ISSN: 0067-2904

New Methodology to Predict Basin or Intrusion from Gravity Data, A Machine Learning Approach.

Ali M. Al-Rahim*, Ahmed A. Al-Rahim
Department of Geology, College of Science, University of Baghdad, Baghdad, Iraq

Received: 22/3/2023 Accepted: 2/6/2023 Published: 30/6/2024

Abstract

Basins and Intrusions structures are essential features in defining and assessing the evolution of tectonic geo-structures. The gravity effects for Basin and Intrusion refer to (such as a salt dome or granitic pluton) structures that are similar in form, shape, and value. Attempts to characterize these structures from gravity data depend on derivation methods such as second horizontal and absolute second horizontal derivative methods. The task of the discriminator is to determine whether the data presented refers to a Basin or Intrusion. Hence, it is just a binary classifier giving the output as 0 (for Basin) or 1 (for Intrusion). The machine learning approach can solve such types of classification with high accuracy and confidence. Machine learning is a field concerned with algorithms that learn from data sets. Classification is a task that requires machine learning algorithms that learn from data sets how to assign a category label to examples from the problem domain. To learn the machine, how to classify the given data into 0 or 1, big data for training is needed. An easy-to-understand example would be classifying gravity data as "Basin, 0" or "Intrusion, 1". Later on, the learned machine can predict any given test data to the state of (0, 1). Therefore, the procedure is simply to prepare a huge synthetic data set (from 2D gravity modeling) for the Basin and Intrusion case. Then, divide the data sets into 80% data for training and 20% for testing. Label this 80% data set with 0 for Basin and 1 for Intrusion. Next, training these 80% data sets using some algorithms specifically designed for binary classification and do not natively support more than two classes. These include Logistic Regression and Support Vector Machines. A confusion matrix is used to evaluate the accuracy of learning. The following step lets the learned machine predict a label for the 20% data set. Python code programming is usually used for this type of analysis. This study uses an orange program for visual programming and data mining for training and predicting. The result of the prediction is perfect for the tested data. Field data for some cases from the Bougure gravity data of Iraq is tested with the learned machine and gives similar results to the absolute second horizontal derivative used. The saved model of the learned machine can be used to predict Basin or Intrusion case studies for future work.

Keywords: Basin or Intrusion, Machine Learning, Logistic Regression, Support Vector Machine.

منهجية جديدة للتنبؤ بالحوض أو الاقحام من بيانات الجاذبية ، نهج التعلم الآلي

علي مكي حسين الرحيم*، احمد علي مكي الرحيم

* Email: alial_rahim@yahoo.com, ali.m@sc.uobaghdad.edu.iq

قسم عام الارض، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

تراكيب الاحواض والاقحامات هي سمات أساسية في تحديد وتقييم تطور الهياكل الجيوتكتونية. تشير قيم الجاذبية للأحواض والاقحامات (قبة الملح او اجسام لصخور نارية) الى تشابه في الشكل والهيئة والقيمة. تجري محاولات توصيف هذه الهياكل من بيانات الجاذبية اعتمادًا على طرق الاشتقاق مثل طرق المشتق الأفقي الثاني والأفقي الثاني المطلق.تتمثل مهمة أداة التمييز في تحديد ما إذا كانت البيانات المقدمة تشير إلى حوض أم اقتحام. ومن ثم ، فهو مجرد مصنف ثنائي يعطي الناتج كـ 0 (للحوض) أو 1 (للاقحام). يمكن لنهج التعلم الآلي أن يحل مثل هذه الأنواع من التصنيف بدقة وثقة عاليتين. التعلم الآلي هو مجال معني بالخوارزميات التي تتعلم من مجموعات البيانات. التصنيف هو مهمة تتطلب استخدام خوارزميات التعلم الآلي التي تتعلم من مجموعات البيانات كيفية تعيين تسمية فئة لأمثلة من مجال المشكلة. من الأمثلة سهلة الفهم هي تصنيف بيانات الجاذبية على أنها "حوض ، 0" أو اقحام ، 1". لتعلم الآلة ، ولتصنيف البيانات المعطاة إلى 0 أو 1 ، هناك حاجة إلى بيانات ضخمة للتدريب. لاحقا ، يمكن للآلة المتدربة أن تتنبأ بأي بيانات اختبار معينة لحالة (0 ، 1). لذلك ، فإن الإجراء هو ببساطة إعداد مجموعة بيانات تركيبية ضخمة (من نمذجة الجاذبية ثنائية الأبعاد) لحالة الحوض والاقحام. بعد ذلك ، قسّم مجموعات البيانات إلى 80٪ بيانات للتدريب و 20٪ للاختبار. تسمى مجموعة البيانات 80٪ هذه بـ 0 للحوض و 1 لاقحام. بعد ذلك ، يتم تدريب مجموعات البيانات هذه بنسبة 80٪ باستخدام بعض الخوارزميات المصممة خصيصًا للتصنيف الثنائي ولا تدعم أصلاً أكثر من فئتين. وتشمل هذه الانحدار اللوجستي ودعم المتجهات. يتم استخدام مصفوفة النتائج لتقييم دقة التعلم. تتمثل الخطوة التالية في السماح للآلة التي تم تعليمها بالتنبؤ بتسمية مجموعة بيانات 20٪. عادةً ما تُستخدم برمجة كود بايثون لهذا النوع من التحليل. تم استخدام البرنامج اورنج للبرمجة المرئية واستخراج البيانات في هذه الدراسة للتدريب والتنبؤ. نتيجة التنبؤ مثالية للبيانات المختبرة. اختبرت بيانات حقلية حقيقية لبعض الحالات من بيانات الجاذبية بوجير للعراق باستخدام الآلة المتعلمة واعطت نتائج مماثلة لتلك المستخدمة في المشتق الأفقي الثاني المطلق. يمكن استخدام النموذج المحفوظ للآلة المتعلمة للتنبؤ بدراسات حالة الحوض أو الاقحام لأي عمل مستقبلي.

1. Introduction

Basins and Intrusions structures are significant features in defining and assessing the evolution of tectonic geo-structures. It is very important to interpret gravity data for hydrocarbon exploration to distinguish between a sedimentary basin (a good possible hydrocarbon prospect) and a granitic pluton (no prospect for hydrocarbons). The gravity effects for Basin and Intrusion structures are similar in form, shape, and value, and both can produce negative gravity anomalies of comparable magnitude. Scientific research indicates multiple cases of failure to explain some negative gravitational anomalies as intrusive bodies of acid igneous rocks, which later turned out to be sedimentary basins [1]. These errors in the interpretation are due to the ambiguity inherent in the gravitational method, whereby objects of varying shapes or dimensions can give similar gravitational anomalies. Mussett and Khan [1] [page 120, Figure 8.17b and illustrated in Figure 1] provided a good example that illustrates such cases, which is called non-uniqueness. Attempts to characterize these structures (Basin or Intrusion) from gravity data depend on derivation methods such as second horizontal [2]. McCann and Till [3] have described how Bott's method can be computerized and show the application of Fourier analysis to the method. Some authors calculate the second horizontal derivative ($\delta^2 g/\delta x^2$), which response exactly like the vertical derivative, except that the maxima and minima are reversed. AL-Rahim and Lima [4] use the absolute second horizontal derivative methods as a criterion to distinguish between (Basin and Intrusion) and apply these criteria to real data from different locations in Iraq. The task of the discriminator

is to determine whether the data presented refers to a Basin or Intrusion. Hence, it is just a binary classifier giving the output as 0 (for Basin) or 1 (for Intrusion). The machine learning approach can solve such types of classification with high accuracy and confidence.

The current study aims to use a data-driven model to predict whether the negative gravity anomaly is related to Basin or Intrusion depending on machine learning approaches. The paper will discuss the machine learning method, workflow procedure, synthetic data generation, short background about the model used in prediction, describe the methodology of using the Orange program as a tool for virtual programming, predict results and finely test the application to real data from Iraq.

2.1 Methodology:

One of the important topics now a day is the Machine Learning (ML) application in different disciplines of scientific approach. Their importance is related to their ability to predict results from the large size of the dataset after examining and training this dataset using newly developed algorithms designed for such purposes, working automatically without human intervention [5]. In ML, most classification problems require predicting a categorical output variable called the target based on one or more input variables called features. The idea is to fit a statistical model that relates a set of features to its target variable and use that model to predict the output of future input observations. Extracting new information from a large data set is one of the most important capabilities in ML for different scientific applications. ML supplies researchers with the tools for discovering new relationships in the huge scientific dataset that is not easily obtained using ordinary methods.

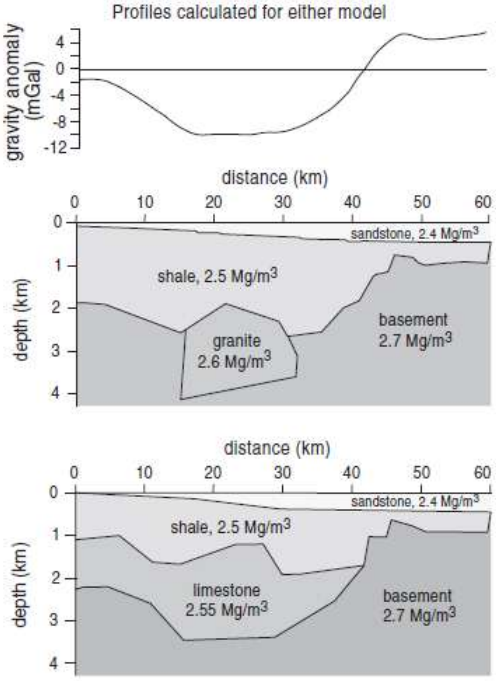

Figure -1 Illustration of the non-uniqueness case for interpreting different geological model that gives the same gravity anomaly effect. (After [1]).

2.2 Types of ML:

Making a data prediction, finding the patterns in data, and/or classifying data are the main machine learning task that usually deals with large size of featured data. A bit different algorithms are used to train these data (learning the machine how to predict specific features

from huge amounts of data). These algorithms determine the machine learning types: Supervised, Unsupervised, and Reinforcement learning (Figure 2).

- Supervised learning: The input is labeled data, and the machine is "supervised" while it's learning, that's means supplying the algorithm with information to help it learn. Remains inputs are the given information used as input features (such as, images, text, series of tabulated data...etc.). The algorithms used during supervised learning include Neural Networks, Decision Trees, Linear Regression, Logistic Regression, and Support Vector Machines.
- Unsupervised learning: doesn't use labeled training sets and data. Instead, the machine search for clear patterns in the data and identify them to make decisions. Common algorithms used in unsupervised learning include Hidden Markov Models, K-means, Hierarchical Clustering, and Gaussian Mixture Models.
- Reinforcement learning: humans learn the closest type. The algorithm learns by interacting with its environment and getting a positive or negative reward. Common algorithms include Temporal Difference, Deep Adversarial Networks, and Q-learning.

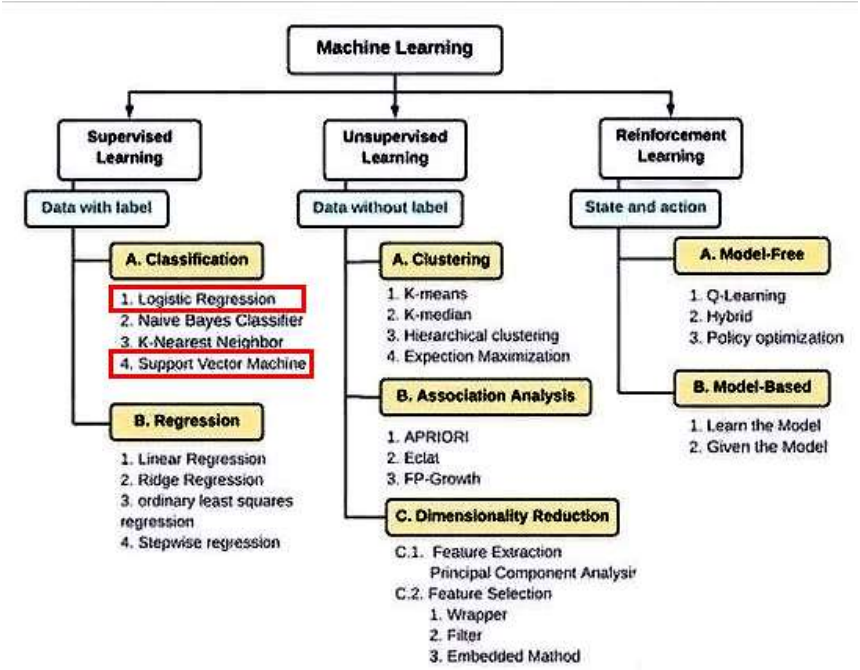

Figure 2: Types of machine learning and it is related to some common algorithms. The red rectangular refers to the algorithms used in the current study.

Bergen et al. [6] (Figure 3) mentioned that most solid Earth geoscience ML applications deal with two types: Supervised Learning and Unsupervised Learning. In supervised learning tasks, such as prediction and classification, the goal is to learn a general model based on known (labelled) examples of the target pattern. In Unsupervised Learning tasks, the goal is instead to learn the data structure, such as sparse or low-dimensional feature representations. Other classes of ML tasks include Semi-Supervised Learning, in which both labeled, and unlabelled data are available to the learning algorithm, and Reinforcement Learning. Deep neural networks represent a class of ML algorithms, including supervised and unsupervised tasks. Deep learning algorithms have been used to learn feature representations, surrogate models for performing fast simulations, and joint probability distributions. It is worth mentioning that deep learning DL is a part of the ML-Neural network method, Figure 4. DL is a promising branch used in fault detection, facies identification, salt and karst detection, and image segmentation. All these are applied to seismic 2D and 3D data.

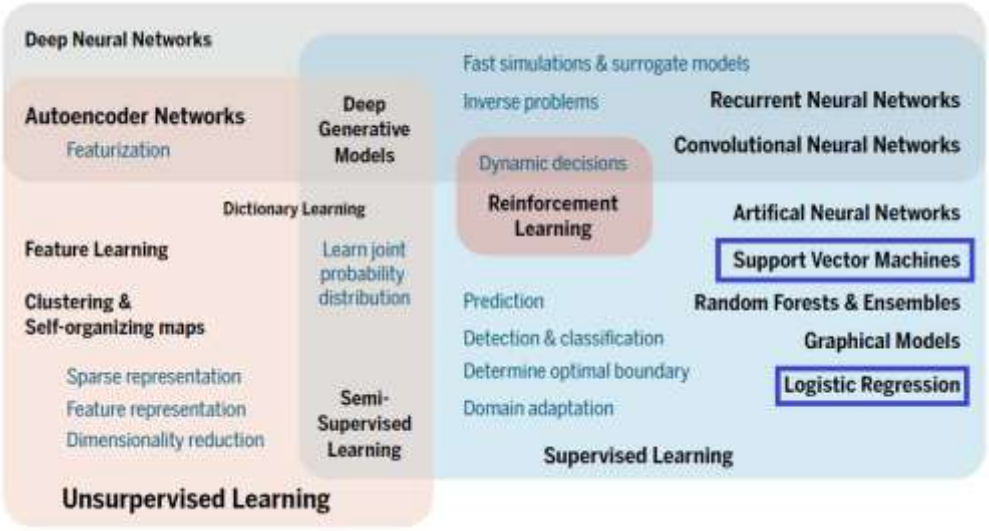

Figure 3: ML methods and their applications in solid Earth geoscience.(After [6]).

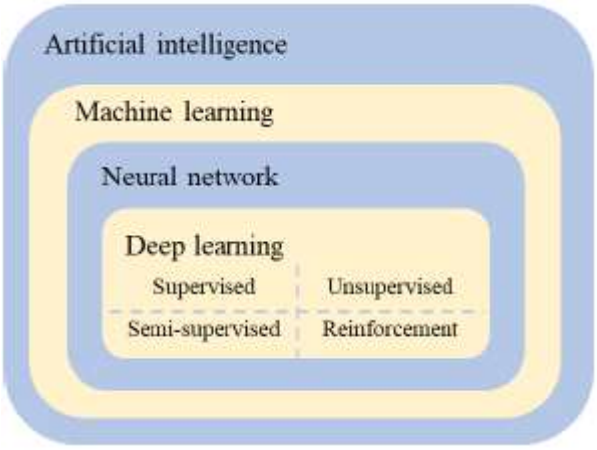

Figure 4: Shows the interference relationship between Artificial Intelligence, Machine Learning, Neural Networks and Deep Learning and the ranking of deep learning approaches. (After [6]).

2.3 Classification types:

Many classification tasks are available; Binary Classification is one of them. This type of classification involves dividing the dataset into two categories. The output variable can only take two values (0 or 1, Y or N, etc.). Figure 5 compares different classification algorithms, their importance and their requirements during the ML code application.

The red rectangular in Figure 2 refers to the algorithms used in the current study: Logistic Regression (LR) and Support Vector Machine (SVM).

- Logistic Regression (LR) is a probability data estimation of one of two categories.
- Support vector machine (SVM) is a binary ranking algorithm that defines the accurate boundary between the training data from two categories. SVMs with linear trends separate classes with an estimated plane, whereas nonlinear trend functions allow for nonlinear decision boundaries between categorized data.

Classification Algorithms Comparison

Source: aisoma.de Algorithm	Primary Problem	Predictors	Power	Raw Implementation	Inter-pretabilty	Regression also	Normalization
k-NN	Multiclass or binary	Numeric	Medium	Easy	Good	No	Required
perceptron	Binary	Numeric	Low	Easy	Good	No	No
Logistic Regr.	Binary	Numeric	Low	Easy	Good	No	No
Linear Discr. Analysis	Binary	Numeric	Low	Medium	Medium	No	No
Naive Bays	Multiclass or binary	Categorical	Medium	Medium	Good	No	Required
Decision Tree	Multiclass or binary	Numeric or categorical	High	Difficult	Good	Yes	No
Random Forest	Multiclass or binary	Numeric or categorical	High	Difficult	Good	Yes	No
Adaboost	Multiclass or binary	Numeric or categorical	High	Medium	Medium	Yes	Usually
SVM	Binary	Numeric or categorical	High	Very Difficult	Medium	No	Yes
Neural Networks	Multiclass or binary	Numeric or categorical	Very High	Very Difficult	Weak	Yes	Yes

AISOMA®

Figure 5: A comparison between different types of classification algorithms, their importance and requirement during applied the ML code [6].

3. Workflow and synthetic data generations:

Figure 6 is an example of a simple ML workflow pipeline. The procedure is divided into three levels. In the data level, to learn the machine how to classify the given data into 0 or 1, big data for training is needed. The synthetic data must be well organized, labeled, and split into trained and evaluated data to test the model. Later on, the learned machine can predict any given test data to the state of (0, 1). Therefore, the procedure is simply to prepare a large synthetic data set (from 2D gravity modeling) for the Basin and Intrusion case. Figure 7 shows examples of the synthetic data created by 2D gravity modelling for Basin and Intrusion structures. See the similar negative gravity effect for both structures. Then, divide the data sets into 80% data for training and 20% for testing. Label this 80% data set with 0 for Basin and 1 for Intrusion. Next, training these 80% data sets using some algorithms specifically designed for binary classification and do not natively support more than two classes. The training algorithm aimed to build a suitable model that could classify the data as 0 or 1. These models include Logistic Regression and Support Vector Machines algorithms. A confusion matrix is used to evaluate the accuracy of learning. The following step lets the learned machine predict a label for the 20% data set. After validating the test model, real data is fed to the model (deployment) to predict their labels which are 0 for Basin and 1 for Intrusion. In this study, the synthetic data is restricted to 100 cases. Fifty cases for each Basin and Intrusion structure. Ninety-four cases are used in training, and six isolated cases for prediction tests.

Python programing language is normally used to execute this workflow of data processing. Newly, the Orange Data Mining program is developed by Bioinformatics Lab at the University of Ljubljana, Slovenia, in collaboration with the open-source community. Orange Data Mining program is a virtual programming tool designed especially for machine learning approach [7] supported with sample workflow for different types of ML applications. The Orange Data Mining program provides widgets for various data management, statistical analysis, visualizations, figures presentations, and ML models for training and predicting data for many science disciplines. Figure 8 shows the processing workflow used in the current study in the Orange Data Mining Program. Figure 9 represents the trained data for the negative gravity effect for both Basin and Intrusion structures. See the overlap and

interference of the profiles, which are regarded as a feature in training and labelled as 0 for Basin and 1 for Intrusion as target data.

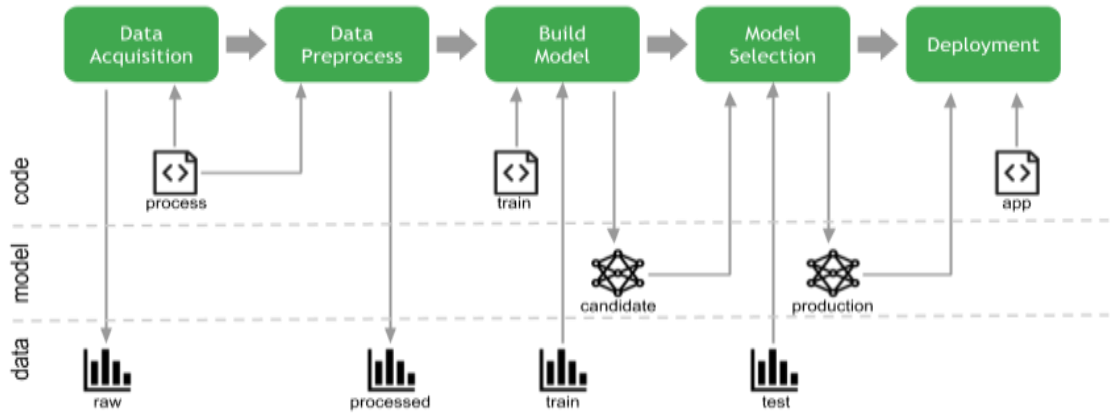

Figure 6: Shows an example for simple ML workflow pipeline.

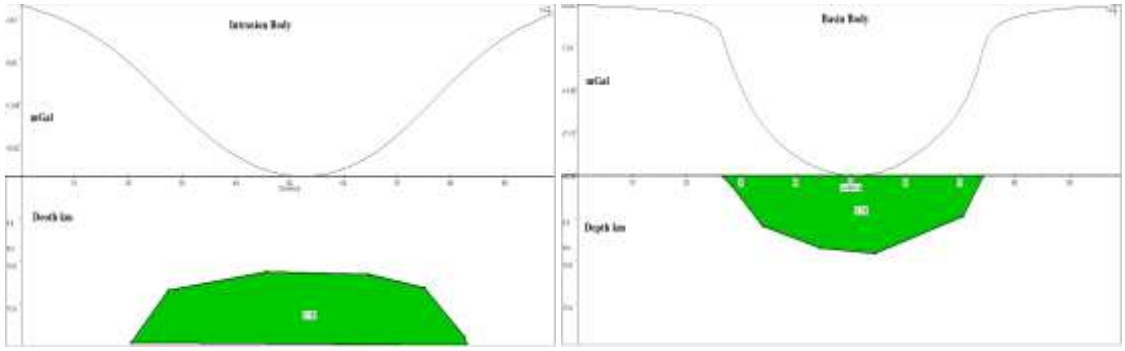

Figure 7: Examples of the used synthetic data created by 2D gravity modelling for Basin and Intrusion structures. See the similar negative gravity effect for both structures.

4. Results:

The statistical results for LR and SVM models can be seen using the confusion matrix (see Figure 8). Figure 10 shows the confusion matrix result for Logistic Regression LR and SVM models. The LR correctly predicted 47 cases for both Basin and Intrusion with 100%. Whilst SVM predict all cases for the Intrusion and 45 cases for the Basin. 2 cases for Basin are predicted by the SVM model as Intrusion, which is 95.9% of the correct result. Both models give highly accurate results. The last step is to test the remaining six cases and indicates whether their results' prediction was correct. Actually, the prediction results are perfect for both Basin and Intrusion cases.

Taken the real gravity data used by [4] for two profiles across a wide circular shape anomaly located in the middle part of the stable shelf and one profile across a depression area (Maa'niyah depression) that is located at the border with Saudi Arabia to predict whether these are related to Basin or Intrusion. Both LR and SVM predict that the circular shape anomaly is related to Intrusion and Maa'niyah depression related to Basin structures which are entirely comparable results.

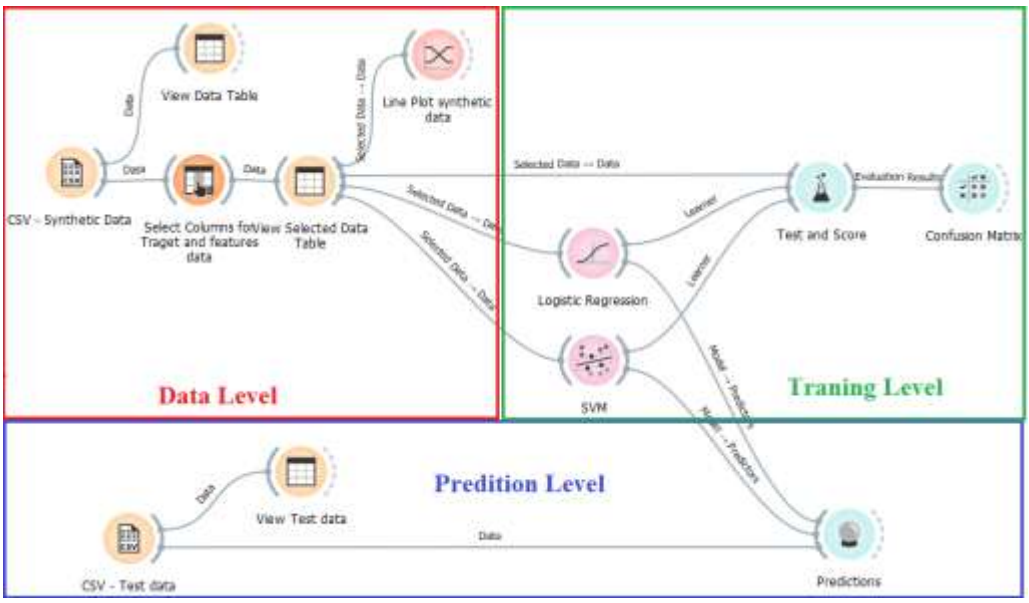

Figure -8 Shows the workflow of processing in Orange Data Mining Program.

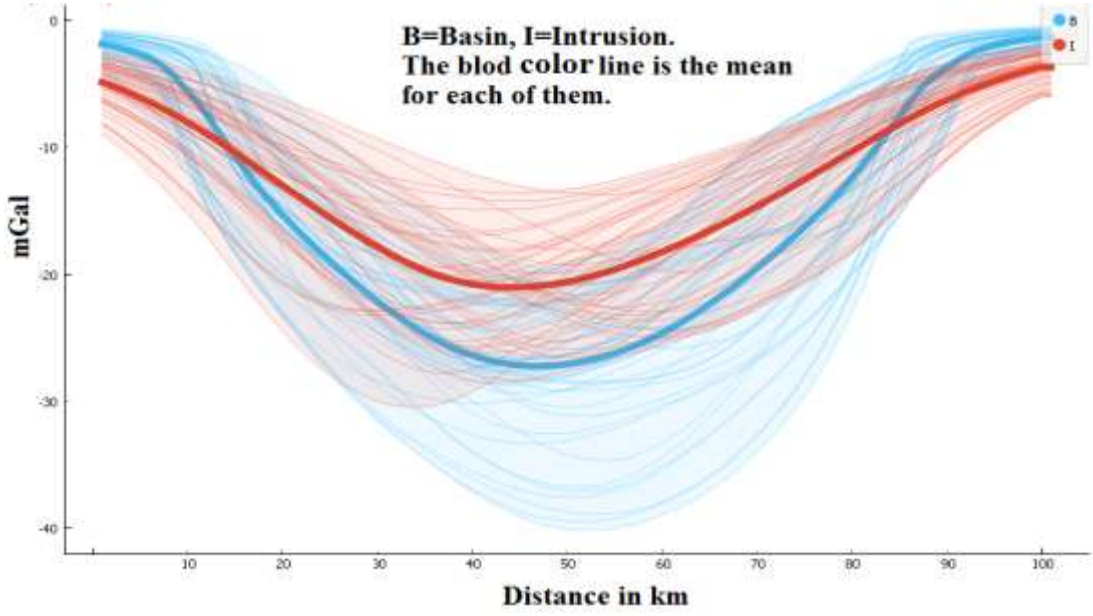

Figure -9 Represent the trained data for the negative gravity effect for both Basin and Intrusion structures. See the overlap and interference of the profiles which regarded as a feature in the training and labelled as 0 for Basin and 1 for Intrusion as a target data.

Logistic Regression

Actual \ Prediction	B	I	Σ
B	47	0	47
I	0	47	47
Σ	47	47	94

B=Basin, I=Intrusion.

(a)

SVM

Actual \ Prediction	B	I	Σ
B	45	2	47
I	0	47	47
Σ	45	49	94

B=Basin, I=Intrusion.

(b)

Figure -10 The confusion matrix result for a) Logistic Regression LR and b) SVM models. The LR correctly predicted 47 cases for both Basin and Intrusion with 100%. Whilst SVM predict all cases for the Intrusion and 45 cases for the Basin. 2 cases for Basin are predicted by the SVM model as Intrusion, which is 95.9% of the correct result. Both models give highly accurate results.

5. Conclusion:

The current study is a new method to predict whether the similar effect of negative gravity anomalies deduced from Basin or Intrusion using the ML classification approach. It gives a decisive decision about the type of object causing the negative gravitational effect, reduces the non-uniqueness in the interpretation of gravity data and reduces the illness of gravity inversion. Models deduced from training can be saved and used to predict any other gravity profiles and ML transfer learning method. Indeed, using more synthetic data and testing another model, like deep learning, is the next step in developing the technique for spherical, cylindrical, and prismatic bodies.

6. Conflicts of Interest: The authors declare no conflict of interest.

Reference:

[1] Reference:Mussett, A. E., and Khan, M. A. *Looking into the Earth an introduction to geological geophysics*. *Cambridge University press*, p.494, 2009.

[2] Bott, M. P. H. "A simple criterion for interpreting negative gravity anomalies", *Geophysics*, vol.27, no.3, pp. 376-381, 1962.

[3] McCann, C., Till, R. "The use of interactive computing in teaching geology and geophysics", *Computers & Geosciences*, vol.2, no.1, pp.59-67, 1976. https://doi.org/10.1016/0098-3004(76)90093-5.

[4] Al-Rahim, Ali M. and Williams A. Lima. "Basin or Intrusion, a New Method to Resolve Non-Uniqueness in Gravity Interpretation". *Iraqi Journal of Science*, vol. 57, no. 1B, pp. 408-491, 2016.

[5] Sen, P.C., Hajra, M., Ghosh, M. Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal, J., Bhattacharya, D. (eds) *Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing*, vol 937. Springer, Singapore, 2020. https://doi.org/10.1007/978-981-13-7403-6_11.

[6] Bergen K. J., Paul A. Johnson, Maarten V. de Hoop and Gregory C. Beroza. "Machine learning for data-driven discovery in solid Earth geoscience". *Science*, 363 (6433), eaau0323,2019. DOI: 10.1126/science.aau0323.

[7] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B. "Orange: Data Mining Toolbox in Python", *Journal of Machine Learning Research* 14(Aug), pp. 2349−2353m 2013.