Transactions on Machine Learning and Artificial Intelligence - Vol. 9, No. 4
Publication Date: August, 25, 2021
DOI:10.14738/tmlai.94.10281.

Namaswa, S., Githiri, J., Mariita, N., K'Orowe, M. O., & Njiru, N. (2021). Application of Data-Driven Discovery Machine Learning Algorithms in Predicting Geothermal Reservoir Temperature from Geophysical Resistivity Method. Transactions on Machine Learning and Artificial Intelligence, 9(4). 01-12.

Application of Data-Driven Discovery Machine Learning Algorithms in Predicting Geothermal Reservoir Temperature from Geophysical Resistivity Method

Solomon Namaswa
Physics Department, Jomo Kenyatta
University of Agriculture and Technology, Kenya
Department of Physics, Multimedia University of Kenya

John Githiri
Physics Department, Jomo Kenyatta
University of Agriculture and Technology, Kenya

Nicholas Mariita
Geothermal Training and Research Institute
Dedan Kimathi University of Technology, Kenya

Maurice O. k'Orowe
Physics Department, Jomo Kenyatta
University of Agriculture and Technology, Kenya

Nicholas Njiru
Department of Computer Science, Multimedia University of Kenya

ABSTRACT

Geophysical methods including seismology, resistivity, gravity, magnetic and electromagnetic have been put in use for geothermal resource mapping at the Great Olkaria Geothermal field for decades. Reservoir temperature distribution and the electrical conductivity of rocks mainly depend on the same parameters such permeability, porosity, fluid salinity and temperature. This research focused on the integration of Olkaria Domes geothermal well testing temperature and geophysical Electromagnetic resistivity data with the aim of establishing an alternative estimation method for temperature of the reservoir through machine learning Analytics. To achieve this, Data-Driven Discovery Predictive Model Algorithm was built using Python programming language on Anaconda framework. The open-source web based application Jupyter Notebook for coding and visualization was used. Decision Tree Regression, Adaptive Booster Regression, Support Vector Regression and Random Forest Regression were used. The model performance was evaluated using R-Score and Mean Absolute Error metrics. Based on these performance score, the best performing model was suggested to predict subsurface temperature from resistivity. Training the model using the DTR algorithm approach provides superior outputs with R^2 of 0.81 and lowest MAE of 29.8. The DTR algorithm could be implemented in determination of subsurface Temperature from resistivity in high temperature hydrothermal fields.

Key words: Geothermal Reservoir, Temperature, Resistivity, Data-driven modelling, Machine Learning

INTRODUCTION

A valuable large volume of heat bearing formations that has the ability to supply a considerable amount of energy if suitable extraction techniques are applied for exploitation is contained in the earth's subsurface (Hu, 2016). For proper exploitation of this resource, a comprehensive and integrated reservoir analysis from exploration to production monitoring need to be implemented (Aragón et al., 2019). Prediction and mapping of reservoir properties in the subsurface requires the integration of knowledge as a result of recognition of the importance and advantages of integrated reservoir studies (Zhu et al., 2020).

The approaches employed in determination of some of subsurface properties such as well drilling for subsurface temperature measurements are time-consuming and expensive. Relating and integrating these reservoir properties with the geophysical rock resistivity can be highly effective, since the reservoir temperature distribution and the electrical conductivity of rocks mainly depend on the same parameters such as porosity, permeability, tortuosity and pore geometry (Adebayo et al., 2019). The electrical conductivity of reservoir rocks and aquifers can be measured at the surface using geophysical survey methods such as Electromagnetic methods. However subsurface temperature can only be obtained through well logging after drilling the well. Therefore, a method for reliably estimating subsurface temperatures from surface electrical conductivity measurements would be beneficial.

DATA-DRIVEN DISCOVERY PREDICTIVE MODELING

Data-Driven Discovery Predictive Modeling is defined as the construction of robust empirical models of real or complex systems with the intension of helping decision makers establish the relationship between the input and output parameters of the system without having a proper understanding of the physical behavior of the system (Mont´ans et al., 2019). This is illustrated in Figure 1.

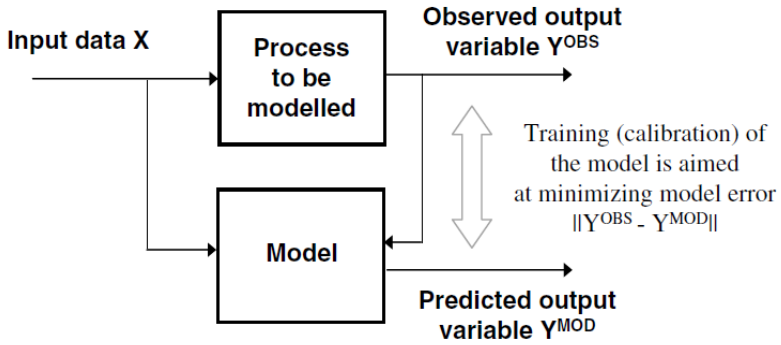


Figure 1: Data Driven Modelling Approach (Abrahart et al., 2008)

The technique is able to build models that characterize the performance of corresponding physical processes through the study of relevant data characterizing the systems of interest by inferring the dependencies between system inputs and outputs using certain learning algorithms without building the complex physical models (Lee et al., 2019).

Machine Learning

Machine learning is the process of iteratively training models to learn from data through experience without being programmed explicitly, without a human intervention unlike traditional programming methodology. ML improves learning when given new unseen data by use of diverse ML algorithms to help non-experts in ML in Decision making. Machine learning algorithms are classified into three types; supervised, unsupervised and reinforced learning algorithms (Schmidt et al., 2019).

Supervised is further divided into Classification and Regression algorithms (CART). Classification algorithms are used for data that are categorical in nature, that is, the expected output is categorized into classes and used in the prediction of classes. In regression is a supervised ML algorithms, the model trained are applied in continuous data values where a continuous real value prediction is expected. The regression problems are further divided into Linear and Non-Linear Regression (Lundkvist, 2014). Linear regression is applied where the data to be learned exhibits a linear relationship with its main goal is to fit a linear equation to the given data in a n-dimensional space. Non-Linear or also known as curvilinear regressions as used in learning data that exhibit non-linearity in data. In Curvilinear models are achieved by setting higher-order of the predictor variables. In supervised learning algorithms, the training model is provided with example data that is human label. Examples of the common popular supervised learning algorithms are Linear Regression, Support Vector Machine (SVM), Naïve-Bayes, Logistic Regression, Classification and Regression Techniques (CART), Decision Trees and K-Nearest Neighbors (Daniya et al., 2020). Unsupervised learning algorithms are provided with unlabeled data to enable the hidden structure or patterns in data. The unsupervised learning is also divided into clustering, dimensionality reduction and association. Clustering is a technique that is used to group observations into natural groupings that are similar to each other as clusters (Njiru et al., 2018). Dimensionality reduction is a technique that is used to reduce the number of features in a dataset to ensure the most important information is conveyed. The dimension of the dataset is reduced for a high- dimensional to a low-dimensional space. Having a very high dimension space makes visualization of the features to become very poor. Association is an algorithm used to find the probability of co-occurrence of items in a data set. Examples of unsupervised algorithms are K-means, Principal Components Analysis (PCA), Apriori algorithms, Linear Discriminant Analysis, Latent Dirichlet Allocation (LDA), Hierarchical clustering, Fuzzy C-means, Singular Value Decomposition (SVD) and Independent Component Allocations (ICA). Reinforcement learning is an algorithm where the machine learning model learns from a dynamic environment through a trial and error method by use of rewards and penalties to the learner based on its performance (Mirza et al., 2019). Regression is a predictive modeling technique that has been proven scientifically to be predicting the future and causal inference. It is a long established statistical procedure that has been adapted in AI and ML because it models re understandable. The choice of a machine learning algorithm in ML is dependent with the type of data. Regression is adapted in numerical data whose value of prediction is a single value hence the choice of the algorithm in this research.

Decision Tree Regression (DTR) is a technique that is used for fitting a sine curve on a model where the continuous value is being sought. It can be used for building classification or regression models. The model observes the features of a dataset to learn the tree structure in the dataset for predicting a real value output. The dataset is broken down into decision nodes

until it reaches the leaf nodes. The Decision nodes can further be divided into either one or two more decision nodes. The top decision node is referred to as the root node and is considered the best predictor. The DTR split the node by use of the core algorithm known as Iterative Dichotomizer 3(ID3). Support Vector Regression (SVR) is an algorithm that is used in prediction of a real value response variable unlike the Support Vector Machine (SVM) which is used in classification to select the hyper plane with the maximum margin. SVR fits a curve in the give data points. The objective of SVR is for numerical prediction. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The fundamental concept behind random forest is a simple but powerful one. In data science speak, the reason that the random forest model works so well is: A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. AdaBoost is a popular boosting algorithm for which the fundamental principle is combining weak estimators and implementing them on improved data versions. The results of each prediction are then combined with the help of weighted majority voting/hard voting. Adaboost Regressor is used for solving regression problems.

METHODS

The MT data were collected by Kenya electricity generating company (KENGEN) between 2005 and 2017 by use of a 5-channel (MTU-5A) MT equipment. The SSMT2000 software was used to process the data in time-series form from the MT. Through Fourier transformation, the data was changed to frequency spectrum. Noisy data points were eliminated by the MT editor. The resulting time-series data were Fourier transformed to the frequency domain. From the Fourier transform band, the robust processing method was used to compute average cross-powers and auto powers, which were then edited to get rid of noisy data by the MT editor program graphically. The data was then saved in Electronic Data Interchange (EDI) files and later used as TEMTD input for joint inversion with the TEM. The TEM equipment used in this survey is from Zonge. A 200 m × 200 m transmitter wire loop was used.

The raw TEM data processed using the program TemxZ where the data collected at same were averaged and late time apparent resistivity computed as a function of time after current turn-off. The MT static shift corrected by TEM data through joint inversion. Resistivity as a function of depth at every well location was generated.

The heat-up temperature profiles data was collected by the Kenya Electricity Generating Company during well completion tests for each well at different times. Cold fluids were pumped into the well at a rate of 1600lpm for a total period of 8 hrs. The well was then shut to allow the well to recover. A temperature profile run was conducted periodically during well recovery. Fifteen wells were randomly selected with the aim of getting a good representation of the entire field. These wells included: OW901, OW902B, OW903, OW905A, OW906, OW908, OW910, OW911A, OW912, OW914B, OW915, OW917, OW918, OW919 and OW921.

Resistivity and temperature datasets were originally provided in excel worksheet. The data was converted into Comma Separated File (CSV) format for modeling. Data from thirteen wells was used for training the model while two wells data observations were set aside for the purposes

of the verification during deployment evaluation. This was in the ration of 80% to 20% for training and testing respectively according to (Chaki, 2015). The training dataset was made up of 297 observations and two features; Resistivity and Temperature. The testing data was important as it was used to evaluate how the model was to handle new data that was outside the training data.

Data-Driven Discovery Predictive Model (3DPM) for Geothermal wells Parameters using Machine Learning computational approach was used in this research (Figure 2). This approach was necessitated by the fact that governing equations have been of fundamental importance across all scientific disciplines and the traditional derivation of governing equations is based on underlying first principles, such as conservation laws and symmetries, or from universal laws, such as gravitation. However, in geothermal systems, the very high variability range of chemical and physical conditions within the Earth represents the primary limit to find theoretical, or even empirical but universal, relationships between temperature and resistivity. Therefore the recourse to first-principles derivations is untenable.

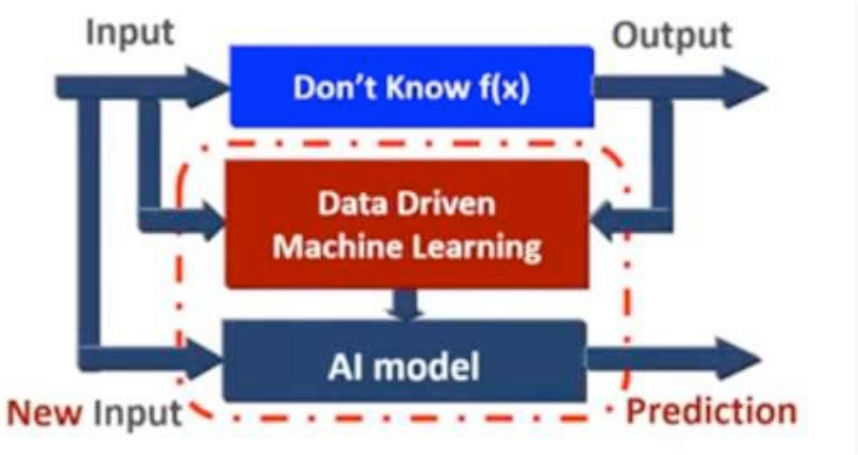


Figure 2: Data-Driven Discovery Predictive Model

A detailed machine learning procedure adopted is described in the flow chart of Figure 3. The model was build using Python programming language and Anaconda framework. NumPy an open source Python library was used in working with arrays. The open-source web based application Jupyter Notebook for coding was used. Matplotlib a graph plotting library in python was used for graph visualization. Pandas was used for reading, analyzing, cleaning, exploring, and manipulating data into comma separated files. SciPy, a scientific computation library was used for optimization and statistical processing. Sklearn was used for predicting temperature attribute from resistivity. Resistivity was used as the input parameter while temperature as the output variable. Using the same set of data, Decision Tree Regression, Adaptive Booster Regression, Support Vector Regression and Random Forest Regression machine learning models were developed in order to establish the best fitting model based on the performance inspectors.

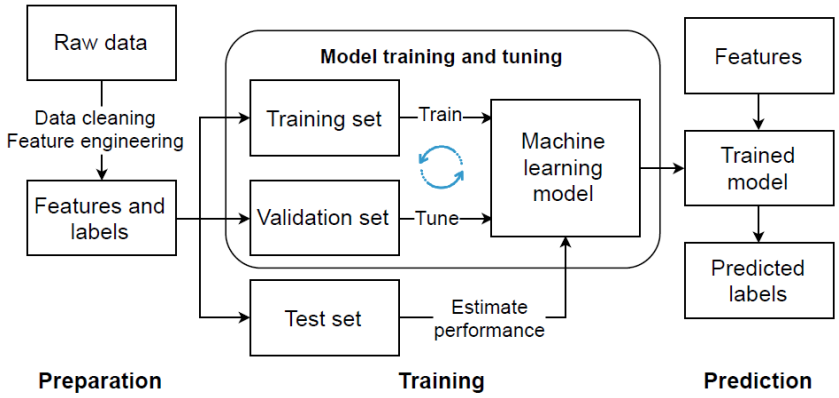


Figure 3: Machine learning prediction flowchart (Steuer, 2018)

RESULTS AND DISCUSSIONS

Exploratory Data Analysis

Before building the model, the exploratory analysis of the dataset was performed. The dataset was made up of 297 Observations and two features, Resistivity and Temp, of data type float. The statistical values for the two features are shown in table 1. There was no missing record in our data.

Table 1: The Dataset Summary Information

	Resistivity	Temperature
Count	297	297
Mean	56.26	215.49
Std	95.03	78.36
min	3	24.5
25%	11	186.6
50%	20	225.1
75%	51	269.1
max	630	364

The histogram in Figure 4 was used to show the shape of statistical data for resistivity. The histogram shows that the data was asymmetrical and skewed toward the left with a tail toward the right, Most of the data lies between the resistivity values between 0 to 200 Ohms and very few points lying above 200 Ohms.

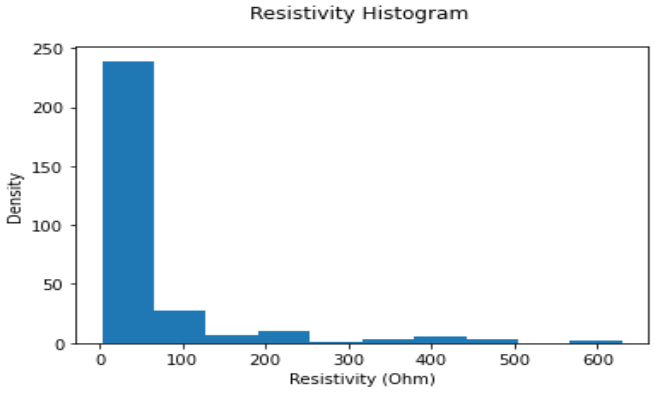


Figure 4: Resistivity Histogram

The Figure 5 shows a right skewed graph. More than 60 percent of our data point's temperature was between approximately 90 Degrees Celsius to 290 Degree Celsius.

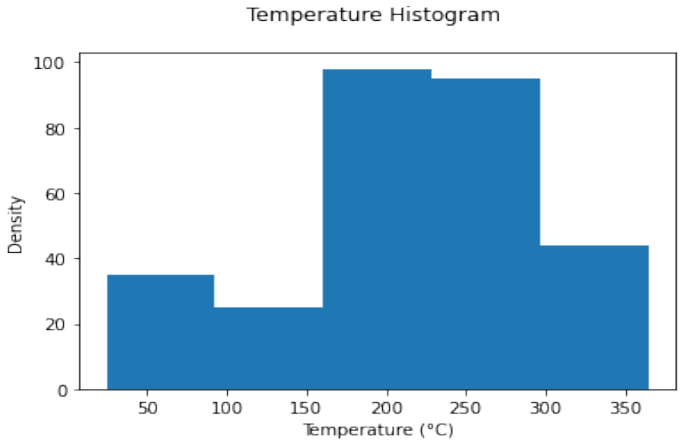


Figure 5: Temperature Histogram

Adaptive Boosting (AdaBoost) Regression

Mean Absolute Error and coefficient of correlation for the AdaBoost regressor are shown in Table 2 below.

Table 2: MAE and R^2 Results

Algorithm	AdaBoost	Random Forest	SVR	DTR	DTR Test Data
MAE	46.99	36.17	34.40	29.8	21.7
R^2	0.66	0.758	0.789	0.81	0.835

From the table, R^2 was 0.66 implying that 66% of the variation between subsurface temperatures is accounted for by the surface resistivity. Mean Absolute Error was 46. The relationship between the predicted values and the actual temperature values of AdaBoost model is shown in scatter plot of Figure 5. From the Figure, there is a fair correlation between the predicted and the actual temperatures especially at temperatures above 300^{o}C as compared to lower temperatures where plots are nonlinear.

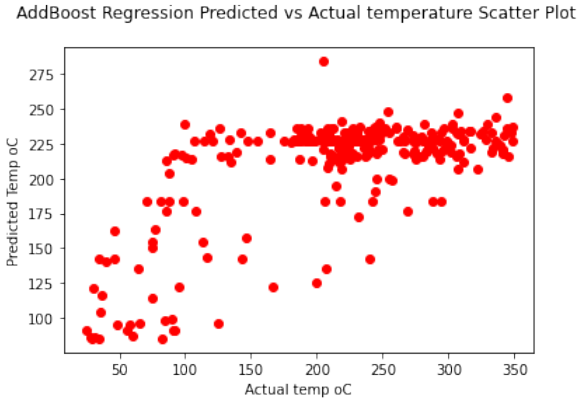


Figure 5: Predicted vs Actual Temperature AdaBoost Scatter Plot

The variability in the predicted vs actual temperature values in the scatter plot indicates that the AdaBoost model was able to learn more features from the data where resistivity values were observed to decrease with depth of 1000 metres below the surface, the trend was reversed where resistivity values started to increase with depth, the AdaBoost predicted temperature showed a slight increase in, however the temperature increase did not match up with the actual temperature increase since they were of lower values shown in Figure 6.

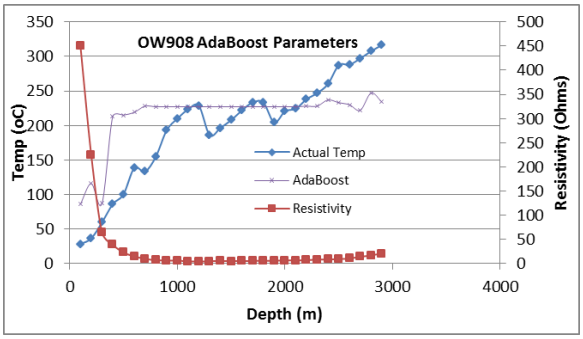


Figure 6: OW908 Actual and AdaBoost Predicted Parameters

Random Forest Regression

From the Table 2, R^2 for random forest regression was 0.789 implying that 78.9% of the variation between subsurface temperatures is accounted for by the surface resistivity. This shows that RFR model performed better that AdaBoost. Mean Absolute Error was 36.17 indicating that the absolute value of the difference between the predicted temperature value and the actual temperature is 36.17.

Figure 7 shows the relationship between the predicted values and the actual temperature values of RFR model. From the scatter plot, there is a better correlation between the predicted and the actual temperatures especially at temperatures above 200°C as compared to lower temperatures where there are more scattered instead of being linear.

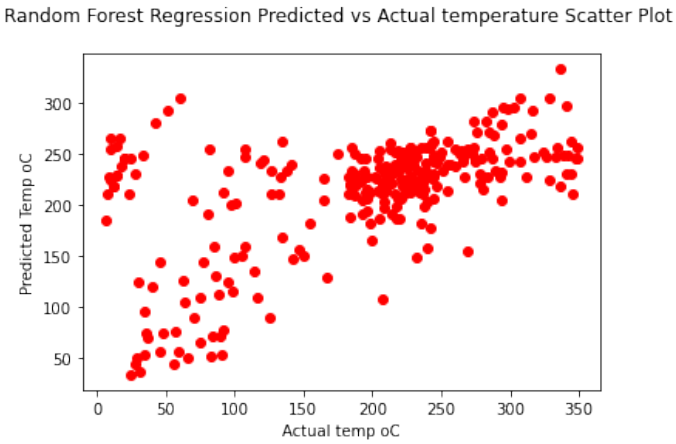


Figure 7: Predicted vs Actual Temperature RFR Scatter Plot

The variability in the predicted vs actual temperature values in the scatter plot indicates that the RFR model was able to learn more features from the data where resistivity values were observed to decrease with depth of 1000 metres below the surface, the trend was reversed

where resistivity values started to increase with depth, the RFR was able to cope up with the increase in temperature as AdaBoost regression model as shown in Figure 8.

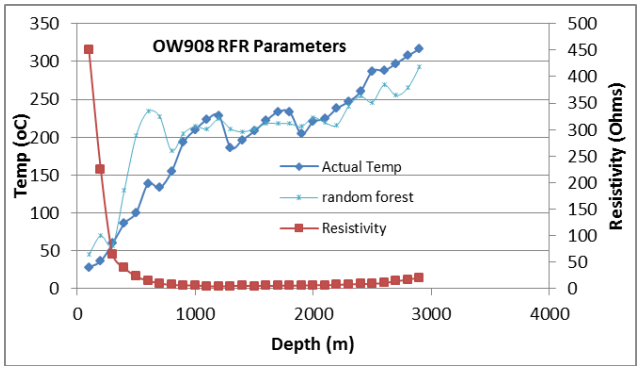


Figure 8: OW908 Actual and RFR Predicted Parameters

Support Vector Regression (SVR)

From the Table 2, R^2 was 0.789. This was a good relationship between temperature and resistivity. SVR model also recorded Mean Absolute Error was 34.4 which was an improvement as compared to AdaBoost and Random Forest Regressors. The relationship between the predicted values and the actual temperature values of SVR model is shown in scatter plot of Figure 9. From the Figure, there is a better correlation between the predicted and the actual temperatures.

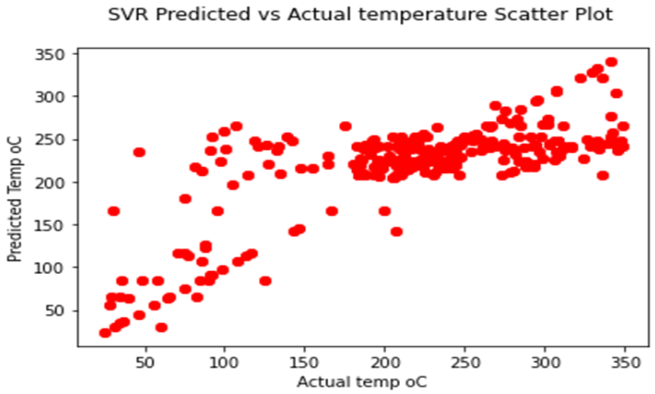


Figure 9: Predicted vs Actual Temperature SVR Scatter Plot

The variability in the predicted vs actual temperature values in the scatter plot indicate that the SVR model was able was able to cope up with the increase in temperature as shown in Figure 10.

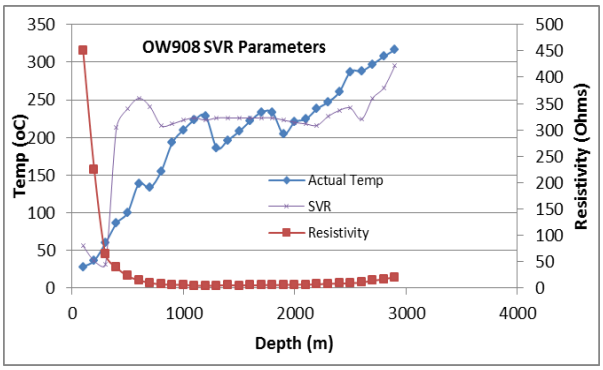


Figure 10: OW908 Actual and SVR Predicted Parameters

Decision Tree Regression (DTR)

Mean Absolute Error and coefficient of correlation for the DTR model are shown in table 3. From the table, R^2 was 0.8147 implying that 81.7% coefficient of correlation between the actual and the predicted temperatures. This value was highest as compared to all the algorithms used. Mean Absolute Error was 29.8 which was also the lowest value. The relationship between the predicted and actual values is depicted in scatter plot of Figure 11.

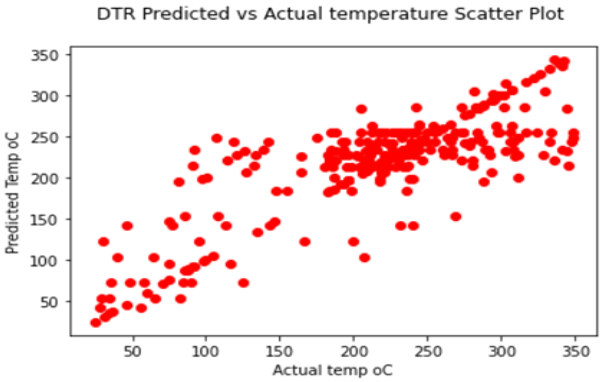


Figure 11: OW908 Actual and DTR Predicted Parameters

As seen from the Figure, the majority of the points fall into a line while there are still quite a few points that fall outside the line indicating that the predicted data well fits the model. Also the DTR model was able to learn the trend of decreasing resistivity with depth up to 1000 metres below surface and they started increasing in values after 1000 metres as shown in Figure 12 and the predicted temperature graph matched well with the actual temperature graph.

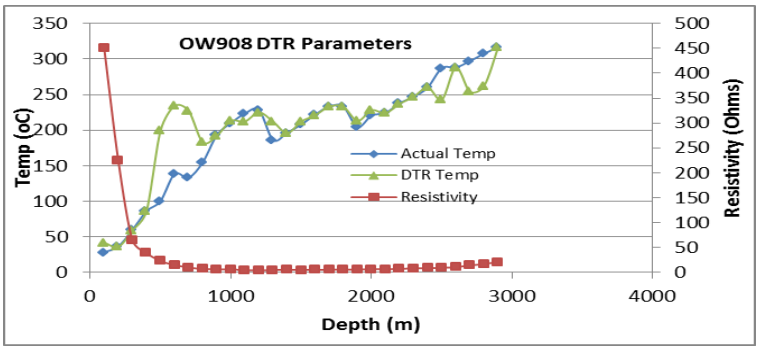


Figure 12: OW908 Actual and DTR Predicted Parameters

Training the model using the DTR algorithm approach provided superior outputs to SVR. This is based on the suggested statistical performance indexes. DTR being the best algorithm for the study was used to test with new data and the R-squared was 0.835 with MAE at 21.7 as shown in Table 2. The Predictive performance of the new data is visualized in Figures 13.

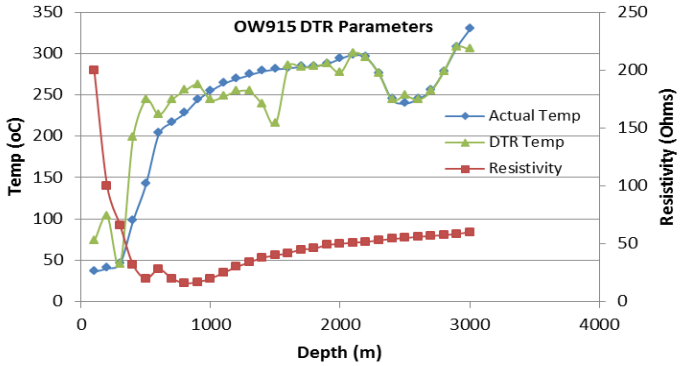


Figure 13: The Predictive performance of the new data, OW915

CONCLUSIONS

Machine learning, a subset of Artificial intelligence can be adopted as a novel technique in geothermal implemented in high temperature hydrothermal fields with three main resistivity zones with the first zone being characterized by a narrow layer of higher resistivity near the surface that represents unaltered region followed by the second broader layer of high conductivity that is interpreted to be due to high conductive hydrothermally altered mineralogy such as zeolites and the third zone being a relatively higher-resistivity zone where resistivity is controlled by the formation of high temperature mineralogy at depth such as epidote.

References

Abrahart, R.J., See L.M. and Solomatine, D. (2008). Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications Water Science and Technology Library 68

Adebayo, A.R., Babalola, L., Hussaini, S.R., Alqubalee, A. and Babu, S.R (2019). Insight into the Pore Characteristics of a Saudi Arabian Tight Gas Sand Reservoir. Energies 2019,12, 4302

Aragón, A.A., Izquierdo-Montalvo, G., Aragón-Gaspar, D.O and Barreto-Rivera, D.N. (2019) Stages of Integrated Geothermal Project. Renewable Geothermal Energy Explorations (2019)

Chaki, S. (2015). Reservoir Characterization: A Machine Learning Approach. *Msc. Thesis, Indian Institute of Technology, Kharagpur*

Daniya, T., Geetha, M & Kumar, K.S. (2020). Classification and Regression Trees with Gini Index. Advances in Mathematics Scientific Journal. 9. 1857-8438. 10.37418/amsj.9.10.53.

Hu, X. (2016). A Coupled Geomechanics and Flow Modeling Study for Multistage Hydraulic Fracturing Of Horizontal Wells in Enhanced Geothermal Systems Applications. *Msc. Thesis,* Colorado School of Mines

Lee, G., Kim, W., Oh, H., Youn, B.D. and Kim, N.H. (2019). Review of statistical model calibration and validation—from the perspective of uncertainty structures. *Struct Multidisc Optim* **60,** 1619–1644 (2019).

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N.C. and Ping, P. (2019) Machine Learning and Integrative Analysis of Biomedical Big Data. *Genes*. 2019; 10(2):87. https://doi.org/10.3390/genes10020087

Mont´ans, F.J., Chinesta, F., Bombarelli, G.R., Kutz, N.J., (2019). Data-driven modeling and learning in science and engineering. Data-Based Engineering Science and Technology Volume 347, Issue 11, pp 845-855

Njiru, N., Opiyo, E., (2018). "Clustering and visualizing the status of child health in kenya: A data mining approach." International Journal of Social Science and Technology I, 2018.

Schmidt, J., Marques, M.R.G., Botti, S. *et al.* Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater* **5,** 83 (2019)

Steuer, F. (2018). Machine Learning for Public Policy Making. How to Use Data- Driven Predictive Modeling for the Social Good Msc. Thesis, Institut Barcelona Estudis Internacionals

Zhu, Z., Lei, X., Xu, N., Shao, D., Jiang, X and Wu, X. (2020). Integration of 3D Geological Modeling and Geothermal Field Analysis for the Evaluation of Geothermal Reserves in the Northwest of Beijing Plain, China. Water 2020,12, 638