Machine Learning: Geotechnical Engineering

Predicting UCS and RQD using Random Forest

A Presentation for Operations Technical Support - Ok Tedi Mining Limited

Presented by Jesse Gabriel

October 27, 2025

Outline

- What & Why Machine Learning (ML)?
- Machine Learning Models and Workflow
- Raw Data and EDA
- Preprocessing
- ML Model Training & Evaluation Approach
- Results: UCS (ESTUCS) RQD Predictions
- Discussion: Further Work
- Discussion: Related Work on Application

What & Why Machine Learning (ML)?

- Machine Learning (ML), a subfield of AI, involves development of algorithms that enable computers to learn and improve task performance from data and experience without explicit programming.
- Types: Supervised, Unsupervised, Reinforcement, Semi-supervised
- Data inputs: Spreadsheet, Image, Text, Software files (e.g. GIS shapefiles), etc.
- ML can handle large datasets and generate useful prediction outputs

INPUT DATA → ML MODEL → OUTPUT

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	HOLEID	PROJECTC	GEOLFROI	GEOLTO	PRIORITY	GT_RockStre	RQD	GT_Weath	ESTUCS	GT_RS_FAI	GT_RS_C_(	GT_RS_C_(	GT_RS_C_F	GT_RS_C_F	GT_RS_C_F
2	GDH1011	TUNNEL	0	5.7	11		0	MW	20		20	0.75	Poor Rock	10	3
3	GDH1011	TUNNEL	5.7	9.1	11		3.45	HW	10		20	0.75	Poor Rock	10	3
4	GDH1011	TUNNEL	9.1	17.8	11		13.79	MW	40		2	13.7931	Fair Rock	18	3
5	GDH1011	TUNNEL	17.8	23.3	11		25.14	MW	40		2	16.90909	Fair Rock	16	4
6	GDH1011	TUNNEL	23.3	26	11		0	HW	3		20	0.75	Very Poor	0	3
7	GDH1011	TUNNEL	26	34.35	11		28.5	HW	5		20	4.275449	Poor Rock	16	5
8	GDH1011	TUNNEL	34.35	36.4	11		17.78	MW	40		20	1.756098	Poor Rock	10	4
9	GDH1011	TUNNEL	36.4	40.8	11		8.97	HW	5		2	12.55981	Fair Rock	14	3
10	GDH1011	TUNNEL	40.8	44.15	11		2.3	MW	30		2	7.5	Fair Rock	18	3
11	GDH1011	TUNNEL	44.15	46.3	11		26.05	MW	20		20	3.906977	Poor Rock	10	5
12	GDH1011	TUNNEL	46.3	52.8	11		14.46	MW	30		3	14.46154	Fair Rock	18	4
13	GDH1011	TUNNEL	52.8	62.15	11		21.62	MW	40		6	7.130125	Fair Rock	12	5
14	GDH1011	TUNNEL	62.15	69.4	11		28.37	MW	30		6	13.7931	Fair Rock	16	5
15	GDH1011	TUNNEL	69.4	75.8	11		49.49	SW	120		12	8.004386	Fair Rock	16	8

	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE
1	GT_RS_C_F	GT_RS_C_F	GT_RS_C_V	GEOLINTE	GT_Alterat	GT_RS_FR/	GT_RS_FR/	GT_RS_AV(	GT_RS_C_(	GT_RS_C_(	GT_RS_RE(	GT_RS_RQ	GT_RS_LIT	GT_Logger	GT_LoggedI	TCR_pct
2	5	3	3	5.7		0	0		1	3	3.9	0	BF	JAS	03-Mar-11	68.42105
3	5	2	4	3.4		0	0		1	3	2.9	0.1	MD	JAS	03-Mar-11	85.29412
4	19	4.8	3	8.7		3	0.52		1	3	5.8	0.8	MD	JAS	03-Mar-11	66.66667
5	19	4.8	3	5.5		2	0.54		1	2	3.7	0.93	MD	JAS	03-Mar-11	67.27273
6	5	1	4	2.7		0	0		1	3	2.1	0	MD	JAS	03-Mar-11	77.77778
7	5	1	4	8.35		0	0		1	3	8.35	2.38	MD	JAS	03-Mar-11	100
8	5	4.8	3	2.05		0	0		1	3	1.35	0.24	MD	JAS	03-Mar-11	65.85366
9	16	1	4	4.4		3	0.77		0.95	3	3.9	0.35	MD	JAS	03-Mar-11	88.63636
10	20	3.5	3	3.35		1	0.33		1	3	3.05	0.07	MD	JAS	03-Mar-11	91.04478
11	5	3	3	2.15		0	0		1	3	2.15	0.56	MD	JAS	03-Mar-11	100
12	20	3.5	3	6.5		3	0.46		1	3	6.5	0.94	MD	JAS	03-Mar-11	100
13	13	4.8	3	9.35		11	1.19		1	2	9.25	2	MD	JAS	03-Mar-11	98.93048
14	12	3.5	3	7.25		10	1.42		1	3	7.05	2	MD	JAS	03-Mar-11	97.24138
15	10	11	2	6.4		14	2.37		0.95	2	5.9	2.92	MD	JAS	03-Mar-11	92.1875

Figure: Sample spreadsheet data (geotechnical log)

Machine Learning Models and Workflow

EDA → PREPROCESSING → MODEL TRAINING → RESULTS → POSTPROCESSING

$$\hat{y} = f_\theta(X) + \varepsilon \qquad \text{with} \quad X = \{x_1, \ldots, x_p\}$$

$\hat{y}$: prediction f_θ : model θ : parameters X : features x_i : feature p : dimensionality ε : noise

```python
from sklearn.ensemble import RandomForestRegressor
# Example: X is feature matrix, y target (UCS or RQD)
# X = data[feature_cols].values
# y = data['UCS'].values
model = RandomForestRegressor(n_estimators=200, random_state=42,....)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

Raw Data and EDA

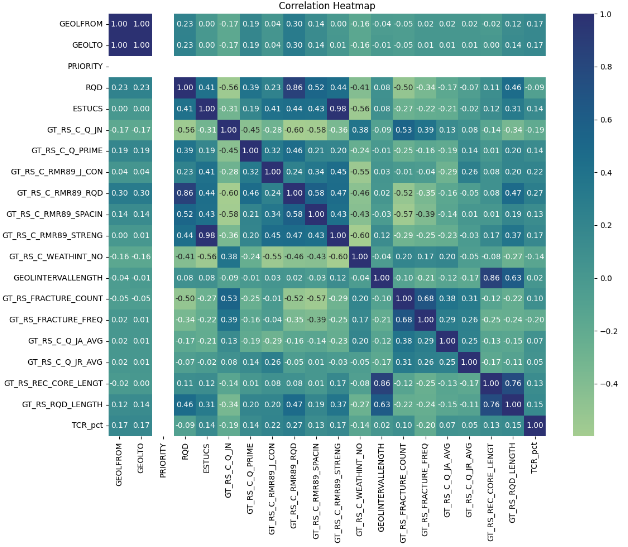
- Geotechnical Log data (13236 rows 31 columns)

Table: Data types and missing values summary

#	Variable	Data Type	Missing	Missing (%)
1	HOLEID	object	0	0.00
2	PROJECTCODE	object	0	0.00
3	GEOLFROM	float64	0	0.00
4	GEOLTO	float64	0	0.00
5	PRIORITY	int64	0	0.00
6	GT_RockStrength	object	1998	15.10
7	RQD	float64	194	1.47
8	GT_Weathering	object	14	0.11
9	ESTUCS	float64	0	0.00
10	GT_RS_FABRIC	object	5010	37.85
11	GT_RS_C_Q_JN	float64	65	0.49
12	GT_RS_C_Q_PRIME	float64	11	0.08
13	GT_RS_C_RMR89_DESC	object	167	1.26
14	GT_RS_C_RMR89_J_CON	float64	89	0.67
15	GT_RS_C_RMR89_RQD	float64	174	1.31
16	GT_RS_C_RMR89_SPACIN	float64	183	1.38
17	GT_RS_C_RMR89_STRENG	float64	71	0.54
18	GT_RS_C_WEATHINT_NO	float64	15	0.11
19	GEOLINTERVALLENGTH	float64	0	0.00
20	GT_Alteration	object	5011	37.86
21	GT_RS_FRACTURE_COUNT	float64	2	0.02
22	GT_RS_FRACTURE_FREQ	float64	130	0.98
23	GT_RS_AVG_DEFECT_RGH	object	5522	41.72
24	GT_RS_C_Q_JA_AVG	float64	609	4.60
25	GT_RS_C_Q_JR_AVG	float64	634	4.79
26	GT_RS_REC_CORE_LENGT	float64	52	0.39
27	GT_RS_RQD_LENGTH	float64	186	1.41
28	GT_RS_LITH_ROCKTYPE	object	6	0.05
29	GT_Logger	object	111	0.84
30	GT_LoggedDate	object	79	0.60
31	TCR_pct	float64	12 379	93.53

Raw Data and EDA

- Correlation matrix heatmap



Preprocessing

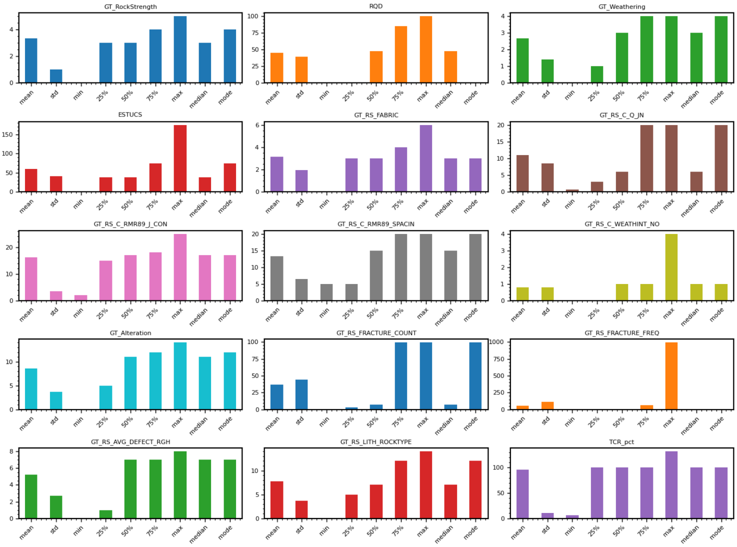
- Feature selection
- Handling outliers (e.g. RQD above 100%)
- Handling missing values
- Computing summary statistics
- Encoding of categorical data

Table: Sample encoded data: `GT_RS_AVG_DEFECT_RGH`

Original Value	Encoded Value	Count
UR	7	4876
PR	1	1879
PS	2	400
US	8	264
SI	3	58
UP	6	51
PP	0	40
SP	4	19
SS	5	10

Preprocessing

- Summary statistics of selected & encoded data (features & targets)



ML Model Training & Evaluation Approach

- **Random Forest Regressor**
- **Hyperparameter Optimization with Optuna**: We define an objective function that tests different hyperparameter combinations for a Random Forest using cross-validated R^2 as the evaluation metric.
- **Parameter Search Space**: Key parameters tuned include n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features, while bootstrap and random_state are fixed.
- **Automated Study Execution**: Optuna runs 50 trials to maximize the model's R^2 score, automatically identifying the best-performing hyperparameters.
- **Final Model Training & Evaluation**: The Random Forest is retrained on the full training data using the best parameters; predictions on the test set are evaluated with R^2 and RMSE metrics.

ML Model Training & Evaluation Approach

- R^2 measures the proportion of variance in the target explained by the model. $RMSE$ quantifies the average magnitude of prediction errors.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

y_i : actual value $\hat{y}_i$: predicted value $\bar{y}$: mean of actual values n : number of samples

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

y_i : actual value $\hat{y}_i$: predicted value n : number of samples

ML Model Training & Evaluation Approach

```python
def objective(trial):
    params = {
        'n_estimators': trial.suggest_int('n_estimators', 200, 800),
        'max_depth': trial.suggest_int('max_depth', 5, 30),
        'min_samples_split': trial.suggest_int('min_samples_split', 5, 20),
        'min_samples_leaf': trial.suggest_int('min_samples_leaf', 5, 15),
        'max_features': trial.suggest_categorical('max_features', ['sqrt', 'log2']),
        'bootstrap': True,
        'random_state': 42
    }
    rf = RandomForestRegressor(**params)
    score = cross_val_score(rf, X_train, y_train, cv=5, scoring='r2').mean()
    return score
study = optuna.create_study(direction="maximize")
study.optimize(objective, n_trials=50, show_progress_bar=True)
best_params = study.best_params
best_params_dict[target] = best_params
```

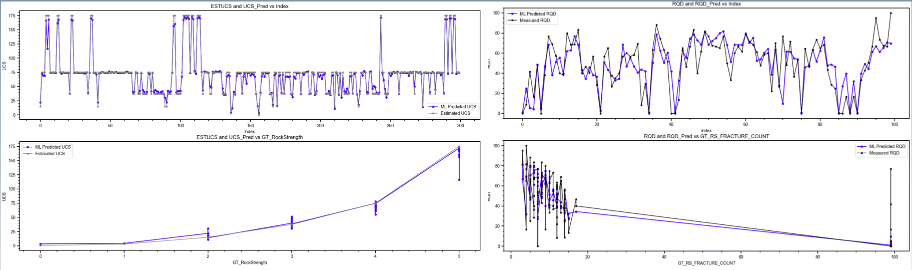
```
[I 2025-10-27 23:36:01,731] Trial 42 finished with value: 0.8305717604676554 and parameters: {'n_estimators': 383, 'max_depth': 12, 'min_samples_split': 10, 'min_samples_leaf': 5, 'max_features': 'sqrt'}. Best is trial 29 with value: 0.830966815135695.
[I 2025-10-27 23:36:04,539] Trial 43 finished with value: 0.8303061779671321 and parameters: {'n_estimators': 360, 'max_depth': 18, 'min_samples_split': 8, 'min_samples_leaf': 6, 'max_features': 'sqrt'}. Best is trial 29 with value: 0.830966815135695.
[I 2025-10-27 23:36:09,066] Trial 44 finished with value: 0.8299069675584141 and parameters: {'n_estimators': 607, 'max_depth': 15, 'min_samples_split': 12, 'min_samples_leaf': 7, 'max_features': 'sqrt'}. Best is trial 29 with value: 0.830966815135695.
[I 2025-10-27 23:36:12,592] Trial 45 finished with value: 0.8310083814162488 and parameters: {'n_estimators': 448, 'max_depth': 13, 'min_samples_split': 11, 'min_samples_leaf': 5, 'max_features': 'sqrt'}. Best is trial 45 with value: 0.8310083814162488.
[I 2025-10-27 23:36:15,966] Trial 46 finished with value: 0.8280852834182111 and parameters: {'n_estimators': 489, 'max_depth': 10, 'min_samples_split': 11, 'min_samples_leaf': 8, 'max_features': 'log2'}. Best is trial 45 with value: 0.8310083814162488.
[I 2025-10-27 23:36:19,501] Trial 47 finished with value: 0.8222201924038999 and parameters: {'n_estimators': 593, 'max_depth': 7, 'min_samples_split': 13, 'min_samples_leaf': 14, 'max_features': 'sqrt'}. Best is trial 45 with value: 0.8310083814162488.
[I 2025-10-27 23:36:21,796] Trial 48 finished with value: 0.8299643731450761 and parameters: {'n_estimators': 306, 'max_depth': 20, 'min_samples_split': 11, 'min_samples_leaf': 7, 'max_features': 'sqrt'}. Best is trial 45 with value: 0.8310083814162488.
[I 2025-10-27 23:36:25,942] Trial 49 finished with value: 0.8303878306658362 and parameters: {'n_estimators': 535, 'max_depth': 24, 'min_samples_split': 15, 'min_samples_leaf': 6, 'max_features': 'sqrt'}. Best is trial 45 with value: 0.8310083814162488.
Best trial: 45. Best value: 0.831008: 100%| | 50/50 [02:52<00:00,  3.45s/it]
RandomForestRegressor(max_depth=13, max_features='sqrt', min_samples_leaf=5,
                      min_samples_split=11, n_estimators=448, random_state=42)
📊 RQD - Test R2: 0.8380, RMSE: 15.7970
['D:\\coding\\vscode\\python\\geotech\\pilot\\models\\rf_RQD.pkl']

🔹 Training model for target: ESTUCS
Train size: 6077, Test size: 1520
[I 2025-10-27 23:36:26,878] A new study created in memory with name: no-name-436cfad8-5eaf-4ee8-81e0-16b952365121
[I 2025-10-27 23:36:30,940] Trial 0 finished with value: 0.36167063313618736 and parameters: {'n_estimators': 679, 'max_depth': 5, 'min_samples_split': 9, 'min_samples_leaf': 15, 'max_features': 'sqrt'}. Best is trial 0 with value: 0.36167063313618736.
[I 2025-10-27 23:36:35,458] Trial 1 finished with value: 0.39012727290636545 and parameters: {'n_estimators': 612, 'max_depth': 9, 'min_samples_split': 7, 'min_samples_leaf': 14, 'max_features': 'sqrt'}. Best is trial 1 with value: 0.39012727290636545.
[I 2025-10-27 23:36:40,330] Trial 2 finished with value: 0.3948139254383867 and parameters: {'n_estimators': 622, 'max_depth': 27, 'min_samples_split': 12, 'min_samples_leaf': 15, 'max_features': 'sqrt'}. Best is trial 2 with value: 0.3948139254383867.
[I 2025-10-27 23:36:42,174] Trial 3 finished with value: 0.39811339884615726 and parameters: {'n_estimators': 224, 'max_depth': 25, 'min_samples_split': 13, 'min_samples_leaf': 11, 'max_features': 'sqrt'}. Best is trial 3 with value: 0.39811339884615726.
[I 2025-10-27 23:36:44,322] Trial 4 finished with value: 0.39647445818925287 and parameters: {'n_estimators': 264, 'max_depth': 21, 'min_samples_split': 16, 'min_samples_leaf': 13, 'max_features': 'log2'}. Best is trial 3 with value: 0.39811339884615726.
[I 2025-10-27 23:36:47,004] Trial 5 finished with value: 0.39768061565131907 and parameters: {'n_estimators': 329, 'max_depth': 12, 'min_samples_split': 5, 'min_samples_leaf': 11, 'max_features': 'sqrt'}. Best is trial 3 with value: 0.39811339884615726.
```

Results: UCS (ESTUCS) & RQD Predictions

- Both models show solid performance on the test set.
- Predicted values of UCS and RQD exported as csv/excel files

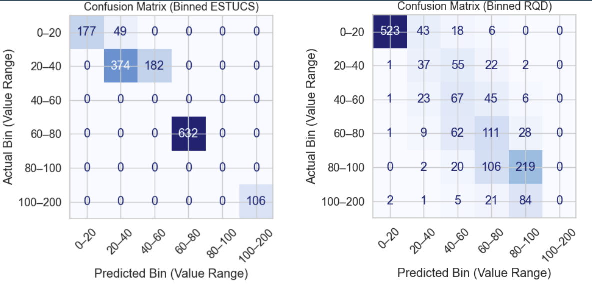
Table: Random Forest Model Summary

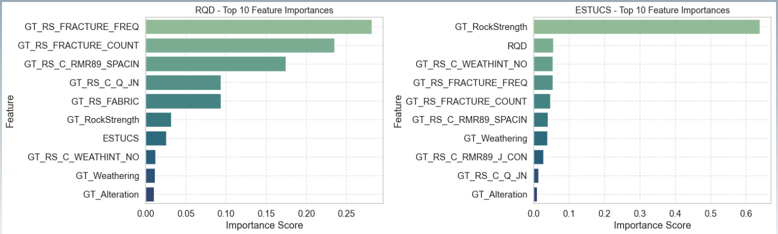
Target	Best Trial	Best Hyperparameters	Train Size	Test Size	R^2 (Test)	RMSE (Test)
RQD	29	n_estimators=458, max_depth=26, min_samples_split=7, min_samples_leaf=5, max_features='log2'	6077	1520	0.8594	14.7149
ESTUCS	41	n_estimators=535, max_depth=16, min_samples_split=8, min_samples_leaf=5, max_features='sqrt'	6077	1520	0.9818	5.3558



Results: UCS (ESTUCS) & RQD Predictions

- Confusion Matrix and Feature Importances





Discussion: Further Work

- Limitation: Proper selection of features required for this work
- Limitation: Need to handle imbalanced data to improve prediction
- Other derived/measured parameters apart from RQD & UCS can be predicted using appropriate models (FoS, displacement, etc.)
- Trained ML model can be deployed within existing/industry software or isolated web/desktop apps to complement standard workflows

Discussion: Related Work on Application

- Related work from Rocscience

geotechnics MDPI

Article

Back-Analysis of Structurally Controlled Failure in an Open-Pit Mine with Machine Learning Tools

Alison McQuillan [1], Amichai Mitelman [2,*] and Davide Elmo [3]

[1] Rocscience Inc., Toronto, ON M5T 1V1, Canada; alison.mcquillan@rocscience.com
[2] Department of Civil Engineering, Ariel University, Ariel 4077625, Israel
[3] Department of Mining Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada; delmo@mining.ubc.ca
* Correspondence: amichaim@ariel.ac.il

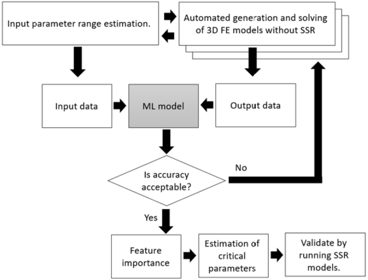


Figure 5. Flowchart of proposed back-analysis methodology.

Acknowledgments: THANK YOU

- Mr. Dauba Dauba (Snr. Geotechnical Engineer), for organizing and facilitating this presentation.
- OTML Operations Technical Support Team for your support and facilitation.
- All participants and attendees, for your engagement and valuable contributions.

- Further questions / discussions: Email: jessegabriel11@gmail.com