})$$

$\mathbf{Fx_1}$ for record split by node4 = 71.2 + (0.1 * - 14.7) = 69.73

$\mathbf{Fx_1}$ for record split by node5 = 71.2 + (0.1 * 4.8) = 71.68

$\mathbf{Fx_1}$ for record split by node6 = 71.2 + (0.1 * 3.8) = 71.58

$\mathbf{Fx_1}$ for record split by node7 = 71.2 + (0.1 * 16.8) = 72.88

- Now as in the step before computing $\mathbf{Residual_2}$ based on the original target(**y(i)**) and the last prediction ($\mathbf{Fx_1}$).

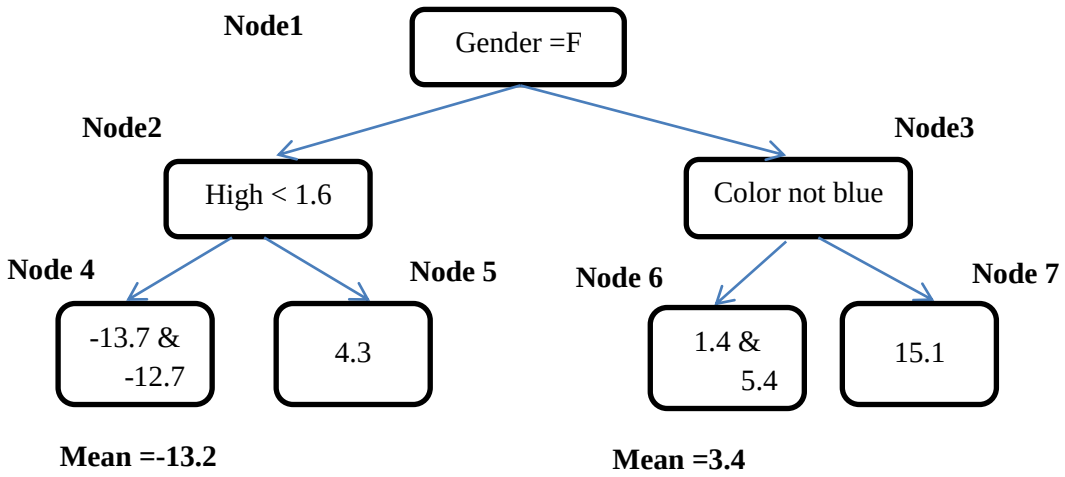
$$\boldsymbol{Residual_2} = \boldsymbol{y(i)} - \mathbf{Fx_1}).$$

Table (2.3): Dataset after applying $\mathbf{Fx_1}$

Height	Eye Color	Gender	Weight	Fx_0	Target: Residual	Fx_1
1.6	Blue	M	88	71.2	15.1	72.88
1.6	Green	F	76	71.2	4.3	71.68
1.5	Blue	F	56	71.2	-13.7	69.73
1.8	Red	M	53	71.2	1.4	71.58
1.5	Green	M	73	71.2	5.4	71.58
1.4	blue	F	57	71.2	12.7	69.73

Step # 3:

- Create a third DT model $\mathbf{FX_2}$ to fit on $\mathbf{Residual_2}$.



- As the steps before repeat compute the **Residuals** and the **Additive Models** until satisfying accurate Predictions or the Residuals do not change.

Table (2.4): Prediction Techniques from Side Data Mining

Techniques	Advantage	Disadvantage
DT [Rustam et. al.,2021].	▪ Simple and easy to build and interpret. ▪ Multi-stage decision-making. ▪ Its performance depends on the way of building on the training set.	▪ Only one feature is examined at each node. ▪ Suitable only for a limited number of features.
RF [Cotfas et. al.,2021] [Rustam et. al.,2021]	▪ Used to solve both classification and regression problems. ▪ Reduces the over-fitting problems. ▪ Robust with selecting training samples. ▪ Provide more accurate results as compared to another classifier.	▪ Hard to interpret, however, its hyper-parameter is an easy turn. ▪ Based on two types of randomization in the selected number of samples and features.

Xgboost [Rustam et. al.,2021].	▪ Used for classification and regression. ▪ Popular and fits the number of DT parallelly, unlike GBM which does this sequentially. ▪ Control over-fitting ▪ Process large datasets. ▪ Scalability feature. ▪ Use log loss function to enhance accuracy.	▪ Steps backward require fixing over-fitting. ▪ The computational time required for large datasets.
ETC [Rustam et. al.,2021].	▪ Higher prediction accuracy by implementing a meta-estimator. ▪ Generate DT from the original training sample. ▪ Same to RF classifier in which both ensemble learning model. ▪ Different from RF in the way of building trees. ▪ Based on math gini index criteria, it selects the best feature to split the data.	▪ Several de-correlated DTs are created because of provided random sample.
GBM [Al-Janabi,2015].	▪ One of the more powerful DM algorithms for prediction. ▪ Uses an additive or average model for boosting the error(loss function). ▪ Uses many weak learner models for each step, the new model tries to minimize the error of the previous model. ▪ The accuracy is the best because its uses a learning rate value (< 1) for reducing the contribution of each BRT.	▪ Work in a sequential style therefore it is slow in processing and analyzing data. ▪ High computation and implementation time.

2.3 Prediction Techniques from Side Neurocomputing

This section shows the main prediction techniques from side neurocomputing and their characteristics as shown in table (2.5).

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks **(**CNN) uses multi-stage to extract many features that could automatically recognize the representation from data. It shows high ability in machine vision techniques and image processing. It can exploit the local or time correlation between data. CNN consists of a set of convolution layers, non-linear processing units, and subsampling layers. CNN is a feed-forward algorithm that has a hierarchical learning model, multiple tasks, and sharing weight. It's lacking interpretation and explanation, not dealing with noise, and may lead to misclassification. It needs a lot of training on data to learn. There are various CNN architectures like LeNet, AlexNet, VGGNet, GoogleNet, ResNet, and ZFNet[Khan et.al.,2020].

2.3.2 Long Short Term Memory

Long Short Term Memory **(**LSTM) is a spatial kind of Recurrent Neural Network (RNN), which results from the back-propagation of errors that is infinite inside the cell, the ability of LSTM to bridge the long temporal interval. It can deal with noise, distributed forms, and continuous values. It can be generalized well, but suffer from problems as same that face the feed-forward, and does not rely on random weight estimation but use small weight initialization[Hochreiter&Schmidhuber,1997][Al-Janabi et. al.,2021].

2.3.3 Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) is a probability classification based on the strong assumption that classifies data whose attributes are autonomous from each other. Although it's simple, this Bayes theorem is quick, accurate, and reliable in a variety of Natural Language

Processing(NLP) classification jobs[Cotfas et. al.,2021]. Handle noise points, and can process missing values during model construction. It is robust to unrelated features and correlated features can damage the performance[Tan,2016] [Ardianto et. al., 2020].

2.3.4 Support Vector Machine

Support Vector Machine (SVM) is a linear supervised learning algorithms model, which gives data points to each object within n-dimension, the variable n denotes the attribute's number. SVM finds the best hyper-plane that separates between the points, it performs a binary classification that suffers a few from over-fitting. Besides, it can perform multiple classifications by combining multi-binary classification functions. Furthermore, it can perform other tasks like regression and outlier detection[Rustam et. al.,2021] [Cotfas et. al.,2021].

2.3.5 Self-Organizing Map

Self-Organizing Map (SOM) employs unsupervised competitive learning to output low-dimensional, discretized descriptions about the given high-dimensional data, and maintains similarity relations among the given data items at the same time. That low-dimensional representation is known as a feature map. SOM is an individual-layer neural network(NN) including units set together with an n-dimensional grid. Hexagonal grids, some three or higher-dimensional spaces are used for multiple applications. SOMs provide low-dimension projection images to distribute high-dimensional data, preserving similar relations between data items, and two-dimensional and rectangular grids and are the most applications used [Miljkovic,2017**]**[Mahdi & Al_Janabi, 2019].

Table (2.5): Prediction Techniques from Side Neurocomputing

Technique	Advantage	Disadvantage
CNN [Khan et.al.,2020]	▪ Using multi-stage extraction features. ▪ Exploit the local or time correlation between data. ▪ Hierarchical learning model. ▪ Multiple tasks, and sharing weights.	▪ Hard to interpret and explain. ▪ Unhanding noise, and may lead to misclassification. ▪ Need a huge of training data to learn. ▪ Hyper-parameter choosing is more sensitive. ▪ Ineffective when used to estimate the object's location, orientation, and pose.
LSTM [Al-Janabi et. al.,2021]	▪ The ability to bridge the long temporal interval. ▪ Can deal with noise, distributed forms, and continuous values. ▪ Can be generalized.	▪ Suffer from some problems that same that appear in the feedforward network. ▪ Do not rely on random weight estimation. ▪ High computation. ▪ Increase the time complexity.
MBC [Cotfas et. al.,2021]	▪ Simple, quick, and accurate in natural language processing classification tasks. ▪ Handle noise. ▪ Robust for unrelated features. ▪ Process missing value.	▪ Correlated features can damage its performance. ▪ Based on the probability principle.
SVM [Rustam et. al.,2021] [Cotfas et. al.,2021]	▪ Binary classification and multi-classification. ▪ Finding the fittest hyper-plane. ▪ Regression and outlier detection task.	▪ Suffer a few from over-fitting. ▪ Based on the try-and-error principle in choosing the main parameters.

SOM [Miljkovic D.,2017]	▪ Discretized descriptions about given high dimensional data, and maintains similarity relations among the given data items at the same time.	▪ Weak prediction can occur if the number of samples is small. ▪ Does not consider the relationship between variables.

2.5 Evaluation Measures

Solar power generation forecasting has recently been developed using a variety of intelligent-based forecasting methodologies. However, several parameters like humidity, ambient temperature, Global Horizontal Irradiance (GHI kW/m 2), Direct Normal Irradiation (DNI kW/m 2), cloud variation, seasonal change, and so on, impact the accuracy of PV output forecasts[Amir & Khan,2021].

$$\boldsymbol{MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{mod,i} - y_{obs,i})^2} \quad \text{...(2.1)}$$

$$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_{mod,i} - y_{obs,i})^2}{N}}} \quad \text{...(2.2)}$$

Where y_{obs} are the original value, and y_{mod} is the value of the model at a time/place i. [Basaran et. al.,2019].

The root of the mean squared error (RMSE) is usually used to measure the difference between the predicted values of the model and the actual values from the system that does being created.

While the coefficient of determination (R^2) shows the percentage variety of prediction values. The rate of the R^2 is between 0 and 1. The measures are described as the following formulas:

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{N}(y_{obs,i} - y_{mod,i})^2}{\sum_{i=1}^{N}(y_{obs,i} - \overline{y_{obs}})^2}} \quad \text{...(2.3)}$$

Chapter Three:

Building Zero to Max Energy Predictor Based on Developing Gradient Boosting Machine (ZME-DGBM)

Chapter Three: Zero to Max Energy Predictor Based on Developing Gradient Boosting Machine (ZME-DGBM)

3.1 Introduction

This chapter presents many sections which show the problem and the way to succeed in manipulating it in the next chapter. Therefore this chapter produces a predictor named ***Zero to Max Energy Predictor Model based on Developing Gradient Boosting Machine(ZME-DGBM***) to predict direct current (i.e., Dc-power) which is the electrical power generated from renewable resources (i.e., solar energy) that does not cause environmental pollution.

ZME-DGBM model consists of many stages applied through a stepwise style; ***the first stage*** presents capturing datasets from scientific cite; that contains the data related to both weather and solar plant.

The second stage is preprocessing which contains multi-steps including; (a) merging between two datasets. (b) splitting readings into intervals and deleting the duplicate (c) applying Pearson correlation to the new dataset.

In the third stage, the ZME-DGBM model is constructed based on developing gradient boosting techniques by replacing its kernel (i.e., Decision Tree function) with multi-parameters optimization functions. The stage begins with dividing the dataset into two sets using five cross-validation methods, the training dataset is used to construct *the* ***ZME-DGBM*** models, while the testing dataset is for evaluating them.

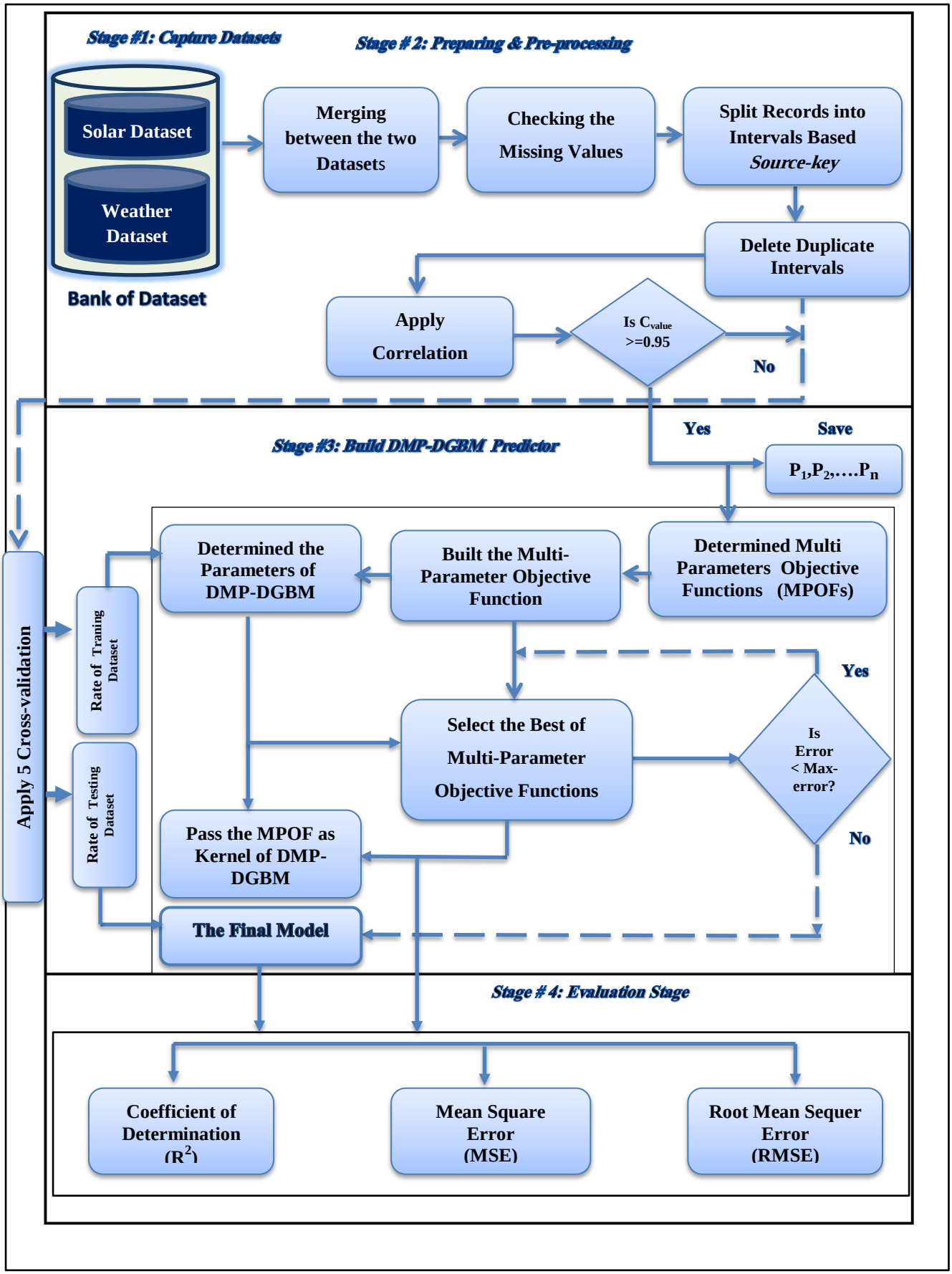
Finally, the results of the ***ZME-DGBM*** are evaluated based on three measures (i.e., coefficient of determination (R^2), Mean Square Error (MSE), and Root Mean Square Error (RMSE).

Figure (3.1) illustrates the sequential stages of the ***ZME- DGBM***, and algorithm (3.1) shows the main stage of it.

We can summarize the main point of this work as follows:

- Capture the datasets from scientific cite; that contains the data related to both weather and solar plant.
- Merging between those datasets, then split the resulting dataset into training and testing based on five cross-validation concepts**.**
- ***ZME-DGBM*** predictor building is based on taking the advantage of MPOFs and using it as a kernel of gradient-boosting techniques to satisfy the goal of this thesis.
- Evaluation of the results of ***ZME-DGBM*** based on three measures.

Also, this chapter presents answers to the questions that are displayed in chapter one through section (1.2.1).



Figure(3.1): Block Diagram of ZME-DGBM

Algorithm#3.1: ZME-DGBM

Input: Solar dataset capture from 6 sensors; Weather dataset capture from 5 sensors

Output: Predict the dc-power generation

// Preparing and Pre-processing stage

1: ***For*** each $feature_i$ in $dataset_1$ // i total number of features in $dataset_1$ i=1,….,6
2: ***For*** each $feature_j$ in $dataset_2$ // j total number of features in $dataset_2$ j=1,….,5
3: Call merge datasets *//Merge based on **Plant-id***
4: ***End for***
5: ***End for***
6: ***For*** i=1 to Nr *// Nr is number of rows in dataset*
7: ***For*** j=1 to Nc *// Nc number of column in dataset*
8: Check missing values
9: Call Split into intervals
10: Call correlation
11: ***End for***
12: ***End for***

// Build DMP-DGBM Predictor

13: ***For*** each id_ interval
14: ***For*** i in range(1: total number of records [id_ interval])
15: Split the dataset into Training and Testing through 5-Cross-Validation
16: ***End for***
17: ***For*** each Training part not used
18: Call DMP-DGBM *//To predict the value of DC-power*
19: ***End for***
20: ***For*** *each Testing part not used*
21: *Test stopping conditions* *//Max number of epoch and max error generation*
22: ***If*** *max error generation< Emax*
23: GO to step 29
24: ***Else***
25: GO to step 17
26: ***End If***
27: ***End for***
28: ***End for***

// Evaluation stage

29: Call Evaluation ZME-DGBM
30: ***End ZME-DGBM***

3.2 Main Stages of ZME-DGBM

This section presents four main stages for building an efficient multi-parameters optimization model.

3.2.1 Data Pre-processing Stage

This section shows the preprocessing steps as explained in the algorithm (3.2):

- Merging the two datasets by seeking the common features between the solar and weather plants and saving them in a single file.
- Checking missing values is applied by dropping any record that has a missing value in each column.
- Splitting data rows into intervals according to the ***Source-key*** feature.
- Execute the Pearson correlation function to find which parameters or features in the dataset most affect determining the Dc-power.

Algorithm#3.2: Pre-processing

Input: Two datasets: solar plant contained features captures from six sensors; weather dataset contained features captures from five sensors

Output: Dataset after preprocessing

// Merging both datasets based on (Plant_id)

1: ***For*** each feature$_i$ in dataset$_1$ // i total number of features in dataset$_1$ i=1,….,6

2: ***For*** each feature**j** in dataset$_2$ // j total number of features in dataset$_2$ j=1,….,5

3: Merge datasets *//Merge based on* ***Plant-id***

4: ***End for***

5: ***End for***

// Checking missing values

6: ***For*** each sample i // i=1..m

7: ***For*** each column j // j=1..n

8: **If** a[j] has a missing value

9: Dropping a[i]

10: ***End for***

11: ***End for***

// Split dataset into multi intervals

Build empty set called Q

For i=1 to the total number of intervals // split on ***Source-key***

    W[i] = read samples

End for

For i=1 to the total number of intervals

    If *w[i]= w[i+1]*

        Delete w[i+1] // delete duplications

    Else

        Q[i]= w[i]

    End if

End for

// Apply correlation

For each sample i

    For each column j

        Compute Correlation $// \boldsymbol{R_{j,j+1}} = \frac{\boldsymbol{covr(j,j+1)}}{\boldsymbol{\sigma_j \sigma_{j+1}}} = \frac{\boldsymbol{Exp(j-\bar{j})(j+1-\overline{j+1})}}{\boldsymbol{\sigma_j \sigma_{j+1}}}$

    End for

End for

End preprocessing

Where R is the Pearson Correlation, the covr is the covariance between columns j and j+1, σ_j is the standard deviation of j, σ_{j+1} and is the standard deviation of j+1, $\bar{j}$ is the average of j, $\overline{j+1}$ is the average of $j+1$, and Exp is the expected value.

3.2.2 Deterministic Multi-parameters Developing GBM Predictor

This stage considers the kernel of the Deterministic Multi-Parameter Objective Function based on the Developing Gradient Boosting Machine (DMP-DGBM) by taking the advantage of multi-parameters objective functions and making it the core of the GBM rather than DT, by applying

four equations; three of these equations have two parameters while the fourth equation has three parameters.

3.2.2.1 Determine the Multi Parameters Objective Functions (MPOFs)

The procedure of **MPOFs** is built to determine the best multi-parameter functions (MPFs) that contain the parameters that have a high correlation with the target.

As a result, choosing the best function that satisfies the maximum prediction and passing it to become a kernel of DMP-DGBM rather than DT. The MPOFs has many steps as illustrated in the steps below and shown in the procedure (MPOFs):

- Apply many thresholds after computing the correlation step, finding that the parameters which have a correlation value with the target much or equal to 0.95 in the dataset are the most affected parameters that enhance the prediction approach.
- Determine the significant parameters (i.e, AC, IRR, and TEM) that affect the decision taken from the predictor.
- Compute the linear combination function to the parameters that are high or equal to the threshold computed in the first step as follows: If Corr (P_i, Target) >= 0.95 Then

$$MPOFs = \sum_{i=1}^{n} Max\ (P_i) \qquad \ldots(3.1)$$

- Apply the equation of multi-parameter functions to the parameters sequentially until getting four MPFs that combine all the significant parameters in one function.
- Find the best MPFs from the four functions.
- Set that function as the kernel of the DMP-DGBM algorithm.

Procedure Multi Parameters Objective Functions(MPOFs)

```
For i=1 to Nr                                   // nr is number of rows in dataset
    For j=1 to Nc                               // nc number of column in dataset
        If correlation-value-target with j >= 0.95 then
            X= a[i,j]
            Zk= Linear combination X            // Zk is MPOFs become as new kernel of DGBM
            // Multi-objective functions of two parameters (AC & IRR)
            For i=1 to Nr
                MPF1= max(AC[i] + IRR[i])
            End for
            Return MPF1
            // Multi- objective functions of two parameters (AC & TEM)
            For i=1 to Nr
                MPF2= max(AC[i] + TEM[i])
            End for
            Return MPF2
            // Multi-objective functions of two parameters (TEM & IRR)
            For i=1 to Nr
                MPF3= max(IRR[i] + TEM[i])
            End for
            Return MPF3
            // Multi-objective functions of three parameters (TEM, AC & IRR)
            For i=1 to Nr
                MPF4= max(AC[i] + IRR[i]+ TEM[i])
            End for
            Return MPF4
        End if
    End for
End for
End MPOFs
```

3.2.2.2 Build DMP-DGBM Predictor

The details of the traditional GBM are explained in the second chapter, section (2.4.6).

The steps of DMP-DGBM are given in algorithm (3.3):

Algorithm #3.3: DMP-DGBM

Input: *D: dataset after preprocessing*

Output: *Predict max DC-power*

Initialization: *Fx: array for predicted values, rc: rows counter, C_{MPOF}: counter of multi parameters optimization functions, MPOFs_MAX: max number of MPOFs, Org_target: array for original target, y: index of target, N: # records in D.*

For i=1 to Tr *// Tr is total number of sample in training dataset*

// Find the initial prediction for all data records in Tr by:

Calculate mean of target values Mean (Dc-power)

While *rc < N*

Fx[0,rc] = Mean(Dc-power)

Org_target[rc] = Tr[y,rc]

Increase row counter: rc = rc + 1

End while

Call MPOFs

While C_{MPOFs} *< = MPOFs_MAX, build optimization model by:*

rc = 0

While *rc < N, Update target values of Tr by:*

Residual [rc] = Org_target [rc] – Fx[C_{MPOFs} -1, ***MPOFs****, rc]*

Tr[y,rc] = Residual [rc]

Increase rows counter: rc = rc + 1

End while

For each function of ***MPOFs:***

While *rc_ MPOFs < # rows in MPOFs, update prediction values by:*

Fx[C_{MPOFs} -1, MPOFs, rc]=

Fx[C_{MPOFs} -1, ***MPOFs****, rc]+ (Sk * MPOFs. predictaed_valus)*

Increase counter of data rows in final MPOFs:

rc_ MPOFs = rc_ MPOFs + 1

End while

Increase *MPOFs* ***counter:*** $C_{MPOFs} = C_{MPOFs} + 1$

End for

End while

End for

Return DMP-DGBM model with array of prediction values Fx.

// Build the testing stages of DMP-DGBM model.

End DMP-DGBM

3.2.4 Evaluation Measures

This section presents the algorithm to compute the four evaluation measures used to determine the performance of predictor DMP-DGBM.

Algorithm# 3.4: Evaluation Measures
Input: Training and testing datasets //data file result from merging step
Initialization: R^2=0, MSE=0, RMSE=0

// Calculate the mean for both actual and predicted values
1: *Compute mean* // $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y(i)$
// Compute a coefficient of determination (R^2)
2: ***For*** each row in the testing dataset
3: ***For*** each column in the testing dataset
4: *Compute R^2* //As equation(2.1)
5: ***End for***
6: ***End for***
7: *R^2=sum all values of R^2*
// Compute Mean Square Error (MSE)
8: ***For*** each row in the testing dataset
9: ***For*** each column in the testing dataset
10: *Compute MSE* //as equation(2.2)
11: ***End for***
12: ***End for***
13: *ME=sum all values of ME*
// Compute Root Mean Square Error (RMSE)
14: ***For*** each row in the testing dataset
15: ***For*** each column in the testing dataset
16: Compute *RMSE* //as equation(2.3)
17: ***End for***
18: ***End for***
19: *RMSE=sum all values of RMSE*
20: ***For*** each row in the testing dataset
21: ***For*** each column in the testing dataset
22: R^2(actual, predicted):
23: MSE (actual, predicted):
24: RMSE (actual, predicted):
25: ***End for***
26: ***End for***
End **Evaluation**

Where $\bar{Y}$ is the average of the actual value y(i), and n is the number of rows in the dataset.

3.5 Summary

This section will explain how the ***ZME-DGBM*** finds an answer to the questions displayed in the first chapter.

- ***How is the ZME-DGBM suitable for the increase in the production of electrical power compared to other techniques?***

The proposed model relied on the substitution of the core for one of the best prediction techniques represented by the gradient boosting machine by a multi-parameter objective function. The parameters of these functions

are highly correlated with the target(i.e, Dc-power). In general; we determined the threshold of maximum affected correlation which is high or equal to 0.95; the value of the threshold is determined based on the try-and-error principle.

The hybrid model is a combination between MPOFs and GBM that leads to results that are characterized by high accuracy.

- ***Can the polynomial optimization function improve the performance of the predictor?***

Yes, the polynomial optimization function was able to improve the performance of the gradient boosting algorithm because it took into account only the most important parameters in power generation. In addition, it was able to avoid the problems of the algorithm, including; determining the number of decision trees, the depth of each tree, and the long time of training it.

- ***What are the parameters affecting the generation of the Dc-power?***

By calculating the correlation between the values obtained from the sensors, it was found that IRR (i.e., Irradiation), Ac (i.e., Ac-power), and TEM (i.e., Temperature) are the most affecting parameters in generating the Dc-power.

Chapter Four:

Implementation (ZME-DGBM)

Chapter Four: Implementation and Results of *(ZME-DGBM)*

4.1 Introduction

This chapter shows the main parameters used in the proposed algorithm and the implementation of the main stages related to designing the model, then gives justification for the results of each stage in that model**.**

4.2 Implementation and Results of ZME-DGBM

This section presents the outcomes for each stage in ***ZME-DGBM*** and presents the main proof of each result, then shows the benefit from the two sides; programming and application.

4.2.1 Description of Datasets

This thesis deals with two types of datasets, the first dataset contains five features including;("Plant_Id", "Source_Key", "Ambient_Temperature", "Module_Temperature", and "Irradiation") have 3183 samples. The second dataset is solar plant involving 6 features;("Plant_Id", "Source_Key", "Dc_Power", "Ac_Power", "Daily_Yield", and "Total_Yield") having 71266 samples.

4.2.2 Collection of Dataset

In this stage, the datasets have been captured from a scientific website that contains data related to both weather and solar plant, each one of them has important parameters related to the thesis target for building and testing the proposed model.

Tables (4.1) and (4.2) show a sample of both datasets and the features that belong to each one.

Table (4.1): Sample of Solar Dataset

T_YIELD	D_YIELD	AC	DC-power	S_KEY	P_ID
7247045.5	5021.5	323.825	3301	zVJPv84UY57bAof	4135001
6382879	4948	451.3714286	4603.428571	1BY6WEcLGh8j5v7	4135001
6316818.286	5386.285714	535.9142857	5466.857143	1IF53ai7Xc0U56Y	4135001
7120640	5336	536.4571429	5472.428571	3PZuoBAID5Wc2HD	4135001
7733013.571	5192.571429	491.7	5014.857143	7JYdWkrLSPkdwr4	4135001
7292147	5387	471.6	4809	McdE0feGgRqW7Ca	4135001
7339696.429	5291.428571	489.5285714	4992.285714	VHMLBKoKgIrUVDU	4135001
7159078	5212	521.1285714	5315.428571	WRmjgnKYAwPKWDb	4135001
7310908.857	5169.857143	489.0428571	4987.571429	YxYtjZvoooNbGkE	4135001
6654833.571	5308.571429	489.4428571	4991.857143	ZnxXDlPa8U1GXgE	4135001

Table (4.2): Sample of Weather Dataset

IRR	TEM	AM_TEM	S_KEY	P_ID
0.434130931	43.36905527	26.98462527	HmiyD2TTLFNqkNe	4135001
0.537777062	43.0649942	27.53423467	HmiyD2TTLFNqkNe	4135001
0.583538421	49.00184933	27.85342613	HmiyD2TTLFNqkNe	4135001
0.723039185	49.10859787	28.36791987	HmiyD2TTLFNqkNe	4135001
0.92426021	54.62204787	29.10705953	HmiyD2TTLFNqkNe	4135001
0.931341062	56.35477893	29.4410646	HmiyD2TTLFNqkNe	4135001
1.016641412	58.59736313	29.58492027	HmiyD2TTLFNqkNe	4135001
0.494514807	55.35197014	29.76473564	HmiyD2TTLFNqkNe	4135001
1.089475536	54.12078233	29.81758227	HmiyD2TTLFNqkNe	4135001
0.371274763	52.61588127	29.74218547	HmiyD2TTLFNqkNe	4135001

4.2.3 Results of the Pre-Processing Stage

This stage includes multi-Handel to extract useful information from the next stage called preprocessing.

Step #1: Apply Merging

This step makes integration between both datasets based on the ***(Plant_id*)** features and gets 10 parameters instead of 11 as shown in table (4.3).

The main purpose of the merging step is to improve the robustness of the results and make them more understandable.

Table (4.3): The Results of Merging the Solar and Weather Dataset

P_ID	AC	DC-power	S_KEY	IRR	TEM	AM_TEM	S_KEY	T_YIELD	D_YIELD
4135001	323.825	3301.000	1BY6WEcLGh8j5v7	0.434	43.369	26.985	HmiyD2TTLFNqkNe	7247045.500	5021.500
4135001	451.371	4603.429	1IF53ai7Xc0U56Y	0.538	43.065	27.534	HmiyD2TTLFNqkNe	6382879.000	4948.000
4135001	535.914	5466.857	3PZuoBAID5Wc2HD	0.584	49.002	27.853	HmiyD2TTLFNqkNe	6316818.286	5386.286
4135001	536.457	5472.429	7JYdWkrLSPkdwr4	0.723	49.109	28.368	HmiyD2TTLFNqkNe	7120640.000	5336.000
4135001	491.700	5014.857	McdE0feGgRqW7Ca	0.924	22.858	25.184	HmiyD2TTLFNqkNe	7158964.000	5192.571
4135001	471.600	4809.000	uHbuxQJl8lW7ozc	0.931	20.428	21.909	HmiyD2TTLFNqkNe	7287002.000	5387.000
4135001	489.529	4992.286	wCURE6d3bPkepu2	1.017	20.427972	21.909288	HmiyD2TTLFNqkNe	7028601	5291.429
4135001	521.129	5315.429	z9Y9gH1T5YWrNuG	0.495	20.427972	21.909288	HmiyD2TTLFNqkNe	7251204	5212.000
4135001	489.043	4987.571	zBIq5rxdHJRwDNY	1.089	20.427972	21.909288	HmiyD2TTLFNqkNe	6583369	5169.857
4135001	323.825	4991.857	zVJPv84UY57bAof	0.434	20.427972	21.909288	HmiyD2TTLFNqkNe	7363272	5021.500

Step #2: Checking Missing Values in the Dataset

The main goal of this work is the prediction of Dc-power generated from solar plants in the next few days based on the current and the history of Dc-power generated from the scientific solar plant, while the prediction based on the law says "The results of any predictor is true if predictor build on fact otherwise the result became virtual". Therefore; in this step, we drop any record of the dataset after merging has missing values to ensure the predictor is built from fact rather than estimation or virtual values.

Step #3 Split Dataset into Multi Intervals

This step presents a split of data rows into multi-intervals based on the ***Source-key***, this is very important to remove the duplicate, where we get 50 different intervals, and each interval contains a sequence of records.

The main purpose of this step is to delete the duplication and enhance the time of implementation by reducing it.

Step #4 Apply the Correlations

The Pearson correlation is considered the most common correlation coefficient used in this step to determine which feature has more effect on the target. This measure computes only the numerical features resulting from step#2 that include only eight features (i.e., "Dc_Power", "Ac_Power", "Daily_Yield","Total_Yield","Ambient_Temperature","Module_Temperature","Irradiation", and "Source-Key"). The result of this step is explained in table (4.4).

Table (4.4): The Results of the Pearson Correlation

	DC-**power**	AC	D_YIELD	T_YIELD	AM_TEM	TEM	IRR
DC-**power**	1.00	1.00	0.09	0.00	0.72	0.95	0.99
AC	1.00	1.00	0.09	0.00	0.72	0.95	0.99
D_YIELD	0.09	0.09	1.00	0.01	0.47	0.20	0.09
T_YIELD	0.00	0.00	0.01	1.00	-0.04	-0.02	0.00
AM_TEM	0.72	0.72	0.47	-0.04	1.00	0.85	0.72
TEM	0.95	0.95	0.20	-0.02	0.85	1.00	0.96
IRR	0.99	0.99	0.09	0.00	0.72	0.96	1.00

Table (4.4) presents three features indicated by a circle that is more important and has a high correlation with the target. The first feature is AC where it correlates equally to 1 with Dc-power this means that the relation is expulsion between them. The second feature is IRR which correlates equally to 0.99 with the target while the third feature is TEM correlation between it and Dc-power equal to 0.95.

By applying many thresholds value based on the try and error principle to find the effects of each feature on generating the maximum Dc-power, finally, we chose features that only correlate equal to or more than the threshold of 0.95 as the most significant features.

4.2.4 Execute the Multi Parameters Objective Functions

In this section, we will use the linear combination of maximum values of the most affected parameters resulting from step # 4 ("Ac-Power", "Module_Temperature", and "Irradiation") to generate results of four equations; MPF1, MPF2, MPF3, and MPF4 and then consider these equations as kernel of DMP-DGBM to satisfy the maximum prediction value of Dc-power.

Tables (4.5) and (4.6) show the computing of four equations for the rate of training and testing through five cross-validations split.

We found the best new kernel equation represented by MPF4 where it satisfies the purpose of the prediction; ***reduces the computations*** and on the other hand, ***reduces the time of implementation***.

Table (4.5): The Result of the Training Dataset in DGBM

Training	Testing	MPF1 =Max (AC+IRR)	MPF2 =Max (AC+TEM)	MPF3 =Max (TEM+IRR)	MPF4 =Max (AC+IRR+TEM)
80%	**20%**	**232.600**	**259.883**	**28.683**	**309.332**
60%	**40%**	220.120	250.910	23.710	254.322
50%	**50%**	202.150	200.783	20.503	172.232
40%	**60%**	195.436	220.232	15.323	189.214
20%	**80%**	123.200	133.853	12.236	122.203

Table (4.6): The Result of the Testing Dataset in DGBM

Training	Testing	MPF1 =Max (AC+IRR)	MPF2 =Max (AC+TEM)	MPF3 =Max (TEM+IRR)	MPF4 =Max (AC+IRR+TEM)
80%	**20%**	**232.467**	**252.403**	**22.661**	**254.452**
60%	**40%**	229.419	138.912	15.852	109.123
50%	**50%**	188.987	130.952	14.202	100.325
40%	**60%**	192.257	220.694	12.235	251.773
20%	**80%**	128.600	123.213	11.235	233.452

As seen in tables(4.5) and (4.6), the four equation values produced in the first row (i.e., 80% of the intervals for training while 20% for testing) are considered the best prediction results. On other hand, it generated maximum Dc-power as compared to other splits, in general:

- In the first column, the MPF1 represents the linear combination of the maximum values to the AC and IRR effective parameters which are equal to 232.600 and 232.467, which are the best as compared to the rate of the rest splits in the same column.

$$MPF1 = \sum_{i=1}^{n} \mathbf{Max}\ (\mathbf{AC}, \mathbf{IRR}) \qquad \ldots(4.1)$$

The values produced by the MPF2 that combines both AC and TEM effective parameters are equal to 259.883 and 252.403, which are also

considered the best as compared to the rate of the rest splits in the same column. $MPF2 = \sum_{i=1}^{n} \mathbf{Max}\ (\mathbf{AC}, \mathbf{TEM})$ …(4.2)

The third MPF3 equation that combines TEM and IRR effective parameters values are equal to 28.683 and 22.661 which also gives the best prediction value as compared to the remainder splits.

$$MPF3 = \sum_{i=1}^{n} \mathbf{Max}\ (\mathbf{TEM}, \mathbf{IRR}) \quad \ldots(4.3)$$

- Finally, in the fourth MPF4 equation, we get the highest and most accurate prediction values as compared to all the rates, which are equal to 309.332 and 254.452, resulting from mixing between all the three previous maximum effective parameters.

As a result, the use of **MPF4** as the kernel of GBM expected makes the accuracy of the model enhancement and makes it more confident.

$$MPF4 = \sum_{i=1}^{n} \mathbf{Max}\ (\mathbf{AC}, \mathbf{IRR}, \mathbf{TEM}) \quad \ldots(4.4)$$

4.2.5 Results of DMP-DGBM Stage

GBM is one of the data mining prediction techniques working with a lot of data and gives high-performance results but on the other side, it has limitations related to its kernel DT making it has high computation and time complexity, therefore replacing its kernel with Multi-Parameter Function is considered a good choice to avoid these limitations.

In other words, determining the effective parameters (i.e., Ac, IRR, and TEM) and then using the Multi-Parameter Functions to generate four equations; three of those equations have two effective parameters while one equation has three effective parameters that enhance the performance of GBM and reduce the computation of software as shown in this section.

In general, the main parameters used to test the performance of the GBM are shown in table (4.7) while the main parameters used to test the DGBM are explained in table (4.8).

Table (4.7): The Parameters Utilized in Traditional GBM

Parameter	Description
Dataset (D)	*D* = 50 intervals
Number of features (m)	m = 8
Maximum number of trees (Tmax)	20
Max depth	5
Learning rate (Sk)	0.1
Min-sample split	2
Activation functions	DT

Table (4.8): The Parameters Utilize in DGBM

Parameter	Description
Dataset (D)	*D* = 50 intervals
Number of features (m)	m = 8
Maximum number of Iterations	100
MPF#1	Max(AC+IRR)
MPF#2	Max(AC+TEM)
MPF#3	Max(IRR+TEM)
MPF#4	Max(AC+IRR+TEM)

Table (4.8) explains the main parameters of DGBM that are required to satisfy the goal of the development. The used dataset represented 50 different intervals, each of them included 8 features and the number of iterations used is equal to 100**.**

- **MPF1** contains the maximum values of AC and IRR parameters.
- **MPF2** also represents the maximum values of AC and TEM.
- **MPF3** represents the maximum values of IRR and TEM**.**
- **MPF4** represents the maximum values of AC, IRR, and TEM**.**

4.3 Comparison between the DMP-DGBM and the Traditional GBM

In this section, we show a comparison between the GBM that used the DT as kernel and DGBM that used four different kernels shown in table 4.8 on the same dataset and same environment. The results of the fifteen intervals for both techniques are shown in table (4.9) which shows a comparison between the original Dc-power and the Dc-power generated through traditional GBM and DMP-DGBM applying on 50 different intervals resulting from pre-processing step #4.

Table (4.9): The Difference between GBM and DMP-DGBM Predictions Values as Compared to the Original

# Interval	Original Dc-power	Results of prediction Dc-power based GBM	Results of prediction Dc-power based DMP-DGBM
Intervl #1	8110	8108.135	8109.421
Intervl #2	8211.143	8209.468	8210.792
Intervl #3	8274.875	8272.956	8274.528
Intervl #4	6829.429	6828.15	6829.362
Intervl #5	7948	7946.856	7947.542
Intervl #6	7889.571	7888.346	7889.428
Intervl #7	8200	8198.407	8199.976
Intervl #8	8391	8389.177	8390.582
Intervl #9	7380.571	7378.885	7379.855
Intervl #10	6981.875	6980.628	6981.086
Intervl #11	7251.714	7249.966	7251.685
Intervl #12	7161.857	7160.175	7161.682
Intervl #13	8404.25	8402.685	8403.829
Intervl #14	5085.857	5084.057	5085.595

Intervl #15	4015.5	4013.685	4015.068
Intervl #16	3541	3539.239	3540.556
Intervl #17	3170.143	3169.011	3169.956
Intervl #18	3592.286	3590.93	3591.519
Intervl #19	5938	5936.919	5937.228
Intervl #20	3690.429	3688.553	3689.995
Intervl #21	2986	2984.113	2985.491
Intervl #22	4983.714	4982.192	4982.725
Intervl #23	4134.857	4133.116	4134.398
Intervl #24	3395.714	3394.706	3395.595
Intervl #25	3834.25	3832.837	3833.389
Intervl #26	5278.857	5277.485	5278.579
Intervl #27	5475.571	5474.373	5475.329
Intervl #28	6005.429	6004.241	6005.102
Intervl #29	5290.125	5288.417	5289.88
Intervl #30	5795	5793.209	5794.405
Intervl #31	7028	7026.28	7027.958
Intervl #32	5680	5678.515	5679.786

Intervl #33	5932	5930.682	5931.087
Intervl #34	4303.143	4301.889	4303.083
Intervl #35	0	0	0
Intervl #36	3219.286	3218.199	3218.813
Intervl #37	4325.5	4324.361	4324.575
Intervl #38	3819.375	3817.978	3819.332
Intervl #39	3608.625	3607.508	3607.69
Intervl #40	3564.571	3563.164	3564.435
Intervl #41	3707.25	3705.288	3706.623
Intervl #42	3864.625	3863.36	3864.144
Intervl #43	3504.125	3502.547	3503.913
Intervl #44	3475.286	3473.441	3475.14
Intervl #45	3568.625	3567.287	3568.197
Intervl #46	3996.875	3994.976	3995.879
Intervl #47	3299.429	3297.601	3298.627
Intervl #48	3721.375	3719.545	3720.55
Intervl #49	3395.857	3394.558	3395.131
Intervl #50	3706.143	3704.379	3705.974

Table (4.9) founding that a result of the DMP-DGBM is considered more accurate and more closely to the fact. Also, figure (4.1) shows the comparison between the results of GBM, the DMP-DGBM, and the original values of the Dc-power.

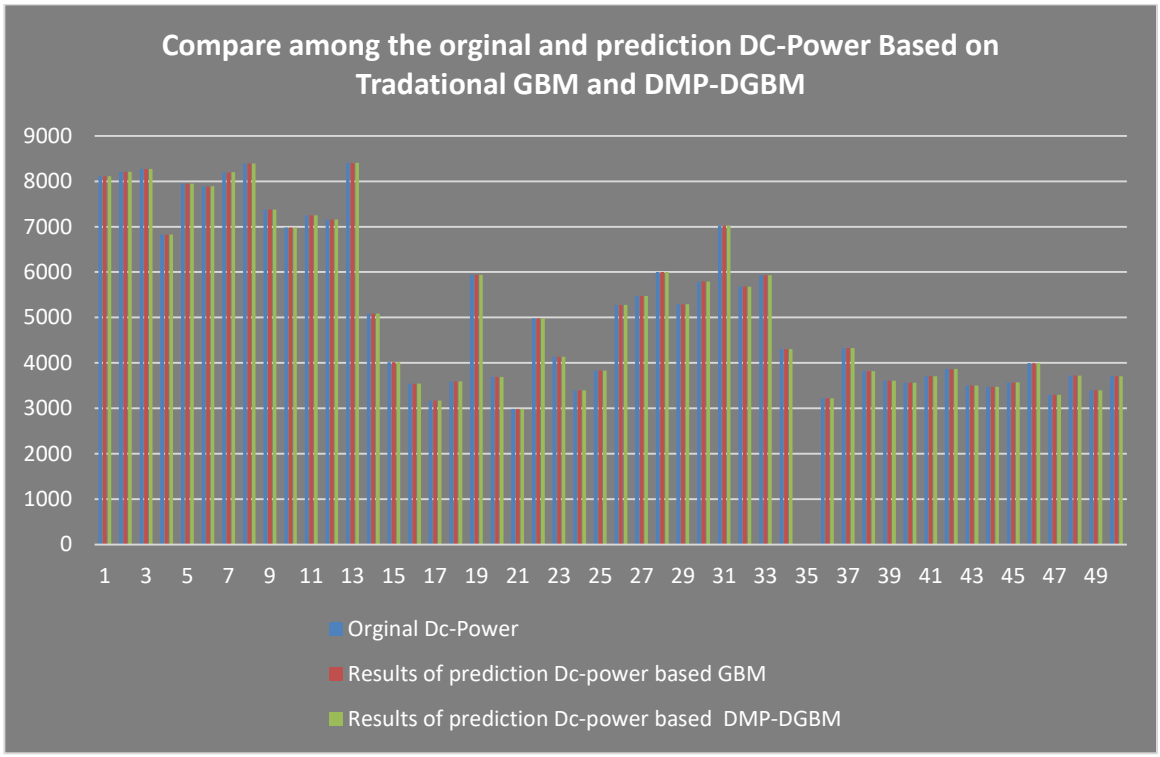


Figure (4.2): A comparison of the Actual and Predicted Dc-power Based on GBM and DMP-DGBM

4.4 Results of the Evaluation Stage

The performances of the DMP-DGBM test in this work are based on three different measures which are **R^2**, **MSE**, and **RMSE**. Founding that the best result was generated by splitting the fifteen different intervals into 80% for training and 20% for testing.

The results of performance for both the training and testing dataset is shown in table (4.10).

Table (4.10): Evaluation Measures of the Training and Testing Dataset

Rate of Training Dataset	Rate of Testing Dataset	Results of the Training Dataset			Results of the Testing Dataset		
		R^2	MSE	RMSE	R^2	MSE	RMSE
80% = 40 intervals	**20%** = 10 intervals	**0.9742**	**0.0099**	**0.0522**	**0.8944**	**0.0193**	**0.0208**
60% = 30 intervals	**40%** = 20 intervals	0.9899	0.1056	0.0559	0.8599	0.0197	0.0859
50% = 25 intervals	**50%** = 25 intervals	0.9499	0.0094	0.0690	0.8099	0.0196	0.0853
40% = 20 intervals	**60%** = 30 intervals	0.9411	0.0085	0.0632	0.2916	0.0194	0.1025
20% = 10 intervals	**80%** = 40 intervals	0.8999	0.0079	0.0526	0.6871	0.0195	0.7693

Table (4.10) show that the results generated by the DGBM algorithm are more accurate and confident than the traditional GBM.

In addition, the next section will show the main questions that found answers through this work.

By seeing the datasets which we dealing with, founded the followings facts:

- The maximum of Dc-power is generated when the AC value is equal to 1405.3 which leads to generating the amount of Dc-power = 1472.068 through applying the MPF4 of the train split 80%, while the minimum value of Dc-power is equal to 1395.507 generated when the AC equal to 1394.286 through applying the MPF1 at the train split of 50%, 40%, and 20%.
- The maximum value of Dc-power is 1472.068 produced when the IRR is equal to 1.222 which is fixed for each 5-cross validation split while getting the minimum value of Dc-power which is equal to 0 when IRR is 0 through the whole testing rate when applying the MPF3.
- The maximum value of Dc-power is 1472.068 produced when TEM is equal to 65.546 through applying the MPF4 where the training rate is 80% and 60% while the minimum Dc-power is 0 when TEM is 0 generated through all the testing rates through applying the MPF3.

Chapter Five:

Conclusions and Future Work

Chapter Five: Conclusion and Future Works

5.1 Introduction

This chapter explains the main conclusions found through implementing the suggested system ***ZME-DGBM***. We focus on the point of how the proposed model handles the challenges shown in the previous chapters (i.e., programming and application challenges).

Also, it will suggest a set of recommendations for other researchers in the future.

5.2 Conclusion

This section will summarize the main conclusions found from implementing ***ZME-DGBM***:

A. Although Renewable Energy is considered one of the main resources for generating energy and it is characterized by multi-features that make it a pragmatic resource to cover the need of humans for energy, but on the other side, it required more effort from humans to get it.

B. Sometimes the data collected from multi-sensors suffer from duplications of their records, which leads to an increase in the computations. In general, the ZME-DGBM handles this problem by splitting the dataset into multi-intervals, each interval represents a stream of data, then create a buffer to save only the different intervals in order to work on it in the next step.

C. Integration of data always leads to an increase in the accuracy of results and makes it more understood by users. Therefore, this work achieves this feature through integrated weather and solar datasets to get more accurate results.

D. The GBM is one of the primitive tools for prediction based on a decision tree as a kernel to generate the decision, but on the other side, it is required to determine multi-parameters such as root, number

of nodes, and depth of the tree. This work handles the limitations of GBM by replacing the kernel (DT) with a mathematical function based on linear combinations. This leads to getting high accuracy results in a short implementation time. In general, the best results are these the study got through using multi-objective functions that have three parameters.

E. The ZME-DGBM gives the best results when splitting the different intervals through a rate of 80% of the intervals to training while 20% of intervals to testing and using MPF4 as kernel of GBM this maximizes the accuracy of the model.

5.3 Future Works

The following points may be good ideas for future work.

A. We can use one of the neurocomputing prediction techniques such as long short-term memory or recurrent neural network.

B. The ZME-DGBM applies through traditional hardware called the central processing unit(CPU) while in the future it appears to benefit implement on other hardware such as graphics processing unit (GPU) or filed programming get array (FPGA).

C. The location of the plant is a very important parameter to increase or decrease the DC generation but this study does not take it into account. Therefore, we recommended of the researchers study the effect of this parameter on the generation of the DC.

References

A. Razmjoo, L. Gakenia Kaigutha, M.A. Vaziri Rad, M. Marzband, A. Davarpanah, M. Denai,(2021), A Technical analysis investigating energy sustainability utilizing reliable renewable energy sources to reduce CO2 emissions in a high potential area, Renewable Energy, Volume 164,Pages 46-57, https://doi.org/10.1016/j.renene.2020.09.042.

Ahmed M.A. Haidar, Adila Fakhar, Andreas Helwig,(2020),Sustainable energy planning for cost minimization of autonomous hybrid microgrid using combined multi-objective optimization algorithm,Sustainable Cities and Society,Volume 62,102391,https://doi.org/10.1016/j.scs.2020.102391.

Al_Janabi, S. (2015). A Novel Agent-DKGBM Predictor for Business Intelligence and Analytics toward Enterprise Data Discovery. *Journal of Babylon University/Pure and Applied Sciences*, *23*(2), 482-507.

Al-Janabi, S., & Mahdi, M. A. (2019). Evaluation prediction techniques to achievement an optimal biomedical analysis. International Journal of Grid and Utility Computing, 10(5), 512-527.

Al-Janabi, S., Alkaim, A., Al-Janabi, E., Aljeboree, A., & Mustafa, M. (2021). Intelligent forecaster of concentrations (PM2. 5, PM10, NO2, CO, O3, SO2) caused air pollution (IFCsAP). Neural Computing and Applications, 1-31. https://doi.org/10.1007/s00521-021-06067-7

Amir, M., & Khan, S. Z. (2021). Assessment of renewable energy: status, challenges, COVID-19 impacts, opportunities, and sustainable energy solutions in Africa. Energy and Built Environment. https://doi.org/10.1016/j.enbenv.2021.03.002

Ardianto, R., Rivanie, T., Alkhalifi, Y., Nugraha, F. S., & Gata, W. (2020). Sentiment Analysis on E-Sports For Education Curriculum Using Naive Bayes and Support Vector Machine. Jurnal Ilmu Komputer dan Informasi, 13(2), 109-122. DOI: https://doi.org/10.21609/jiki.v13i2.885

Bahareh Oryani, Yoonmo Koo, Shahabaldin Rezania, Afsaneh Shafiee,(2021), Barriers to renewable energy technologies penetration: Perspective in Iran, Renewable Energy,Volume 174,Pages 971-983, https://doi.org/10.1016/j.renene.2021.04.052.

Basaran, K., Özçift, A., & Kılınç, D. (2019). A New Approach for Prediction of Solar Radiation with Using Ensemble Learning Algorithm. Arabian Journal for Science and Engineering. doi:10.1007/s13369-019-03841-7

C.V. Diezmartínez, (2021), Clean energy transition in Mexico: Policy recommendations for the deployment of energy storage technologies,Renewable and Sustainable Energy Reviews,Volume 135, 110407, https://doi.org/10.1016/j.rser.2020.110407

Cotfas, L. A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D. S., & Tajariol, F. (2021). The Longest Month: Analyzing COVID-19 Vaccination Opinions Dynamics from Tweets in the Month following the First Vaccine Announcement. IEEE Access, 9, 33203-33223. doi**:** 10.1109/ACCESS.2021.3059821.

Das, H. S., & Roy, P. (2019). A deep dive into deep learning techniques for solving spoken language identification problems. In Intelligent Speech Signal Processing (pp. 81-100). Academic Press https://doi.org/10.1016/B978-0-12-818130-0.00005-2.

Dogan, A., & Birant, D. (2020). Machine Learning and Data Mining in Manufacturing. Expert Systems with Applications, 114060. doi:10.1016/j.eswa.2020.114060

Guozhou Zhang, Weihao Hu, Di Cao, Wen Liu, Rui Huang, Qi Huang, Zhe Chen, Frede Blaabjerg, (2021) , Data-driven optimal energy management for a wind-solar-diesel-battery-reverse osmosis hybrid energy system using a deep reinforcement learning approach,Energy Conversion and Management,Volume 227,113608, https://doi.org/10.1016/j.enconman.2020.113608.

Han, J., & Kamber, M. (2006). Data mining: concepts and techniques, 2nd. University of Illinois at Urbana Champaign: Morgan Kaufmann.

Hao, J. (2020). Deep Reinforcement Learning for the Optimization of Building Energy Control and Management (Doctoral dissertation, University of Denver).

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hossny, K., Magdi, S., Soliman, A. Y., & Hossny, A. H. (2020). Detecting explosives by PGNAA using KNN Regressors and decision tree classifier: A proof of concept. Progress in Nuclear Energy, 124, 103332. doi:10.1016/j.pnucene.2020.103332

https://doi.org/10.1016/j.enconman.2019.04.064.

Khan, A., Sohail, A., Zahoora, U. et al. (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53, 5455–5516. https://doi.org/10.1007/s10462-020-09825-6

Luger, G. F. (2005). Artificial intelligence: structures and strategies for complex problem solving. Pearson education**.**

Mahdi, M. A., & Al_Janabi, S. (2019, April). A novel software to improve healthcare base on predictive analytics and mobile services for cloud data centers. In International conference on big data and networks technologies (pp. 320-339). Springer, Cham https://doi.org/10.1007/978-3-030-23672-4_23.

Miljkovic, D. (2017). Brief review of self-organizing maps. 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). doi:10.23919/mipro.2017.7973581

Ning Li, Zhanguo Su, Houssem Jerbi, Rabeh Abbassi, Mohsen Latifi, Noritoshi Furukawa,(2021),Energy management and optimized operation of renewable sources and electric vehicles based on microgrid using hybrid gravitational search and pattern search algorithm,Sustainable Cities and Society,Volume 75,103279,ISSN 2210-6707,https://doi.org/10.1016/j.scs.2021.103279.

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos one*, *16*(2), e0245909 https://doi.org/10.1371/journal.pone.0245909.

Shin'ya Obara, Yuji Ito, Masaki Okada,(2018), Optimization algorithm for power-source arrangement that levels the fluctuations in wide-area networks of renewable energyEnergy,Volume 142,Pages 447-461, https://doi.org/10.1016/j.energy.2017.10.038.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education India.

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158, 1533-1543. https://doi.org/10.1016/j.enbuild.2017.11.039

Utkucan Şahin,(2020),Projections of Turkey's electricity generation and installed capacity from total renewable and hydro energy using fractional nonlinear grey Bernoulli model and its reduced forms,Sustainable Production and

Consumption,Volume 23,Pages 52-62,https://doi.org/10.1016/j.spc.2020.04.004.

Vinoth Kanna, I., & Pinky, D. (2020). Solar research–a review and recommendations for the most important supplier of energy for the earth with solar systems. In International Journal of Ambient Energy (Vol. 41, Issue 8, pp. 962–968). Taylor and Francis Ltd. https://doi.org/10.1080/01430750.2018.1472658

الخلاصة

شهد العالم خلال العقود الاخيرة توسع كبير في عالم التكنولوجيا والالكترونيات بالإضافة الى التطور الهائل الحادث في الصناعات المختلفة مما ادى الى زيادة الاحتياج الى الطاقة الكهربائية بشكل ملحوظ وكانت الطاقة المتجددة المتولدة من مصادر صديقة للبيئة مثل الطاقة(الشمسية، المياه، طواحين الهواء وغيرها) هي الحل البديل لتوفير تلك الطاقة وخاصة بانها طاقة نظيفة لا تسبب انبعاث غاز ثاني اوكسيد الكاربون الذي يسبب تلوث الهواء والبيئة بشكل عام.

تقدم هذه الرسالة نموذج برمجي لإنتاج اكبر كمية من الطاقة من خلال تطوير لواحدة من افضل تقنيات التنبؤ واستخدام دوال هدف متعددة المعاملات. حيث يتألف النموذج المقترح المسمى ZME-DGBMمن عدة مراحل ويعالج اثناء ذلك العديد من التحديات الاساسية التحدي الاول هو برمجي حيث على الرغم من ان ال GBM هي افضل تقنية تنبؤ الا ان جزء اتخاذ القرار بها وهو ال DT مما يجعلها بطيئة نوعا ما وتحتاج الى تحديد عدد الاشجار وعمق الشجرة وعدد العقد الختامية في كل شجرة. تجاوز النموذج المقترح هذه المشكلة باستبدال ال Kernel ل GBM بدوال هدف متعددة المعاملات. اما التحدي الثاني فهو يتعلق بالتطبيق نفسه والتضمن كيفية انتاج اكبر كمية من الطاقة الكهربائية من الخلايا الشمسية بأعلى دقة واقل وقت تنفيذ.

يتألف النموذج المقترح من اربع مراحل: مرحلة استحصال البيانات واجراء معالجة اولية عليها تضمنت : فحصها اذا كانت تحتوي على قيم مفقودة، حساب الترابط ما بين الخصائص والهدف وتقسيم البيانات الى فترات مختلفة وحذف الفترات المتكررة. مرحلة بناء المتنبئ الذي اعتمد على استبدال ال Kernel ل GBM بأربع دوال مختلفة متعددة المعاملات حيث كانت تلك هي المعاملات الاعلى ترابط مع الهدف واعتمدت قيمة عتبة 0.95 كتحديد اهمية الخاصية (المعاملات)، المرحلة الاخيرة تم تقييم نتائج المتنبئ باستخدام عدة مقاييس لتحديد دقة النتائج التي تم التوصل اليها وتم استخدام كلاً من " MSE, RMSE, R2 " .

امتاز النموذج المقترح بإعطاء افضل النتائج باستخدام دالة ذات ثلاث معاملات ل Kernel ال GBMوكانت تلك المعاملات هي (AC, TEM, and IRR) حيث كان مقياس (R2=0.9742) بينما (MSE=0.0099) و (RMSE= 0.0522) استغرق تنفيذ النظام .80 Ms على حاسبة Core i5 بلغة برمجة Python 3.8.13.

وزارة التعليم العالي والبحث العلمي
جامعة بابل
كلية العلوم للبنات
قسم علوم الحاسبات

نظام تنبؤ ذكي لتعظيم الطاقة المتجددة بالاعتماد على تطوير ميكانيكية تعزيز الانحدار (DGBM)

رسالة
مقدمة الى مجلس كلية العلوم للبنات - جامعة بابل وهي جزء من
متطلبات نيل شهادة الماجستير في العلوم/علوم الحاسبات

مقدمة من قبل

زينب خيرالله الجنابي

بإشراف

أ.د. سماهر حسين علي الجنابي

١٤٤٤هـ ٢٠٢٢م