Suitability of data repurposing - assessing sedimentological data as predictors for gold grade estimation in the Witwatersrand Goldfields

T. Mombe[1], G.T. Nwaila[1], and S.E. Zhang[1,2]

[1]University of the Witwatersrand, South Africa
[2] Geological Survey of Canada, Canada

Machine learning has only recently proven relevant in geoscience, with several studies alluding to it being a revolutionary resource estimation tool capable of significantly enhancing prediction accuracies. This paper demonstrates that machine learning algorithms can generate highly accurate predictive models for gold grade estimation. However, the extent to which sedimentological data are useful in this regard relies heavily on the quality of the data. This paper reports on the results of data repurposing in resource estimation to assess the suitability of legacy sedimentological data for the task of in-situ gold grade using South Africa's Witwatersrand goldfield. Analysing the differences between the 'good' and 'bad' datasets revealed a prominent data quality issue that may have resulted from errors during sampling. Overall, this paper produces a resource estimation methodology that can be replicated and used on any Witwatersrand-type gold deposits.

INTRODUCTION

Resource estimation, as a form of spatial data modelling, is the foundation upon which all mine plans are built. Traditionally, most studies (Rahimi *et al*., 2018; Marwanza *et al*., 2018) have taken a purely geostatistical approach, with most methodologies entailing some form of kriging (Zhang *et al*., 2021). These geostatistical methods typically assume data retrieved from small samples to be representative of large ore deposits – under the assumption of spatial autocorrelation of regionalised variables (Kaplan and Topal, 2020). The suitability of certain geostatistical techniques for in-situ metal grade estimation is, however, contextual and application specific (Lindsay *et al*., 2021). Furthermore, a trade-off exists between sample size, cost, and computation time (Zhang *et al*., 2021). These estimation processes also require considerable manual data manipulation, which increases the possibility of unforeseeable errors and analytical partiality (Zhang *et al*., 2021; Kaplan and Topal, 2020). The last few decades have seen a steady introduction of machine learning (ML) across various industries such as manufacturing, healthcare, and finance; mining operations are among the latest (Jung and Choi, 2021). This paper documents investigations of poor model performance of a sedimentology-based gold grade prediction project in the Witwatersrand Basin (South Africa). Our findings could theoretically be applied to other multivariate or multidimensional data types. A set of machine learning algorithms (MLAs) were implemented to understand the suitability of the sedimentological data as covariates of the gold grade, which in turn, permitted us to analyse model performance. We determined that a large fraction of the dataset is unsuitable for ML-based grade estimation using sedimentological covariates as features. This is because such data seems to have been sampled primarily near or off-reef, which is not an issue for grade estimation using geostatistics, but is a data quality issue for ML-based grade estimation.

Our study demonstrates the necessity to understand data quality under data repurposing, which is the current dominant supply of data for ML projects in geosciences and mining, and specifically to adopt a modern data science-type of project management model that caters for characteristics of data repurposing.

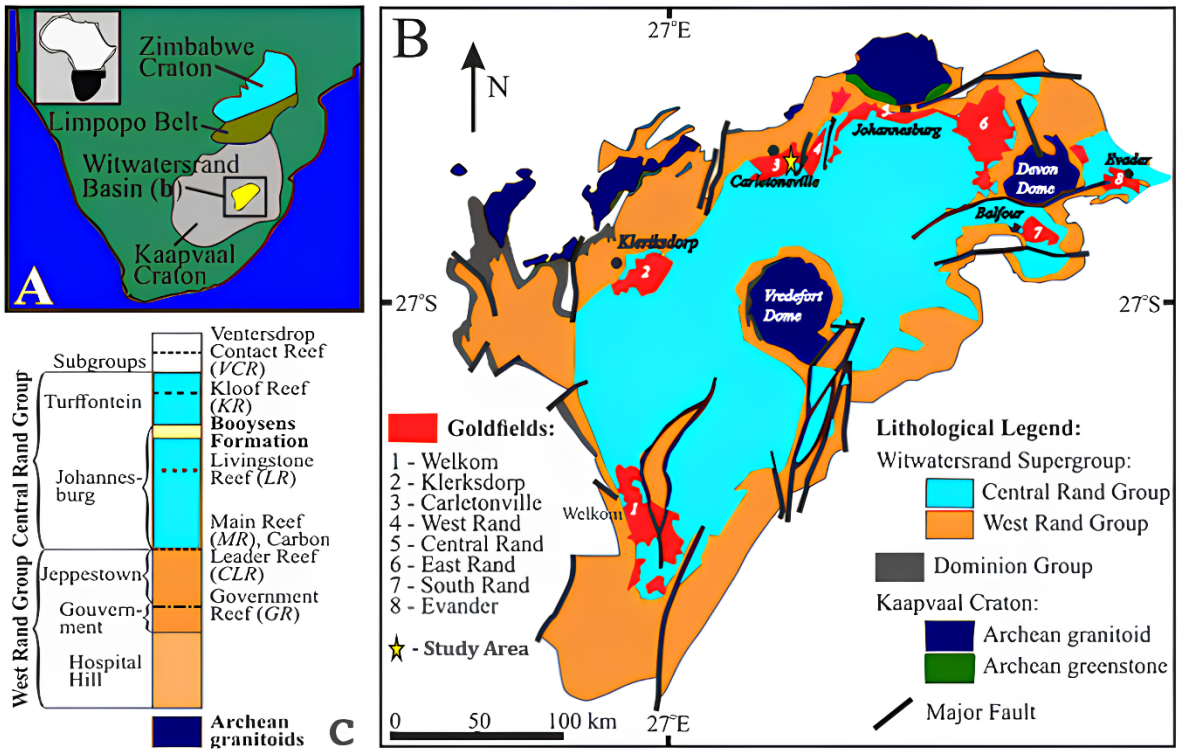

Figure 1. Locality map of the study area. Included is the location of the Witwatersrand Basin (A); its goldfields (B); and its stratigraphic column (C) with the position of the thin Kloof Reef outlined (adapted from McCarthy, 2006; Frimmel and Nwaila, 2020).

The data used in this project originates from Sibanye-Stillwater's Kloof gold mine, which extracts from the Kloof Reef of the Witwatersrand Basin and is located in South Africa's Gauteng Province (Frimmel and Nwaila, 2020; Nwaila *et al.*, 2020) (Figure 1). The ~2.9 Ga intracratonic Witwatersrand Supergroup is located at the centre of the Archean Kaapvaal Craton (Figure 1A; Robb and Meyer, 1995; Frimmel *et al.*, 2009). The emplacement of the Bushveld Complex (~2.06 Ga), in conjunction with the Vredefort meteorite impact (~2.02 Ga), caused a series of low-grade metamorphic events that prompted fluid flow through the basin's units and consequently contributed to gold remobilisation (Safonov and Prokof'ev, 2006; Phillips and Powell, 2011). The sedimentary characteristics are thought to have a significant control on gold deposition and were found to be covariates of the gold grade (e.g., Frimmel and Nwaila, 2020; Zhang *et al.*, 2021; Frimmel, 2023). The basin's gold-bearing conglomerate reefs are mined in eight principal goldfields (Figure 1B) and a few minor occurrences. The Kloof Reef is situated amidst the Central Rand Group sediments (Figure 1C; Frimmel *et al.*, 2009). These are primarily characterised by an arenaceous sequence that includes quartzites, quartz-pebble conglomerates, minor shales and quartzwackes (Frimmel *et al.*, 2009; Tucker *et al.*, 2016). The Central Rand Group comprises the bulk of the Witwatersrand Basin's gold mineralisation (Frimmel *et al.*, 2009). The mineralisation occurs as submicroscopic crystalline grains within detrital pyrite grains; some gold grains have been described as rounded or flattened flakes (Tucker *et al.*, 2016).

MACHINE LEARNING, DATA QUALITY AND PROJECT MANAGEMENT

ML allows a computer to learn a specific task without being explicitly programmed (Cate *et al.*, 2017). Depending on the problem, data-driven ML tasks are solved using either supervised or unsupervised

learning techniques. Both supervised and unsupervised MLAs follow a general methodology that entails data pre-processing, training, testing and application (Cate *et al.*, 2017). Data collection and preparation precede any modelling (Hall *et al.*, 2019). In the case of data repurposing, data collection would have been guided by a former, different purpose (e.g., geostatistical modelling) than for ML-based tasks. Repurposing data for ML-based tasks increases the already high-risk nature of data science, because: (1) a substantial amount of work must be completed before project feasibility can be determined; and (2) data is not specified to requirements of the task (e.g., unknown predictive power of covariates). Consequently, project management models for data science projects must cater for high-risk experimentation, such as the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman *et al.*, 1999). However, such models are not used in traditional geoscience projects. To determine the quality of data for ML-based tasks, data is explored through exploratory data analysis, predictive modelling and performance assessment (Wickham and Grolemund, 2017; Hall *et al.*, 2019). The data understanding, model assembly and model audit phases of CRISP-DM can be iterated until a model with the best predictive performance is obtained (Hall *et al.*, 2019). This high-performance model can then be deployed once testing is complete and the necessary documentation has been developed (Wickham and Grolemund, 2017; Hall *et al.*, 2019).

METHODOLOGY

Data Pre-processing

The dataset comprised the following predictive features (sedimentological properties): the Channel Width and Internal Waste, whose properties are described in centimetres (cm); the Percentage Conglomerate, which is expressed as a percentage of the volume of the sample that is estimated to be of conglomerate-type lithology; and the Basal Contact, which is a categorical property that describes the nature of the contact between the sample and the underlying lithology. The data label, Au (g/t), describes the target property in grams per ton (gold grade) as determined through fire assay. The Channel Width and Basal Contact are human-derived (qualitative) features, while the Internal Waste and Percentage Conglomerate are derived through a calculation. The first iteration of the CRISP-DM was initiated by analysing and cleaning the dataset to ensure it was in the most desirable form for predictive modelling. The dataset was filtered to exclude points with values equal to 0 for the target gold feature and any missing values. Categorical features (i.e., Basal Contact) were encoded into qualitative integers. Furthermore, the data's distribution and various statistical summaries were subsequently investigated - this included creating visualisations (Figure 2) that give us more information on the data from a large-scale and small-scale point of view. In addition, the dataset was filtered to remove outliers, such that any points that fell outside the upper and lower limits of two percentile-based thresholds were removed while ensuring that > 90% of the original data were retained. To ensure that all features would contribute equally during model fitting, the features were rescaled by removing the mean and scaling to unit variance:

$$y = \frac{(x - mean)}{std} \quad [1]$$

where y is the rescaled feature, x is the original, and std is the feature's standard deviation. All predictive modelling steps were performed with rescaled predictive features while the target feature remained unscaled.

Predictive Modelling

The MLAs used in this study are: Random Forest (RF), AdaBoost, K-Nearest Neighbour (KNN) and Elastic Net (EN). Tree-based methods are desirable for grade estimation because no data transformation is necessary (Nwaila *et al.*, 2022). RF is a non-parametric algorithm that uses an ensemble of decision trees (Breiman, 2001). Decision trees are recursively partitioned, hierarchical and branching structures. Branch nodes represent features, while branches represent decision rules, and each leaf represents an outcome. Partitioning of the data is based on feature values. The splitting criterion of a node is metric-driven to maximise the difference between the resulting leaves. For regression, the mean squared error metric is used to measure the goodness of fit of the leaf to its surrounding samples. The depth of the

tree is a hyperparameter. A forest consists of a number of de-correlated trees, each of which is built using bootstrap sampled features. Subsequently, the output is averaged from the results of the entire forest (Breiman, 2001). This ensembling process converts a weak learner (decision trees) into a strong one, which lowers output variance without increasing model bias. AdaBoost is a boosted (and also ensemble) non-parametric algorithm that builds a series of trees, each of which attempts to mitigate the errors associated with the previous tree in the series by adjusting the weights of each tree's output (Koduri *et al.*, 2019). KNN is a non-parametric algorithm that predicts labels of an unknown target by averaging over 'K' number of labels of the closest neighbours in feature-space (Novitasari *et al.*, 2019). The EN regression is a linear parametric algorithm that uses a penalised least-squares estimation technique (Liu and Li, 2017). It uses a combination of the ridge (L_2 penalty) and the LASSO (L_1 penalty) regression methods (Liu and Li, 2017).

The dataset was split into a training and testing set, with 30% being reserved for testing, and the remaining 70% for training. Model selection was performed using grid search; see Table 1. Algorithm selection was performed to rank MLAs, and cross-validation was performed using the 10-fold method to measure model performance (Kapageridis, 2005). Each fold acts as the training set K – 1 times and the testing set 1 time. The coefficient of determination (CoD) or R^2 score was the metric used to assess performance. Multiple passes of the CRISP-DM were used to iterate the models. We also explored the effects of partitioning the dataset by feature similarity, which used the K-means clustering algorithm (Nainggolan *et al.*, 2019), combined with the elbow method. Predictions were thereafter performed on each cluster using a cluster-specific model.

Table I. Model parameters used in the grid search

Algorithm	Parameter Grid
RF	max depth = {80, 90, 100, 110}, max features = {2, 3, 4}, min samples per leaf = {3, 4, 5}, min samples per split = {8, 10, 12}, number of trees = {100, 200, 300, 1000}
EN	α = np.logspace(-5, 2, 8), L_1 ratio = {.2, .4, .6, .8}
KNN	number of neighbours = {2, 3, 4, 5, 6, 7, 8, 9}
AdaBoost	learning rate = {.001, 0.01, .1}, number of classifiers = {500, 1000, 2000} , base classifier = decision tree with maximum depth = {80, 90, 100, 110}

Algorithm Performance and Data Investigations

Once the first and second CRISP-DM iterations were completed, further attempts were conducted to investigate the data and attain improved prediction results – this included residual analysis: the dataset's residuals were calculated by subtracting the predicted target (gold grade) values from the original target values (y), and the relative errors (RE) were thereafter computed:

$$RE = \left|\frac{Residuals}{y}\right| \quad [2]$$

The dataset was then evaluated in sets according to its relative errors – that is, data points were subset using a criterion of $\Delta RE = 0.1$. Subsequently, predictive modelling was performed on each sub-dataset, and the phased prediction results were collected. This process was then used to isolate datapoints that improve the prediction performance and those that do not. ***Another post-prediction investigation was feature selection:*** In this step, the predictive features were removed one by one, and the predictive modelling process was repeated with the remaining features. This was to assess if the prediction scores/accuracy improved when a certain sedimentological feature was missing or whether its absence lowered the scores. Finally, feature importance statistics were computed using the tree-based algorithms' best parameters (RF and AdaBoost). Each predictive feature was analysed to determine how greatly it affects the predictive quality of the overall model using an information gain metric called the Gini importance. These statistics were visualised graphically and assessed to understand which sedimentological features were considered most important for gold grade prediction.

RESULTS

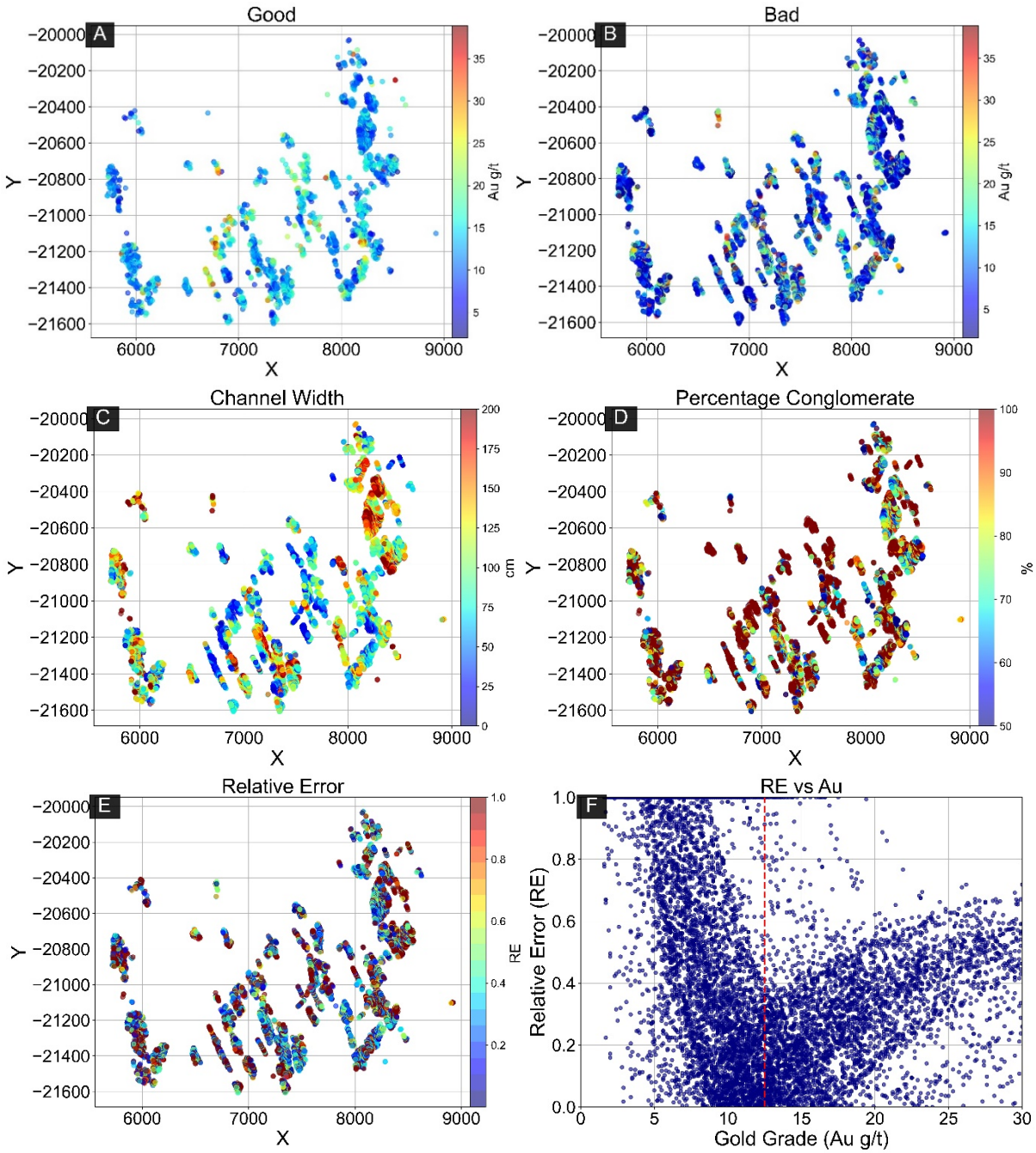

Figure 2. Spatial maps illustrating the dataset's distribution, colour-coded by gold grade for the 'good' (A) and 'bad' (B) datasets. In (C) and (D) the full dataset is colour-coded by the Channel Width and Percentage Conglomerate features. (E) illustrates the relative error (i.e., good vs bad) distribution, while (F) illustrates the relationship between the relative error and gold grade for the full dataset.

The dataset comprised 10 682 datapoints after pre-processing. The first CRISP-DM iteration returned poor results, with test CoD scores of 0.02, -0.02, 0.04 and 0.04 for the RF, KNN, EN and AdaBoost algorithms, respectively (i.e., 2%, -2%, 4% and 4% accuracy). These results prompted analysis of the dataset using subsetting by RE score; predictive modelling was thereafter conducted in intervals, i.e., the RE = 0 to 0.1 column refers to datapoints with RE between 0 and 0.1 from the first CRISP-DM iteration (Table 2). Through experimentation, we decided to examine the contrast between a best-

predicting fraction of the dataset ('good' dataset - RE between 0 and 0.2) and the remainder ('bad' dataset); Figure 2E illustrates the spatial positions of these subsets.

Table II. CoD test scores according to relative error intervals

Test CoD values for each RE subset											
RE Interval	**0 to 0.1**	**0.1 to 0.2**	**0. to 0.3**	**0.3 to 0.4**	**0.4 to 0.5**	**0.5 to 0.6**	**0.6 to 0.7**	**0.7 to 0.8**	**0.8 to 0.9**	**0.9 to 1**	**>1.0**
RF	0.87	0.61	0.48	0.29	0.12	0.02	-0.05	0.62	0.72	0.52	0.28
KNN	0.87	0.58	0.35	0.37	0.01	0.09	-0.08	0.58	0.77	0.41	0.27
EN	0.72	0.43	0.32	0.31	0.14	0.04	0.00	0.28	0.64	0.62	0.16
AdaBoost	0.80	0.55	0.41	0.41	0.09	0.03	-0.07	0.75	0.86	0.65	0.32

The 'Good' Dataset

The well-predicting portion of the data comprised 2325 datapoints (21.8% of the dataset). This subset returned accuracy scores of 84%, 80%, 81%, and 59% for the RF, KNN, EN and AdaBoost algorithms, respectively (Figure 3). To further understand the predictive power of the covariates, feature selection was implemented, and the predictive modelling phase was repeated with the remaining features. On average, prediction accuracies improved when the Conglomerate Percentage or the Internal Waste features were removed (Table 3). The RF's CoD score improved by 1% when either feature was dropped, while the KNN's score improved by 3% and 1% when the Conglomerate Percentage and Internal Waste features were removed. Removing the Basal Contact feature reduced the prediction accuracies by about 10%, on average, while removing the Channel Width reduced them by about 20% (Table 3). Removing both the Internal Waste and Conglomerate Percentage features resulted in a significant drop in model performance compared to when only one feature was missing. Generating feature importance statistics for the RF algorithm (Figure 5) revealed that it considered the Conglomerate Percentage the most important feature. Furthermore, the Channel Width and Internal Waste features were roughly equal in importance, while the Basal Contact was considered insignificant. The AdaBoost algorithm considered the Conglomerate Percentage most important, followed by the Internal Waste.

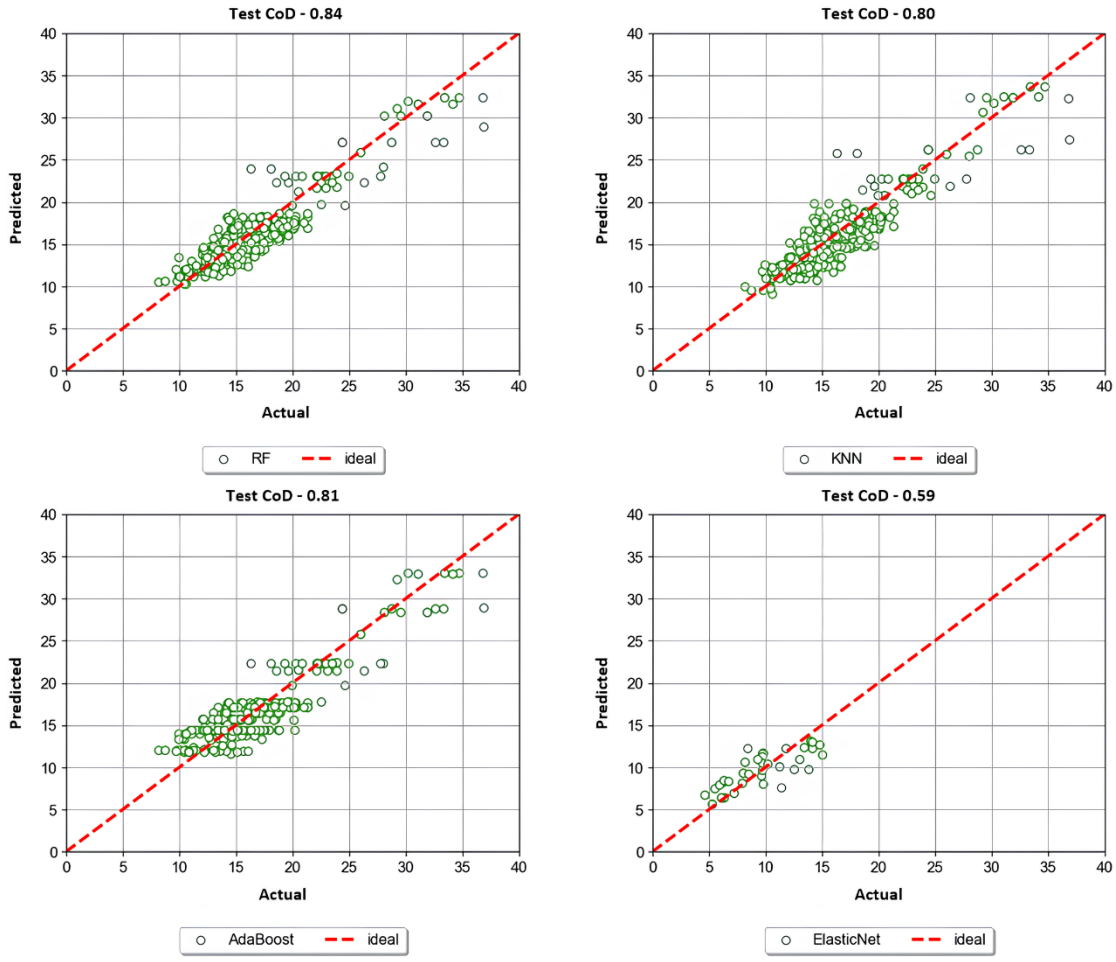

Figure 3. Graphical illustrations of the 'good' dataset's test scores post-clustering: Random Forest = 84 %, K-Nearest Neighbours = 80 %, AdaBoost = 81 % and Elastic Net = 59 %.

The 'Bad' Dataset

The poor-predicting portion of the data comprised 8357 datapoints. This subset returned accuracy scores of 2%, -1%, 0%, and 2% for the RF, KNN, EN and AdaBoost algorithms, respectively (Figure 4). Feature selection did not noticeably improve or reduce the prediction scores. The differences are within 1% to 2%, on average, for most of the algorithms. However, removing both the Internal Waste and Conglomerate Percentage features resulted in a consistent drop in model performance, consistent with the result obtained with the 'good' dataset. The RF algorithm considered the Channel Width as most important, while the Internal Waste and Conglomerate Percentage had lower importance (Figure 5). The AdaBoost algorithm also considered the Channel Width the most important feature, and the Internal Waste came second. Meanwhile, the Basal Contact feature was not considered for both datasets.

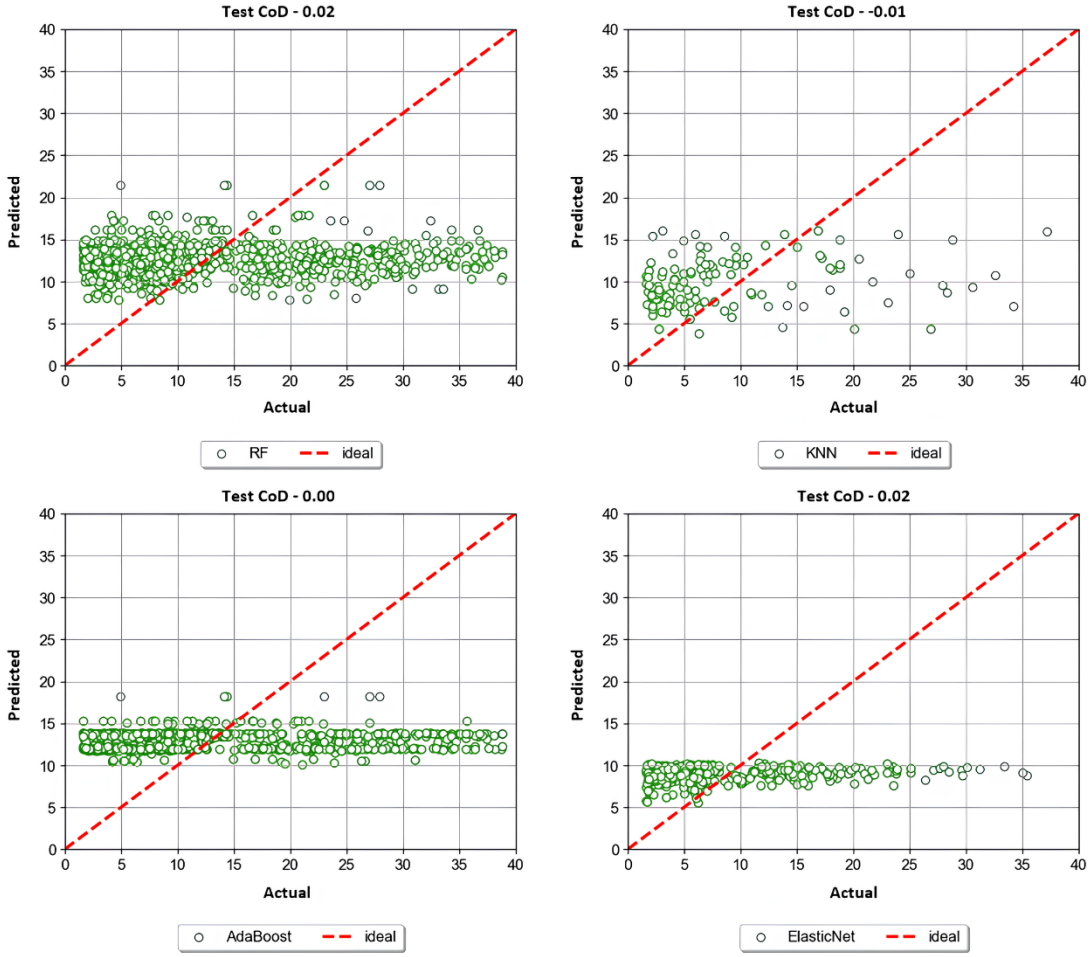

Figure 4. Graphical illustrations of the 'bad' dataset's test scores post-clustering: Random Forest = 2 %, K-Nearest Neighbours = -1 %, AdaBoost = 0 % and Elastic Net = 2 %.

DISCUSSION

Filtering the dataset using the RE metric proved to be productive because it enabled delineation of data quality through predictive modelling performance (the 'good' and 'bad' datasets, characterised by prediction performance). The 'good' dataset produced high accuracy scores (Figure 3), with the highest being 79% (RF) prior to clustering. Further predicting the 'good' dataset using clustering proved beneficial because it boosted its accuracy score to 84% (RF). However, per-cluster prediction did not significantly influence the 'bad' dataset's prediction scores. The feature importance results (Figure 5) show that the most important features for the 'good' dataset are the Conglomerate Percentage and the Internal Waste. In contrast, the 'good' dataset's feature selection results (Table 3) demonstrate that the Conglomerate Percentage and Internal Waste features are also responsible for poor predictions since prediction accuracies remained high or increased when either feature was removed. In general, this could be the result of interactions with other features. In this case, the Conglomerate Percentage and Internal Waste are, by design, complementary features in the sense that the portion of the orebody that is not conglomerate is considered Internal Waste. Hence, at least in principle, having one or the other of Internal Waste and Conglomerate Percentage should be sufficient for predictive modelling - this reduces feature space dimensionality and can therefore improve predictive modelling performance. Nevertheless, it seems that points that predicted poorly within the 'good' dataset were a result of noisy/incorrect information recorded in the Conglomerate Percentage and Internal Waste.

Table III. 'Good' dataset's feature selection test scores

Test CoD (30% Test Size)				
Dropped Feature	**RF**	**KNN**	**EN**	**AdaBoost**
Basal Contact	0.69	0.68	0.46	0.65
Conglomerate Percentage	0.85	0.84	0.52	0.81
Internal Waste	0.85	0.82	0.48	0.81
Channel Width	0.59	0.52	0.45	0.61
Internal Waste and Conglomerate %	0.58	0.56	0.41	0.55

There is a consistent difference between the feature importance statistics for the 'good' and 'bad' datasets from both the RF and AdaBoost algorithms. However, the AdaBoost algorithm consistently categorised Internal Waste as less important; this may again allude to this feature being noisy. The boosting function in the AdaBoost algorithm uses a predictor-corrector approach to ensembling, which makes the AdaBoost algorithm very sensitive to noise (Koduri *et al*., 2019). This implies that there is a structural difference between the 'good' and 'bad' datasets, which could be a recording and data entry consistency issue, or a sample issue (including sample characteristics and sampling methods). It seems that at least some of the difference must be at the sample level, since there is a clear spatial association and gold grade difference between the 'good' and 'bad' datasets (Figure 2A and B**)**. The 'bad' dataset's samples generally contained lower gold grades and were distributed more on the periphery of spatial groups of datapoints (Figure 2E). The Basal Contact feature is categorised as least important in all feature importance statistics; however, when it was removed in the feature selection step, the prediction scores significantly lowered. It can therefore be assumed that the Basal Contact has an inconsistent control on gold occurrence, maybe through interactions with other sedimentological attributes. If this behaviour is not an artefact of sampling, then it is a physical property of the samples. Yang *et al*., (2021) conducted research that found that gold mineralisation was significantly concentrated on lithological contacts, while Jolley *et al*., (2004) encountered gold mineralisation associated with contacts in various parts of the Witwatersrand Basin. Overall, sedimentological properties have proved significant for gold grade prediction after data quality-based delineation. In addition, the tree-based algorithms (RF and AdaBoost) seem to produce the best results for the dataset used in this research.

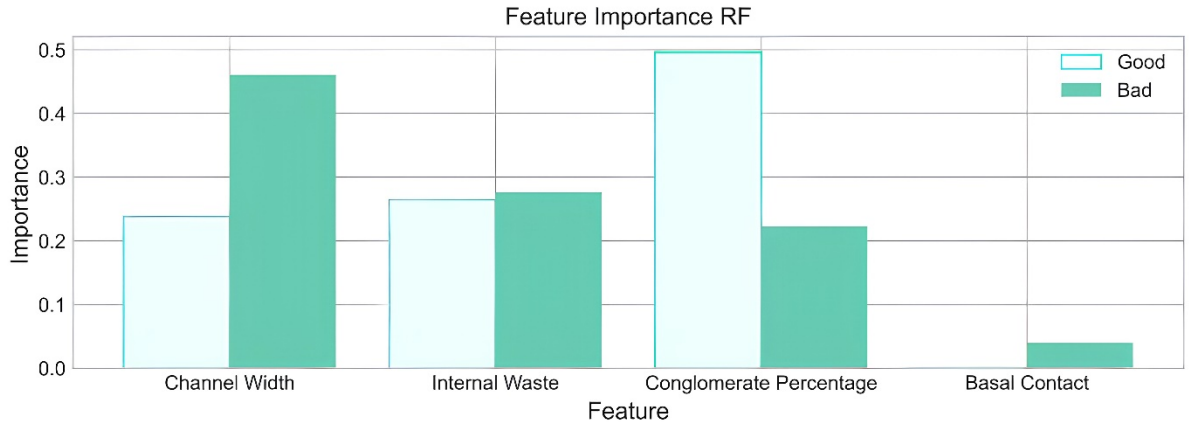

Figure 5. Graphical illustrations of the RF algorithm's feature importance statistics for the 'good' and 'bad' datasets.

Importance and reliability of sedimentological data in resource estimation

This study has reiterated that multidimensional ancillary data can be valuable for resource estimation. The feature importance analysis suggests that the sedimentological properties are useful for gold grade prediction but not uniformly throughout the dataset. Moreover, the feature selection step revealed that excluding certain features (i.e., Channel Width) significantly lowers prediction accuracies across most MLAs. Subsequently, it can be assumed that a larger set of sedimentological properties would provide

a strong database for resource estimation and most likely result in higher model performance. The reliability of using sedimentological data for resource estimation is once again highly dependent on the quality of the geological database. Although sedimentological properties improve model predictions on average, the extensive quality assurance and quality control (QA/QC) process that goes into cleaning the dataset significantly reduces its size, as seen in this research (only 21.8% of the data produced good predictions). Since the dataset was originally designed for geostatistics, only the gold grade data is of high quality. In our case, because we repurposed the data, the quality of the predictive features was important but far below the quality of the Au grade. This emphasises the need to engineer data properly for modern uses such as ML, whose needs and considerations differ from traditional techniques, such as geostatistics.

Limitations and recommendations

The primary challenge experienced during the course of this study was the quality of the data provided for predictive modelling in the context of data repurposing. The initial poor prediction results prompted the development of a complex QA/QC workflow to investigate and eventually extract well-predicting datapoints. This, however, proved to be a very productive tangential adventure, which often happens in data-driven research. As we have demonstrated, data reconstruction using ML is a very effective method to detect poor-quality data, which in this case, exhibited clear and consistent trends (e.g., that they are all of low grade). Consequently, only 22% of the data produced good results – this was another limitation because only a small portion of the data could be salvaged, and the gold grades predicted may not be an accurate reflection of the larger deposit because of the loss of data coverage in the spatial domain. To aid the integration of ML with geoscience, ML techniques should be integrated into modern mining software, ensuring accessibility and ease of use. However, because of the number of artificial intelligence algorithms that may be suitable for mining applications, the more practical approach would be for mining companies to adopt trans-disciplinary expertise alongside their traditional teams. This ensures that all data gathered are engineered to purpose. Studies (e.g., Kapageridis *et al.*, 1999) have already attempted to build ML estimation software that only requires an assay database; the software then automatically performs analysis and makes predictions. Ideally, once resource estimation is complete, ML techniques should also be able to use the generated model/estimates to create block models, determine cut-off grades and perhaps design mining sequences. A benefit of this approach would be the potential to more completely utilise the data, which in our case, involved the use of ancillary data. Such data are of less or insignificant value in traditional geostatistical workflows.

CONCLUSIONS

This study demonstrates the need to manage data-driven projects with an adequate and specific model, which in this case was the CRISP-DM model. These models recognise the importance of iteration, the high-risk nature and tangential outcomes typical of geodata science projects. In our case, the project evolved in a tangential direction because the supplied data does not strictly meet quality standards under data repurposing. Consequently, a series of experiments were conducted to understand the structure of the data and investigate the origin of data quality issues. Thereafter, we developed a methodology to differentiate between good- and poor-quality data and managed to generate highly accurate ML models. We demonstrated that ancillary data is useful in resource estimation, but only with high-quality data. This study has revealed that data quality considerations of trans-disciplinary approaches are vastly different from traditional approaches. It is thus important to engineer data to meet rapidly changing data usage patterns. The sedimentological properties did not necessarily have a strong inferential relationship with the gold grade within the entire dataset. However, the feature importance and selection stages proved that some sedimentological properties, especially Channel Width, are highly effective in predicting in-situ gold grade, while the Conglomerate Percentage/Internal Waste are negligible. The strongest association, as proxied by the highest prediction performance, is associated with samples that are more central in each spatial grouping. Therefore, the difference in prediction performance can be attributed to sampling location, where high-predicting datapoints were sampled on-reef, and the low-predicting data were largely sampled off-reef (as proxied by their gold grades). In addition, this study found that the RF is the best MLA for gold

grade prediction in the Witwatersrand Basin, followed by the AdaBoost algorithm, with the EN being the least suitable. The adoption of ML in geosciences is in its infancy, as exemplified by the state of the data and its suitability for ML in general. However, it has enormous potential for process automation and analytical objectivity, making it a robust and progressive approach that exploits new technologies, thus moving the mining industry into the new high-tech world.

REFERENCES

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32.

Cate, A., Perozzi, L., Gloaguen, E., and Blouin, M. (2017). Machine learning as a tool for geologists. *The Leading Edge*. 36(3), 215-219.

Chapman, P., Kerber, R., Clinton, J., Khabaza, T., Reinartz, T., and Wirth, R. (1999). The CRISP-DM process model. *The CRISP–DM Consortium*. 91, 1-310.

Frimmel, H.E., Lehrmann, A.Z.B., Hallbauer, D., and Frank, W. (2009). Geochemical and geochronological constraints on the nature of the immediate basement next to the Mesoarchean auriferous Witwatersrand Basin, South Africa. *Journal of Petrology*. 50 (12), 2187-2220.

Frimmel, H.E. and Nwaila, G.T. (2020). Geologic evidence of syngenetic gold in the Witwatersrand goldfields, South Africa. *Society of Economic Geologists*. 23, 645-668.

Frimmel, H.E. (2023). Linking Archaean climate change with gold metallogeny. *Canadian Journal of Earth Sciences*. 60, 802-817.

Hall, P., Gill, N., and Schmidt, N. (2019). Proposed guidelines for the responsible use of explainable machine learning. *Proceedings of the 33rd Conference on Neural Information Processing Systems*. pp. 1-11.

Jolley, A.J., Freeman, A.R., Barnicoat, A.C., Phillips, G.M., Knipe, R.J., Fox, N.P.C., Strydom, D., Birch, M.T.G., and Henderson, I.H.C. (2004). Structural controls on Witwatersrand gold mineralisation. *Journal of Structural Geology*. 24, 1067-1086.

Jung, D. and Choi, Y. (2021). Systematic review of machine learning applications in mining: exploration, exploitation, and reclamation. *Minerals*. 11(2), 1-22.

Kapageridis, I.K. (2005). Input space configuration effects in neural network-based grade estimation. *Computational Geosciences*. 31, 704–717.

Kapageridis, I.K. and Denby, B. and Hunter, G. (1999). Integration of a neural ore grade estimation tool in a 3D resource modeling package. *In proceedings of the International Joint Conference on neural networks*. Washington DC, USA.

Kaplan, U.E. and Topal, E. (2020). A new ore grade estimation using combine machine learning algorithms. *Minerals*. 10(10), 1-17.

Koduri, S.B., Gunisetti, L., Ramesh, C.R., Mutyalu, K.V., and Ganesh, G. (2019) Prediction of crop production using Adaboost regression method. International Conference on Computer Vision and Machine Learning. 1228, 1-11.

Lindsay, J.J., Hughes, S.R.H., Yeomans, C.M., Andersen, J.C.Ø., and McDonald, I. (2021). A machine learning approach for regional geochemical data: Platinum-group element geochemistry vs geodynamic settings of the North Atlantic Igneous Province. *Geoscience Frontiers*. 12(3), 1-18.

Liu, W. and Li, Q. (2017). An efficient Elastic Net with regression coefficients method for variable selection of spectrum data. PLoS ONE. 12(2), 1-13.

Marwanza, I., Nas, C., Badaruddin, S.P., and Azizi, M.A. (2018). Copper resource estimation in PT X Batu Hijau, Regency of West Sumbawa, West Nusa Tenggara Province using geostatistical method. *Proceedings of the IOP Conference Series*. Earth and Environmental Science. pp. 0-10.

McCarthy, T.S. (2006). The Witwatersrand Supergroup, in Johnson, M.R., Anhaeusser, C.R., and Thomas, R.J., eds. *The geology of South Africa: Pretoria, Geological Society of South Africa, Johannesburg and Council for Geoscience*. 51, 155–186.

Nainggolan, R., Perangin-angin, R., Simarmata, E., and Tarigan, F.A. (2019). Improved the performance of the K-Means cluster using the sum of squared error (SSE) optimized by using the elbow method. Journal of Physics: Conference Series. 1361, 1-7.

Novitasari, H.B., Hadianto, N., Sfenrianto, M., Rahmawati, A., Prasetyo, R., Miharja, J., and Gata, W. (2019). K-nearest neighbor analysis to predict the accuracy of product delivery using administration of raw material model in the cosmetic industry (PT Cedefindo). Journal of Physics: Conference Series. 1367, 1-7.

Nwaila, G. & Zhang, S., Bourdeau, J., and Ghorbani, Y. (2022). Spatial conditioning of machine learning algorithms for geoscience applications. *Conference: SGA 2022 - The Critical Role Of Minerals In The Carbon-Neutral Future*. 1, 279-282.

Nwaila, G.T., Zhang, S.E., Tolmay, L.C.K., and Frimmel, H.E. (2020). Algorithmic optimization of an underground Witwatersrand-type gold mine plan. *International Association for Mathematical Geosciences*. 1-23.

Phillips, G.N. and Powell, R. (2011). Origin of Witwatersrand gold: a metamorphic devolatilization–hydrothermal replacement model. *Applied Earth Science*. 120(3), 112-129.

Rahimi, H., Asghari, O., and Meysami, F. (2018). Investigation of linear and non-linear estimation methods in highly-skewed gold distribution. *Journal of Mining and Environment*. 9(4), 967-979.

Safonov, Y.G. and Prokof'ev, V.Y. (2006). Gold-bearing reefs of the Witwatersrand Basin: A model of synsedimentation hydrothermal formation. *Geology of Ore Deposits*. 48(6). 415-447.

Taylor, R.D. and Anderson, E.D. (2018). Quartz-pebble-conglomerate gold deposits: Scientific Investigations Report 2010–5070–P. *U.S. Geological Survey*. 1-34.

Tucker, R.F., Richard, P.V., and Viljoen, M.J. (2016). A review of the Witwatersrand Basin - The world's greatest goldfield. *Episodes*. 39(2), 105-133.

Wickham, H, and Grolemund, G. (2017*). R for Data Science: import, tidy, transform, visualize, and model data*. O'Reilly Media.

Yang, Y., Li, Y., Deng, X., and Yan, T. (2021). Structural controls on the gold mineralization at the eastern margin of the North China Craton: Constraints from gravity and magnetic data from the Liaodong and Jiaodong Peninsulae. *Ore Geology Reviews*. 139, 104522.

Zhang, S.E., Nwaila, G.T., Tolmay, L., Frimmel, H.E., and Bourdeau, J.E. (2021). Integration of machine learning algorithms with gompertz curves and kriging to estimate resources in gold deposit. *National Resources Research*. 30(3), 1-18.

Tariro Mombe

PhD Student (Geology)
University of the Witwatersrand

Miss Mombe is a dedicated geoscientist specializing in resource estimation and geometallurgy. She has a strong academic foundation, including a Master of Science (MSc) in Geology from the University of the Witwatersrand. Her MSc research looked at in-situ gold grade prediction using machine learning algorithms in the Witwatersrand Basin. She is currently a PhD candidate at the Wits Mining Institute's Sibanye_Stillwater Digital Mining Laboratory (DigiMine), where she is responsible for the portable XRF. Her PhD research looks at an integrated machine learning and metal accounting approach to facilitate real-time tracking of critical raw materials. She is resolute in pursuing innovative research that seamlessly integrates machine learning and AI into geoscience.