

**LAPORAN
UJIAN TENGAH SEMESTER (UTS)**

**MATAKULIAH
KECERDASAN BUATAN DI INDUSTRI MINYAK DAN GAS**



**Universitas
Pertamina**

Oleh:

Mukhammad Sholikhuddin 101321020

**FAKULTAS TEKNOLOGI EKSPLORASI DAN PRODUKSI
PROGRAM STUDI TEKNIK PERMINYAKAN
UNIVERSITAS PERTAMINA
2023**

DAFTAR ISI

COVER.....	1
DAFTAR ISI.....	2
DAFTAR GAMBAR.....	4
ABSTRAK.....	5
BAB I PENDAHULUAN	6
1.1 LATAR BELAKANG	6
1.2 RUMUSAN MASALAH	7
1.3 TUJUAN PENELITIAN	7
1.4 BATASAN MASALAH.....	7
BAB II STUDI LITERATUR	8
2.1 PREDIKSI ROP (<i>RATE OF PENETRATION</i>).....	8
2.2 REGRESSION DENGAN MENGGUNAKAN <i>MACHINE LEARNING</i>.....	9
BAB III METODOLOGI.....	10
3.1 ALUR PENGERJAAN MODEL <i>LINEAR REGRESSION</i>	11
3.2 ALUR PENGERJAAN MODEL <i>RANDOM FOREST</i>.....	12
BAB IV HASIL DAN PEMBAHASAN	13
4.1 DATA PREPOCESSING.....	13
4.1.1 Melakukan <i>Import Libraries</i> Dan <i>Datasets</i> Yang Digunakan	13
4.2 DATA ANALYSIS.....	13
4.2.1 Melakukan <i>Exploratory Data Analysis</i>	13
4.2.2 Melakukan <i>Multivariate Data Analysis</i>.....	15
4.2.3 <i>Feature Selection</i>	16
4.2.4 Menentukan Feature dan Output Matrix	18
4.2.5 Menentukan Training dan Test Sets.....	18
4.3 MODEL SELECTION AND EVALUATION.....	18
4.3.1 Model Selection	18
4.3.2 <i>Learning and Validation Curve</i>	20
4.3.4 Hyperparameter Tuning.....	24

4.3.5 Plotting	25
4.4 MODEL INSPECTION	26
4.5 PREDICTION USING MODEL	27
BAB V KESIMPULAN DAN SARAN.....	29
5.1 KESIMPULAN	29
5.2 SARAN	29
DAFTAR PUSTAKA	30
LAMPIRAN	31

DAFTAR GAMBAR

Gambar 1 Alur Penggerjaan Model Linear Regression	11
Gambar 2 Alur Penggerjaan Model Random Forest	12
Gambar 3 Exploratory Data Analysis Dataset UP-5	14
Gambar 4 Exploratory Data Analysis Dataset UP-6	14
Gambar 5 Heatmap Dataset UP-5	15
Gambar 6 Heatmap Dataset UP-6	15
Gambar 7 Heatmap setelah Feature Selection Dataset UP-5	17
Gambar 8 Heatmap setalah Feature Selection Dataset UP-5	17
Gambar 9 Feature Matrix dan Output Matrix	18
Gambar 10 Menentukan Training dan Test Sets	18
Gambar 11 Learning Curve Model Linear Regression Dataset UP-5	21
Gambar 12 Learning Curve Model Random Forest Dataset UP-5	21
Gambar 13 Learning Curve Model Linear Regression Dataset UP-6.....	22
Gambar 14 Learning Curve Model Random Forest Dataset UP-6	22
Gambar 15 Validation Curve Model Linear Regression Dataset UP-5.....	23
Gambar 16 Validation Curve Model Random Forest Dataset UP-5	23
Gambar 17 Validation Curve Model Linear Regression Dataset UP-6.....	24
Gambar 18 Validation Curve Model Random Forest Dataset UP-6	24
Gambar 19 Permutation Feature Importance	27
Gambar 20 Prediction Using Model Linear Regression	28
Gambar 21 Prediction Using Model Random Forest.....	28

ABSTRAK

Paper ini membahas penggunaan teknologi *machine learning* untuk membantu prediksi dan optimasi ROP (*Rate of Penetration*) di industri minyak dan gas, dengan membandingkan antara model *linear regression* dan *random forest regressor*. ROP merupakan salah satu parameter penting dalam proses pengeboran sumur minyak dan gas yang berpengaruh pada waktu, biaya, dan efisiensi produksi. Dalam paper ini, akan dibahas pengumpulan data historis dan faktor-faktor operasi yang mempengaruhi ROP. Selanjutnya, data tersebut akan digunakan untuk membangun model prediksi ROP menggunakan teknologi machine learning. Dalam hal ini, model *linear regression* dan *random forest regressor* akan dibandingkan untuk mengetahui model yang paling optimal dalam membantu prediksi ROP. Hasil penelitian menunjukkan bahwa penggunaan *random forest regressor* lebih baik dalam membantu prediksi ROP dibandingkan dengan *linear regression* model. Hal ini disebabkan karena *random forest regressor* mampu menangani data yang kompleks dan memiliki kemampuan untuk menangani interaksi antara variabel independen. Selain itu, paper ini juga membahas tentang manfaat penggunaan teknologi *machine learning* dalam industri minyak dan gas, seperti penghematan waktu dan biaya, peningkatan efisiensi produksi, dan peningkatan keselamatan kerja. Dalam kesimpulannya, penggunaan teknologi *machine learning*, khususnya *random forest regressor*, dapat membantu meningkatkan akurasi prediksi dan optimasi ROP di industri minyak dan gas.

Kata kunci: *Linear Regression, Machine Learning, Oil and Gas Industry, Random Forest Regressor, ROP.*

BAB I

PENDAHULUAN

1.1 Latar Belakang

Industri minyak dan gas merupakan salah satu industri yang sangat penting dalam memenuhi kebutuhan energi dunia. Salah satu kegiatan yang sangat penting dalam industri minyak dan gas adalah pengeboran sumur, dimana ROP (*Rate of Penetration*) menjadi salah satu parameter yang sangat penting dalam menentukan waktu, biaya, dan efisiensi produksi. ROP sendiri merupakan kecepatan pengeboran sumur yang diukur dalam satuan *feet per hour* (ft/h).

Peningkatan dan optimalisasi nilai ROP menjadi suatu permasalahan yang penting dalam industri minyak dan gas, mengingat nilai ROP yang optimal dapat mempercepat waktu pengeboran dan menurunkan biaya produksi. Namun, prediksi ROP yang akurat dan optimal masih menjadi tantangan karena dipengaruhi oleh banyak faktor, seperti litologi formasi, kecepatan rotasi, tekanan hidrostatik, dan parameter operasi lainnya.

Untuk memprediksi ROP, dapat digunakan prediksi korelasi berbasis fisika dan matematika. Namun, data kegiatan pengeboran yang diperoleh dari sumur memiliki jumlah data yang sangat banyak dalam satu dataset, sehingga memerlukan waktu yang lama dan tenaga yang besar untuk melakukan analisis secara manual.

Oleh karena itu, penggunaan teknologi *Machine Learning* dapat menjadi solusi dalam mengatasi masalah tersebut. Dalam hal ini, *Machine Learning* dapat digunakan untuk memprediksi suatu variabel permianyan seperti ROP dengan memanfaatkan data historis dan faktor-faktor operasi yang mempengaruhi ROP. *Machine Learning* dapat memproses data dalam waktu yang lebih cepat dan akurat, sehingga dapat membantu meningkatkan prediksi dan optimalisasi nilai ROP.

Dalam pemodelan menggunakan *Machine Learning*, salah satu teknik yang dapat digunakan adalah *linear regression* dan *random forest regressor*. Kedua teknik tersebut dapat digunakan untuk memprediksi nilai ROP, namun masing-masing memiliki kelebihan dan kelemahan. Oleh karena itu, perbandingan antara kedua model tersebut perlu dilakukan untuk mengetahui model yang lebih optimal dalam membantu prediksi dan optimalisasi nilai ROP.

1.2 Rumusan Masalah

1. Bagaimana bentuk pemodelan ROP (*Rate of Penetration*) pada lapangan UP- 5 dan UP-6 dengan menggunakan *machine learning*?
2. Bagaimana menentukan evaluasi dan investigasi dari hasil prediksi menggunakan model?
3. Bagaimana menentukan parameter utama yang paling mempengaruhi prediksi ROP (*Rate of Penetration*)?
4. Bagaimana dengan hasil yang didapatkan menggunakan model *linear regression* dan *random forest regressor* mana yang mempunyai hasil terbaik diantara kedua model tersebut?

1.3 Tujuan Penelitian

1. Menentukan pemodelan ROP (*Rate of Penetration*) pada lapangan UP-5 dan UP-6 dengan menggunakan *machine learning*.
2. Menentukan evaluasi dan investigasi dari hasil prediksi menggunakan model.
3. Menentukan parameter utama yang paling mempengaruhi prediksi ROP (*Rate of Penetration*).

1.4 Batasan Masalah

1. Menggunakan bahasa pemrograman Python yang mudah dalam pengaplikasian dan Pustaka libraries yang dapat diakses.
2. Terdapat dataset yang disediakan dalam bentuk file CSV (*Comma Separated Value*).
3. Dataset yang tersedia memiliki variable-variabel yang dibutuhkan berkaitan dengan BYM (*Burgoyne and Young Model*).

BAB II

STUDI LITERATUR

2.1 Prediksi ROP (*Rate of Penetration*)

Salah satu tujuan utama optimalisasi pengeboran adalah untuk mengurangi waktu total pengeboran dan tetap menjaga resiko serendah mungkin. Salah satu cara untuk mencapainya yakni dengan cara memilih variabel pengeboran yang optimal sebelum menjalankannya. Untuk merumuskan masalah optimalisasi pada kegiatan pengeboran diperlukan model prediksi yang akurat. Model tersebut memiliki tujuan untuk menilai beberapa variabel penting yang mempengaruhi kinerja pengeboran serta dengan variabel yang dapat diukur (Barbosa, 2019).

Prediksi *rate of penetration* (ROP) adalah salah satu kegiatan yang dilakukan dalam industri minyak dan gas untuk memperkirakan laju pengeboran sumur. Prediksi ROP ini sangat penting karena dapat membantu mempercepat waktu pengeboran dan menurunkan biaya produksi. Untuk itu akurasi model ROP menjadi sangat penting (Soares and Gray, 2019).

Memodelkan ROP sebagai fungsi matematis dari beberapa variabel tidaklah mudah, karena hal tersebut merupakan permasalahan yang sangat non-linear yang mana pemodelan ROP secara tradisional mempunyai keterbatasan (Soares et al, 2016). Untuk mempermudah melakukan prediksi ROP, dapat digunakan teknologi Machine Learning yang memanfaatkan data historis dan faktor-faktor operasi yang mempengaruhi ROP, seperti litologi formasi, kecepatan rotasi, tekanan hidrostatik, dan parameter operasi lainnya.

Oleh karena itu, pada era saat ini banyak peneliti yang telah mulai menggunakan teknik *machine learning* untuk memprediksi ROP (misalnya, ANN, SVM, random forest) yang mana hal tersebut dikarenakan kemampuannya yang terkenal sebagai penaksir fungsi universal (Hornik et al, 1989).

Mengoptimalkan *rate of penetration* (ROP) merupakan hal yang sangat penting dalam industri minyak dan gas. ROP yang optimal dapat membantu mengurangi biaya produksi, mempercepat waktu pengeboran sumur, serta meningkatkan efisiensi operasional.

2.2 Regression dengan Menggunakan *Machine Learning*

Regression dengan menggunakan *machine learning* sebagai prediksi ROP dapat menjadi solusi untuk meningkatkan efisiensi produksi dalam industri minyak dan gas. Dalam prediksi ROP, data yang digunakan meliputi data sejarah pengeboran sumur, parameter pengeboran, dan data geologi (Dupriest and Koederitz, 2005). Model *machine learning* akan menggunakan data tersebut untuk mempelajari pola dan korelasi antara variabel-variabel tersebut dan memprediksi nilai ROP yang optimal. Model *machine learning* dapat mempertimbangkan banyaknya faktor yang mempengaruhi ROP secara simultan, sehingga dapat memberikan hasil prediksi yang lebih akurat.

Beberapa model *machine learning* yang dapat digunakan untuk memprediksi ROP adalah *linear regression* dan *random forest*. Model-model ini memproses data historis dan faktor operasi untuk membuat prediksi ROP dengan tingkat akurasi yang tinggi. Pada umumnya, model-model *machine learning* ini dilatih dengan menggunakan data historis ROP dan faktor-faktor operasi yang dikumpulkan dari sumur-sumur sebelumnya. Setelah model terlatih, model tersebut dapat digunakan untuk memprediksi ROP pada sumur-sumur berikutnya dengan faktor operasi yang sama atau mirip (Barbosa, 2019).

Linear regression merupakan salah satu model *machine learning* yang sering digunakan dalam prediksi ROP. Model ini memperhitungkan korelasi antara variabel-variabel dengan asumsi bahwa hubungan antara variabel-variabel bersifat linear. Namun, model ini kurang dapat menangkap pola yang kompleks dalam data.

Random forest regression merupakan model *machine learning* yang lebih kompleks dan dapat menangkap pola yang lebih kompleks dalam data. Model ini memperhitungkan hubungan antara variabel-variabel secara non-linear dan dapat memperhitungkan interaksi antara variabel-variabel.

Dalam industri minyak dan gas, menggunakan model regression dengan *machine learning* sebagai prediksi ROP dapat membantu perusahaan untuk meningkatkan efisiensi produksi dan mengurangi biaya operasional. Namun, pemilihan model regression yang tepat tergantung pada sifat data dan tujuan prediksi yang ingin dicapai (Gandelman, 2012).

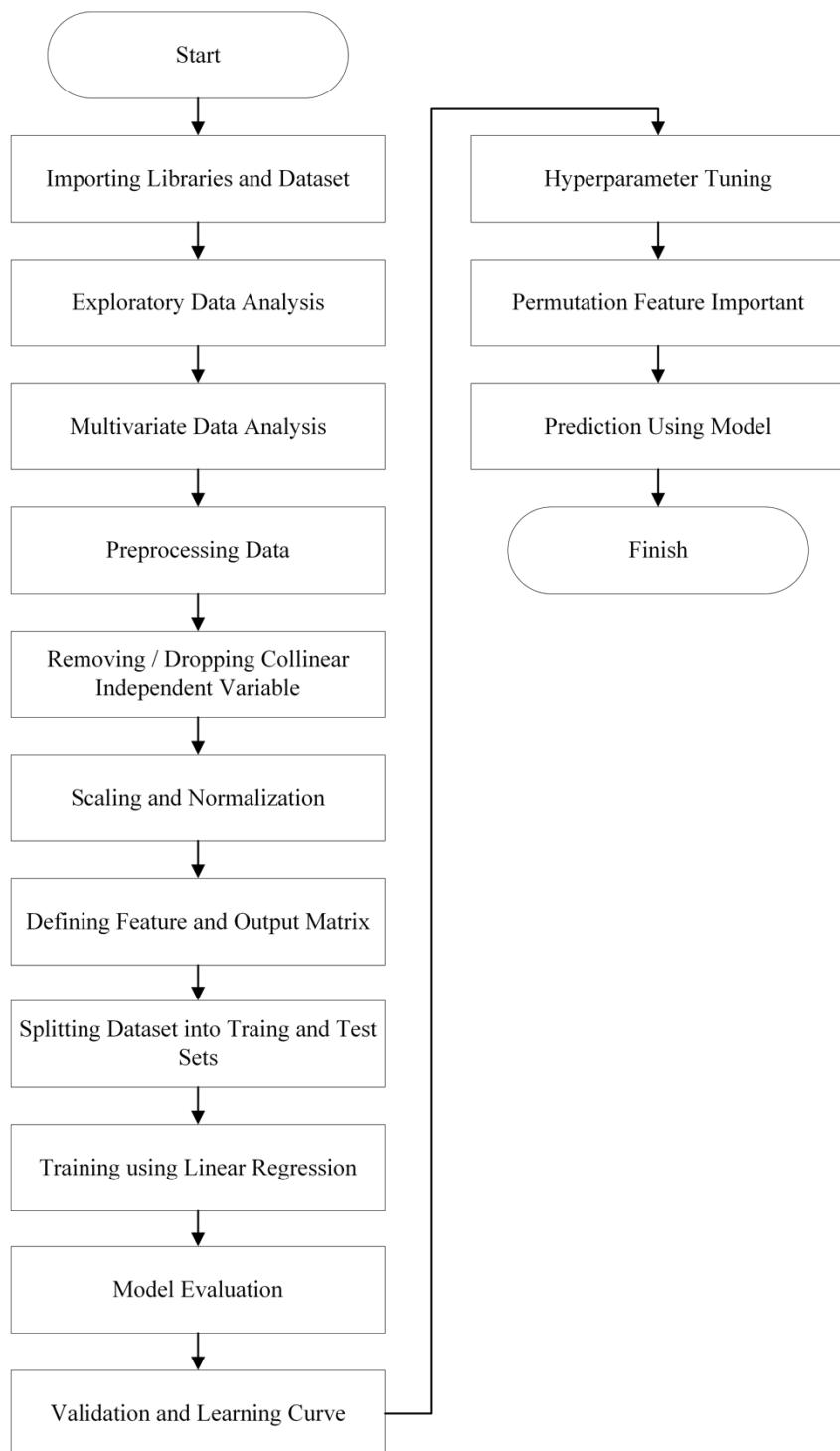
BAB III

METODOLOGI

Alur penggerjaan model *linear regression* dan *random forest* dimulai dengan mengimpor *library* dan *dataset* yang diperlukan. Setelah itu, dilakukan *exploratory data analysis* (EDA) untuk mengeksplorasi dataset dan memahami karakteristik variabel-variabel di dalamnya. Selanjutnya, dilakukan *multivariate data analysis* untuk mengevaluasi korelasi antar variabel dan mengidentifikasi outlier dan missing value pada dataset. Setelah itu, dilakukan *preprocessing data*, yaitu tahapan untuk mempersiapkan data sebelum dilakukan modelling. *Preprocessing* ini meliputi pembersihan data, pengisian missing value, dan konversi tipe data. Kemudian, dilakukan *dropping collinear independent variable*, yaitu menghapus variabel yang memiliki korelasi yang tinggi dengan variabel lain. Selanjutnya, dilakukan *scaling and normalization*, yaitu mengubah skala variabel-variabel agar memiliki rentang nilai yang serupa untuk mempermudah proses training. Kemudian, dilakukan *definisi feature dan output matrix*, yaitu menentukan variabel *input (feature)* dan variabel *output (target)* pada dataset. Setelah itu, dataset dibagi menjadi *training* dan *test sets* untuk melakukan training dan evaluasi model. Kemudian, dilakukan training menggunakan *linear regression* dan *random forest* untuk menghasilkan model.

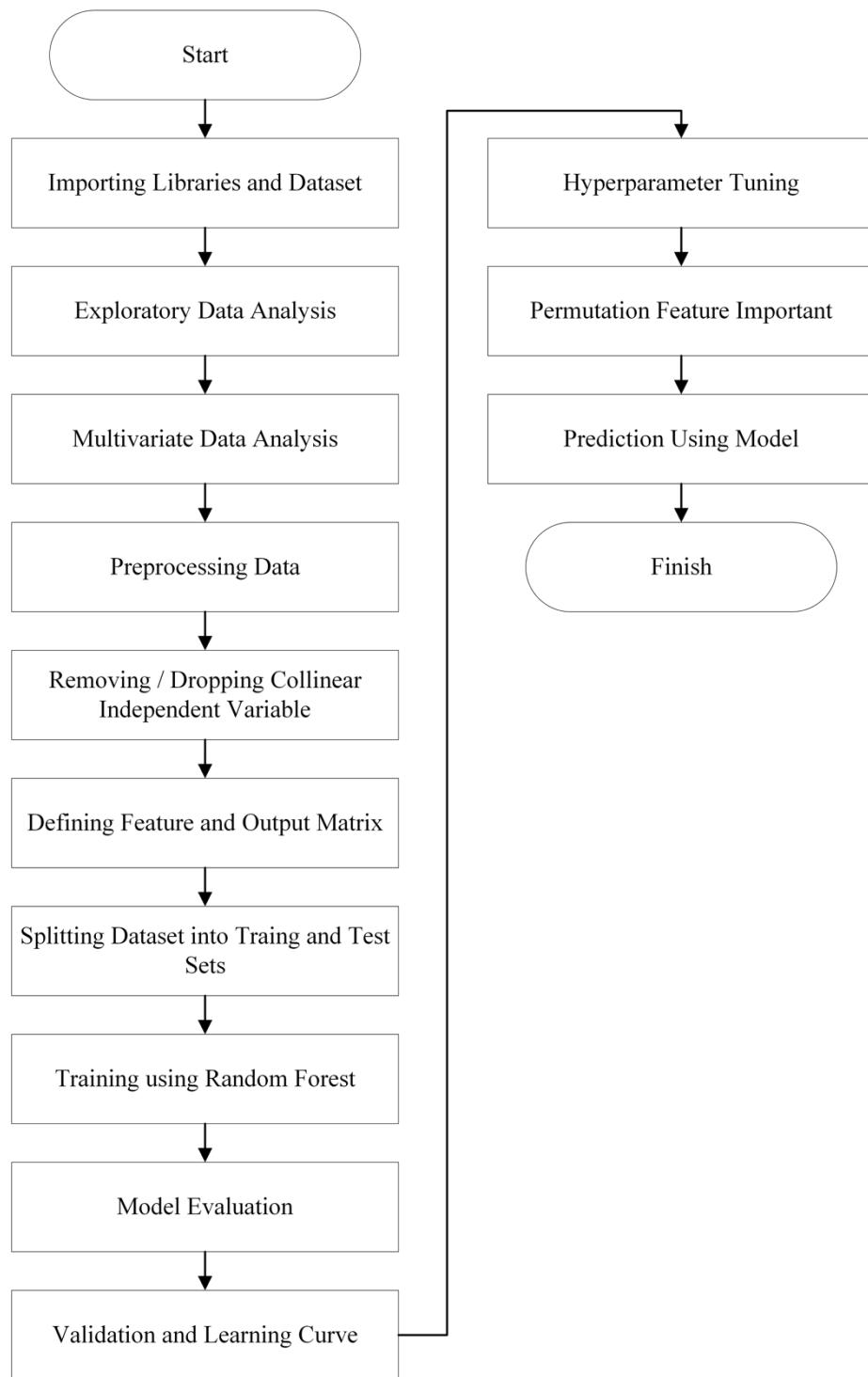
Setelah model selesai dilatih, dilakukan evaluasi model dengan mengukur performa model menggunakan *metrics* seperti MSE, RMSE, dan R-squared. Selain itu, dilakukan validasi model dengan membuat learning curve untuk menguji performa model pada data yang berbeda. Selanjutnya, dilakukan *hyperparameter tuning*, yaitu memilih nilai *hyperparameter* yang optimal untuk model. Kemudian, dilakukan *permutation feature importance* untuk mengevaluasi pengaruh masing-masing *feature* terhadap model. Setelah itu, model dapat digunakan untuk melakukan prediksi menggunakan data yang belum pernah dilihat sebelumnya. Terakhir, proses diakhiri dengan menghasilkan output yang diperlukan dan mengevaluasi performa model pada data baru.

3.1 Alur Pengerjaan Model *Linear Regression*



Gambar 1 Alur Pengerjaan Model *Linear Regression*

3.2 Alur Pengerjaan Model *Random Forest*



Gambar 2 Alur Pengerjaan Model *Random Forest*

BAB IV

HASIL DAN PEMBAHASAN

Dalam menjalankan algoritma model regression untuk menentukan *rate of penetration* (ROP) yang telah dikerjakan, dilakukan berbagai tahapan pengerjaan. Pengerjaan dilakukan menggunakan *platform* Microsoft Visual Studio Code. Adapun tahapan pengerjaannya ialah sebagai berikut:

4.1 Data Preprocessing

4.1.1 Melakukan *Import Libraries* Dan *Datasets* Yang Digunakan

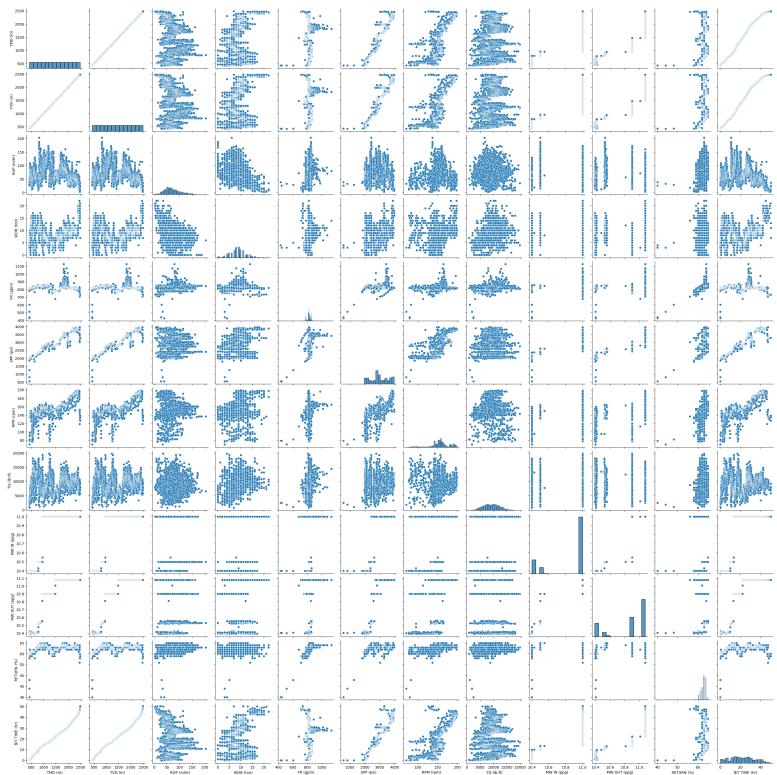
Pada pengerjaan kali ini, libraries yang digunakan yakni Pandas, Numpy, Seaborn, dan Matplotlib.pyplot. Selain itu juga menggunakan sklearn.model selection yang berupa *learning curve*, dan *validation curve*. Serta menggunakan *libraries* scikit-learn dalam membantu proses *preprocessing data*.

4.2 Data Analysis

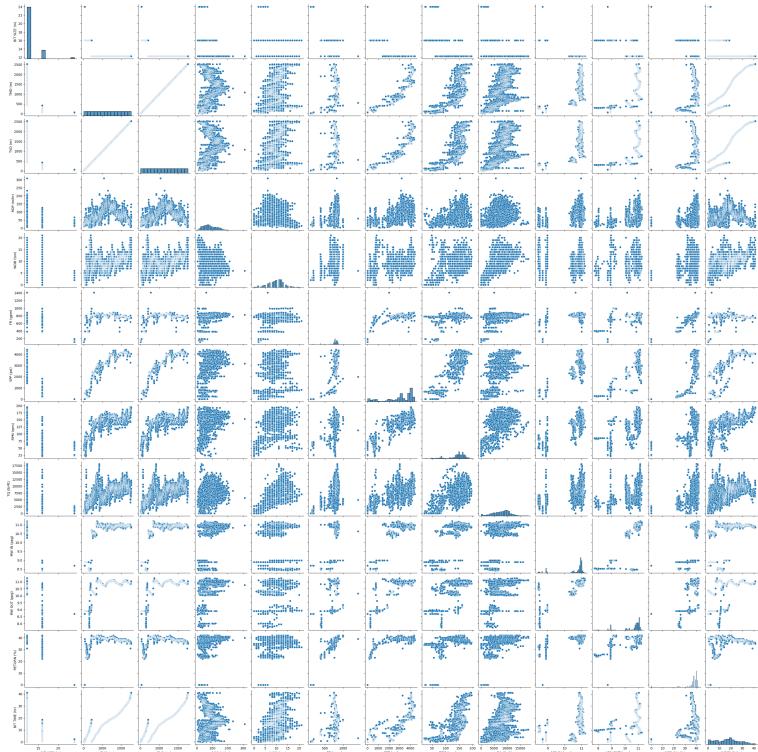
Data analysis yang dilakukan yakni berupa *exploratory* dan *multi variate data analysis*.

4.2.1 Melakukan *Exploratory Data Analysis*

Tujuan dari *exploratory data analysis* adalah untuk memahami data secara lebih mendalam dan menemukan pola atau tren dalam data yang dapat membantu dalam pengambilan keputusan. Dalam tahap ini, dilakukan eksplorasi visual dan statistik dari data, sehingga dapat ditemukan outlier, missing values, dan distribusi data. Dengan melakukan eksplorasi data yang baik, kita dapat menentukan tipe model yang tepat, memilih fitur yang sesuai, mengetahui ketergantungan antara fitur dan variabel target, serta menentukan strategi *preprocessing* yang tepat.



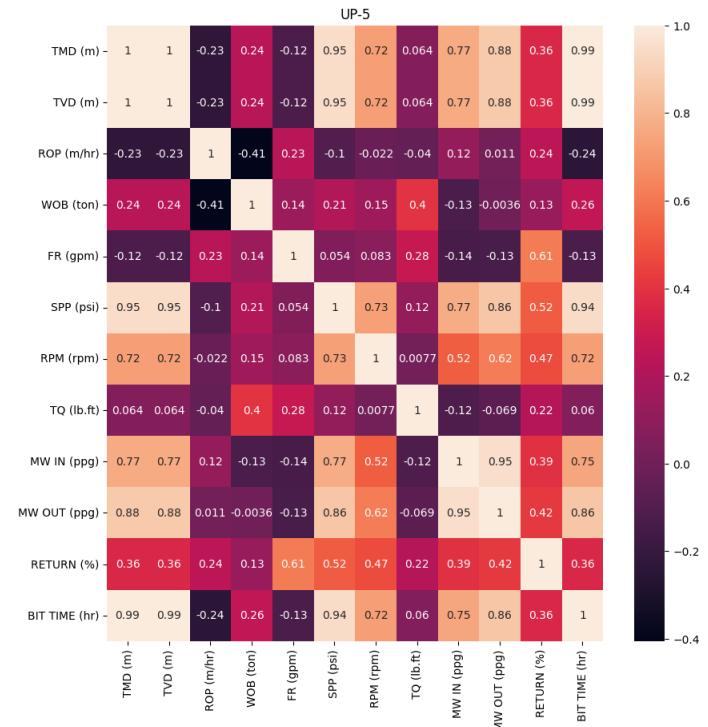
Gambar 3 Exploratory Data Analysis Dataset UP-5



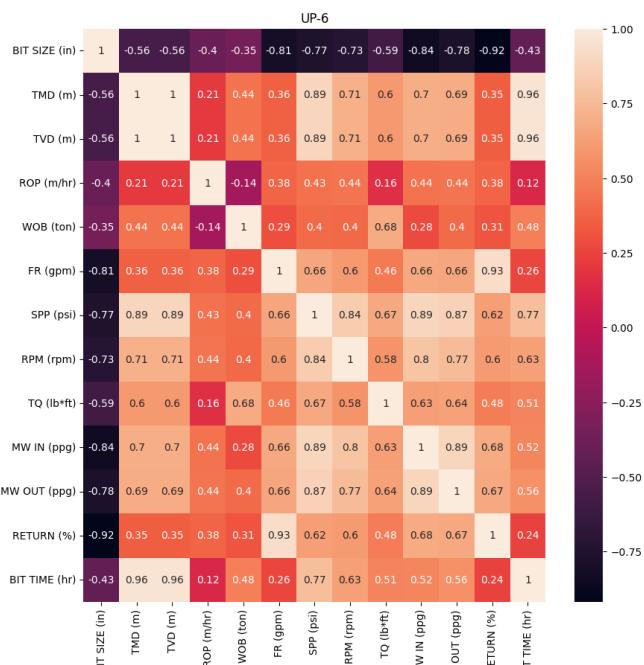
Gambar 4 Exploratory Data Analysis Dataset UP-6

4.2.2 Melakukan *Multivariate Data Analysis*

Pada *multivariate data analysis*, digunakan heatmap untuk melihat adanya interaksi antar *feature* yang menyebabkan *collinearity problems*.



Gambar 5 Heatmap Dataset UP-5



Gambar 6 Heatmap Dataset UP-6

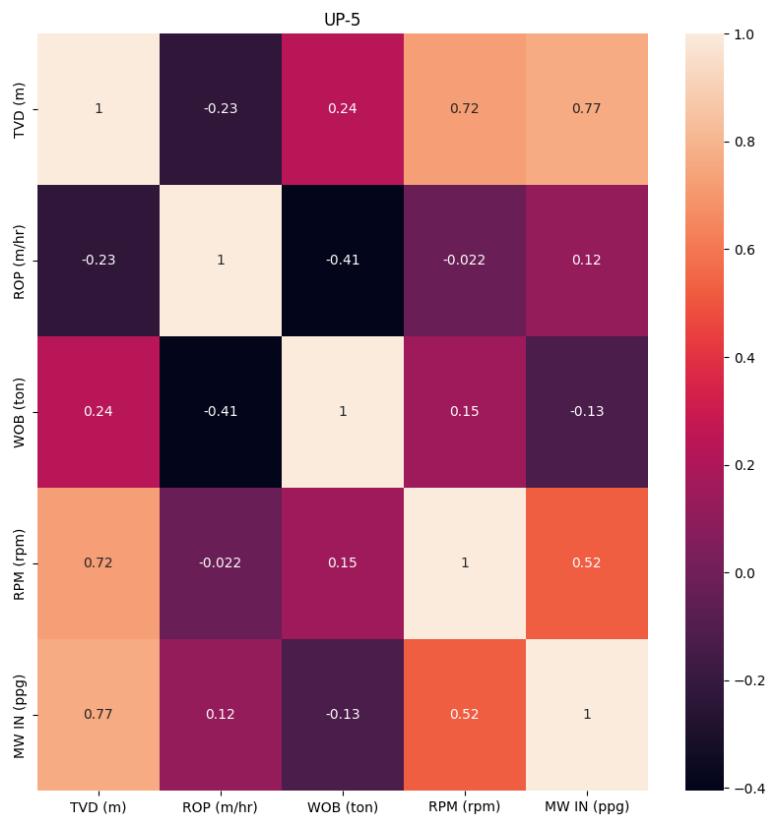
4.2.3 Feature Selection

Dari heatmap yang ditampilkan, terdapat banyak feature yang berinteraksi satu sama lain, maka dari itu dilakukan feature selection untuk memilih feature yang tidak digunakan. Pemilihan feature selection ini juga mempertimbangkan dari paper yang berjudul (*Machine learning methods applied to drilling rate of penetration prediction and optimization - A review*, Barbosa 2019), yang mana penulis menyebutkan frekuensi input yang digunakan dalam melakukan pemodelan ROP diantaranya yakni terdapat WOB, RPM, Depth, Flow Rate, Mud Weight, Bit Diameter, UCS, dan Bit Tooth Wear.

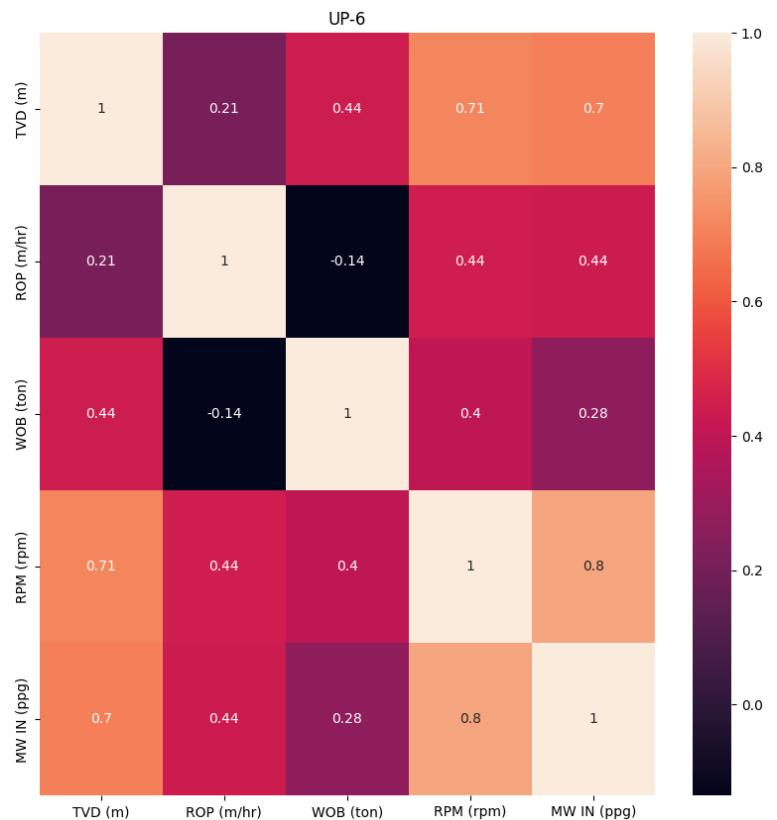
Berdasarkan pertimbangan dari adanya collinear independent variables, maka feature yang di drop adalah sebagai berikut:

- a. TMD (m)
- b. FR (gpm)
- c. SPP (psi)
- d. TQ (lb.ft)
- e. MW OUT (ppg)
- f. Return (%)
- g. Bit Time (hour)

Setelah dilakukan *feature selection*, maka tampak *heatmap* yang dihasilkan sadalah sebagai berikut:



Gambar 7 Heatmap setelah *Feature Selection* Dataset UP-5



Gambar 8 Heatmap setalah *Feature Selection* Dataset UP-5

4.2.4 Menentukan Feature dan Output Matrix

Feature matrix (X) merupakan *matrix* yang berisikan variabel-variabel independen yang digunakan untuk menentukan ROP. Sedangkan *output matrix* (y) berisikan data dari ROP.



```
● ● ●  
1 X = data_normal_up5.drop(['ROP (m/hr)'], axis=1)  
2 y = data_normal_up5['ROP (m/hr)']
```

Gambar 9 *Feature Matrix dan Output Matrix*

4.2.5 Menentukan Training dan Test Sets

Dengan menggunakan `train_test_split` dari `sklearn.model_selection`, dilakukan penentuan dari set *training* dan set *test* dimana set *test* ditentukan hanya merupakan 20% dari keseluruhan data. Dalam penentuan ini, `random_state` dimatikkan atau di set 1 agar tidak terjadi perubahan model.



```
● ● ●  
1 from sklearn.model_selection import train_test_split  
2 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
```

Gambar 10 Menentukan *Training* dan *Test Sets*

4.3 Model Selection and Evaluation

4.3.1 Model Selection

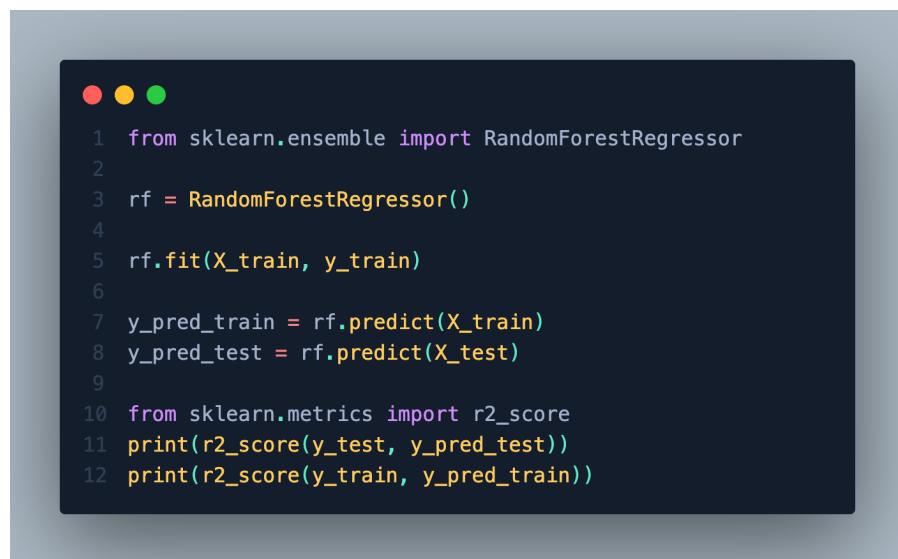
Pada algoritma ini, model yang digunakan ialah dua model yakni *linear regression* dan *random forest regressor*. Model ini dipilih berdasarkan kemudahan penggunaan dan akurasi yang tinggi. Selain itu tujuan menggunakan dua jenis yakni untuk dapat melihat hasil perbandingan diantara kedua model tersebut mana yang mempunyai nilai akurasi yang tinggi. Pada model *random forest regressor*, *feature scaling* tidak perlu dilakukan sedangkan pada *linear regression* diperlukan scaling dan normalization terlebih dahulu.

```
● ● ●  
1 from sklearn.linear_model import LinearRegression  
2  
3  
4 regressor = LinearRegression()  
5 regressor.fit(X_train_scaled, y_train)
```

```
● ● ●  
1 # Model Evaluation on Training Data  
2  
3 y_pred_train = regressor.predict(X_train_scaled)  
4  
5 from sklearn.metrics import r2_score  
6 r2_score(y_train, y_pred_train)  
7  
8 from sklearn.metrics import mean_squared_error  
9 mean_squared_error(y_train, y_pred_train, squared=False)  
10  
11 # Model Evaluation on Testing Data  
12  
13 y_pred_test = regressor.predict(X_test_scaled)  
14  
15 from sklearn.metrics import r2_score  
16 r2_score(y_test, y_pred_test)  
17  
18 from sklearn.metrics import mean_squared_error  
19 mean_squared_error(y_test, y_pred_test, squared=False)  
20
```

Dari hasil regresi menggunakan *Linear Regression* yang telah di fit pada data X_train_scaled dan y_train dilakukan prediksi terhadap nilai y. Kemudian dengan menggunakan r2, dilakukan perhitungan score antara nilai y_train terhadap y_pred_train dan y_test terhadap y_pred_test yang diprediksi dengan nilai aktualnya. Didapatkan nilai sebesar 0.74 untuk y_train dan 0.70 untuk y_test. Hal ini menunjukkan bahwa score dari nilai y_train maupun y_test belum cukup akurat untuk dataset pada UP-5. Sedangkan nilai yang didapatkan untuk y_train sebesar 0.56 dan y_test sebesar 0.54 untuk dataset UP-6 hasil yang didapatkan tidak seakurat yang didapatkan pada dataset UP-5.

Selain menggunakan model *linear regression* pada Dataset percobaan UP-5 dan UP-6 juga dilakukan menggunakan *random forest regressor*.



```
● ● ●
1 from sklearn.ensemble import RandomForestRegressor
2
3 rf = RandomForestRegressor()
4
5 rf.fit(X_train, y_train)
6
7 y_pred_train = rf.predict(X_train)
8 y_pred_test = rf.predict(X_test)
9
10 from sklearn.metrics import r2_score
11 print(r2_score(y_test, y_pred_test))
12 print(r2_score(y_train, y_pred_train))
```

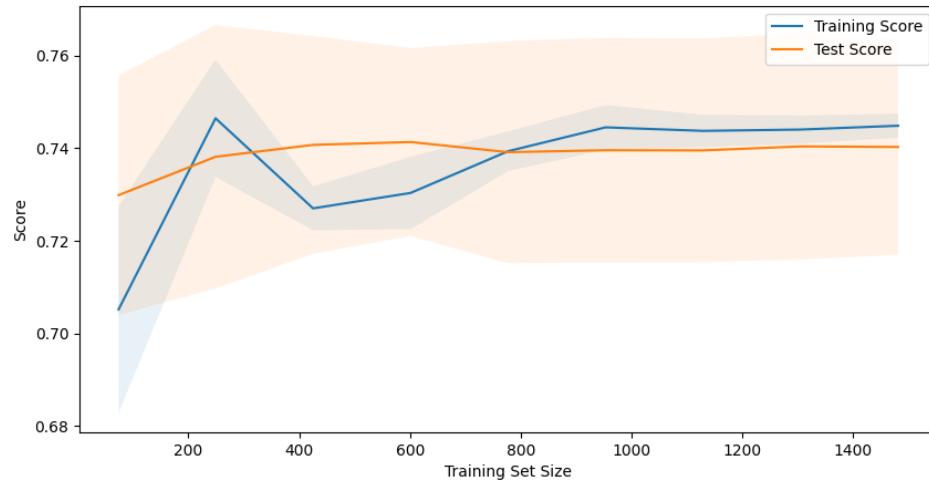
Dari model *random forest regressor* didapatkan nilai r2 score 0.97 untuk y_train dan 0.82 untuk y_test. Dari score yang didapatkan y_train sudah sangat akurat, sedangkan masih dibutuhkan optimasi score kembali bagi y_test untuk dataset UP-5. Pada sataset UP-6 nilai y_train yang didapatkan sebesar 0.99 dan sebesar 0.97 untuk y_test dari masing-masing score yang didapatkan sudah sangat akurat mendekati nilai 1 yang berarti dapat dikatakan sempurna.

4.3.2 Learning and Validation Curve

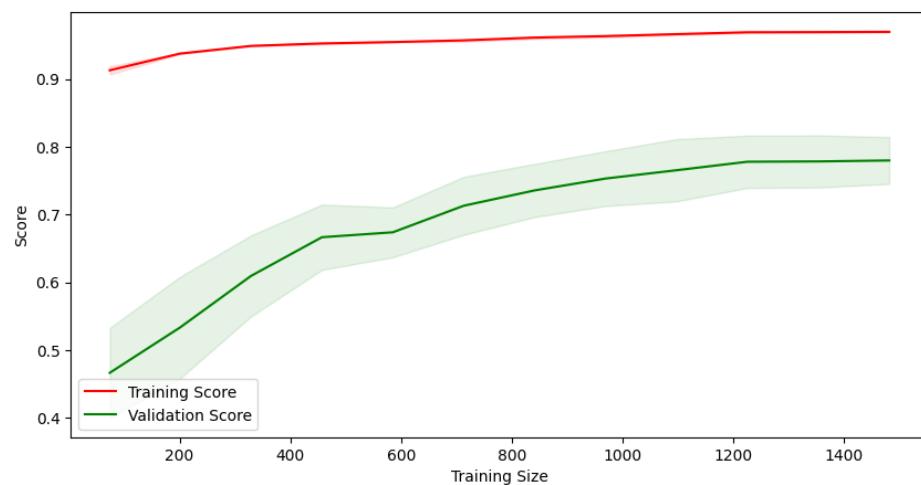
Dalam melakukan evaluasi terhadap model, digunakan *learning curve* dan *validation curve*. Kedua kurva tersebut dibentuk untuk melihat akan adanya bias dan variance. Bias sendiri merupakan rata-rata error yang dihasilkan dari training set, sedangkan variance merupakan indikasi keberagaman data. Tujuan dari *learning curve* dan *validation curve* adalah untuk mengevaluasi kinerja model *machine learning* dan menemukan titik optimal dalam jumlah data dan kompleksitas model.

Learning curve adalah grafik yang menunjukkan bagaimana kinerja model meningkat seiring dengan peningkatan jumlah data pelatihan yang digunakan. Dengan memplot *learning curve*, kita dapat melihat apakah

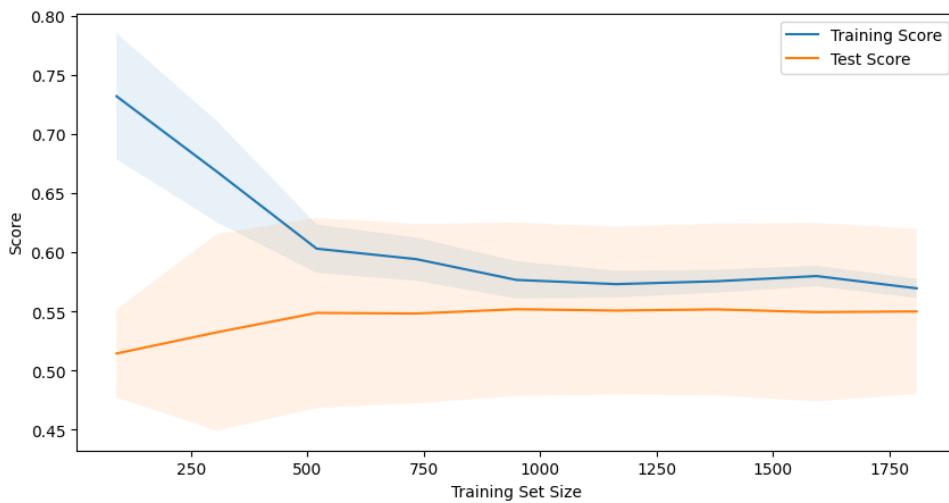
model mengalami *overfitting* atau *underfitting*. Jika *learning curve* untuk data pelatihan dan data validasi bertemu di titik tertentu dan tidak ada perbedaan yang signifikan, maka itu menunjukkan bahwa model sudah cukup baik dan penambahan data pelatihan tidak akan memberikan manfaat yang signifikan.



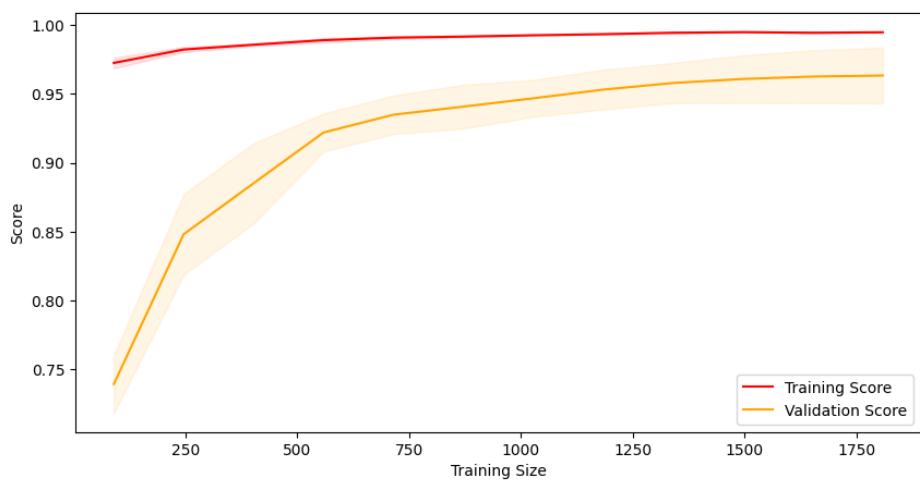
Gambar 11 Learning Curve Model Linear Regression Dataset UP-5



Gambar 12 Learning Curve Model Random Forest Dataset UP-5



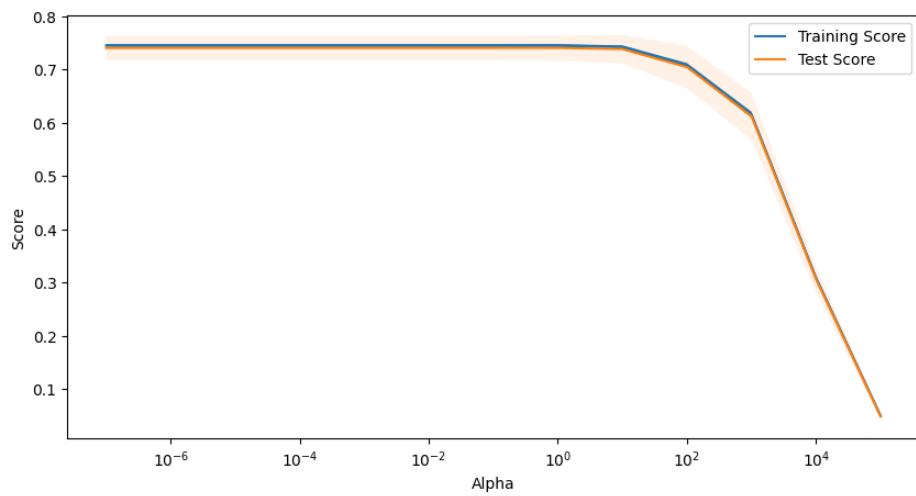
Gambar 13 Learning Curve Model Linear Regression Dataset UP-6



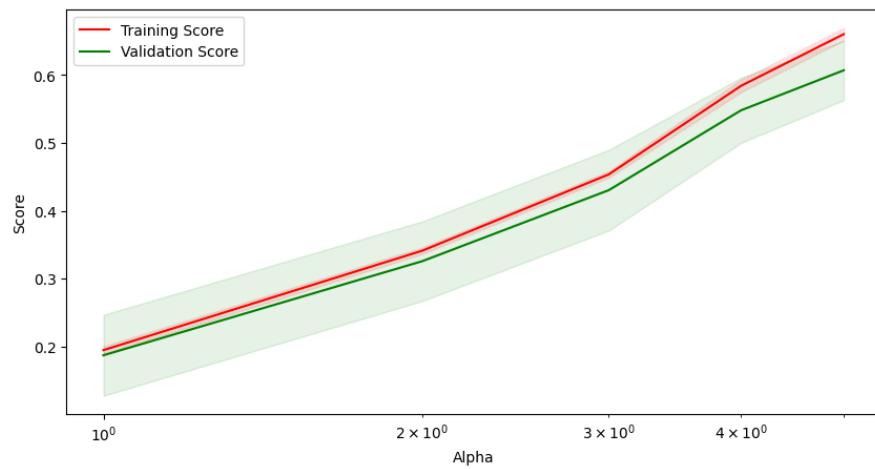
Gambar 14 Learning Curve Model Random Forest Dataset UP-6

Pada saat menggunakan model *random forest* dari kurva yang dihasilkan, pada dataset UP-5 maupun UP-6 tampak *validation score* semakin meningkat, sedangkan *training score* tidak berubah atau konstan. Hal ini menunjukkan bahwa peningkatan jumlah dataset pada training set masih bisa dilakukan untuk meningkatkan nilai score r². Sedangkan pada saat menggunakan model *Linear Regression* didapatkan pada dataset UP-6 tampak *train score* semakin menurun, sedangkan *test score* tidak berubah kembali atau konstan. Hal ini menunjukkan bahwa peningkatan jumlah dataset pada *training set* dapat berpengaruh terhadap semakin menurunnya nilai score r² begitu juga pada dataset UP-5.

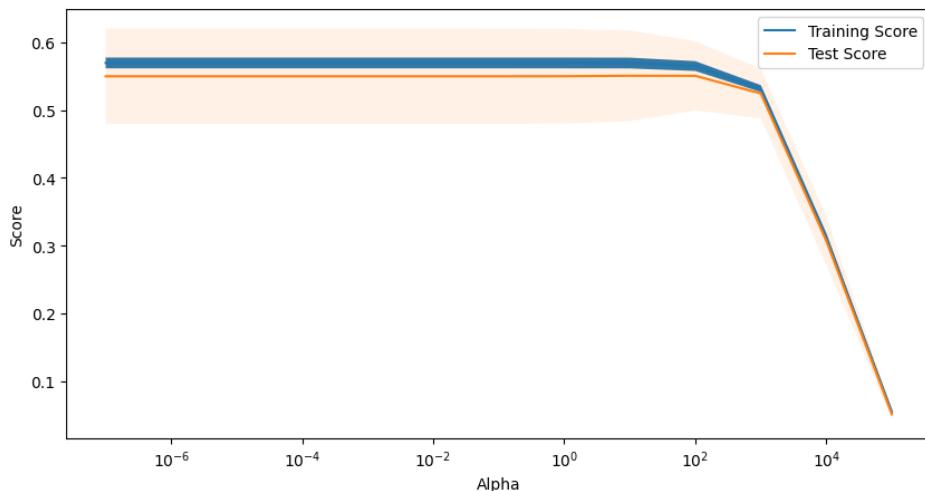
Validation curve adalah grafik yang menunjukkan bagaimana kinerja model berubah seiring dengan peningkatan kompleksitas model. Dengan memplot validation curve, kita dapat mengetahui titik optimal dalam jumlah parameter model yang dapat digunakan. Jika *validation score* cenderung stabil dan tidak ada perbedaan yang signifikan pada *validation score* antara model dengan jumlah parameter yang lebih sedikit dan lebih banyak, maka itu menunjukkan bahwa penambahan parameter model tidak akan memberikan manfaat yang signifikan.



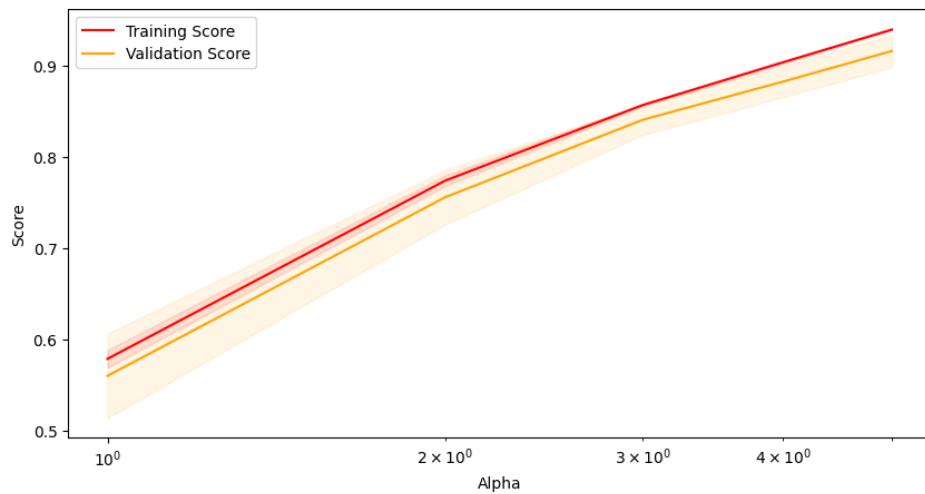
Gambar 15 Validation Curve Model Linear Regression Dataset UP-5



Gambar 16 Validation Curve Model Random Forest Dataset UP-5



Gambar 17 Validation Curve Model Linear Regression Dataset UP-6



Gambar 18 Validation Curve Model Random Forest Dataset UP-6

4.3.4 Hyperparameter Tuning

Hyperparameter tuning merupakan proses mengoptimalkan performa model *machine learning* dengan mengatur parameter-parameter yang diatur sebelumnya. Salah satu teknik untuk melakukan *hyperparameter tuning* adalah Grid Search CV dan Random Search CV. Grid Search CV adalah teknik untuk melakukan pencarian *hyperparameter* secara sistematis dengan menguji kombinasi nilai *hyperparameter* yang berbeda dan menemukan nilai yang menghasilkan performa terbaik pada model *machine learning*.

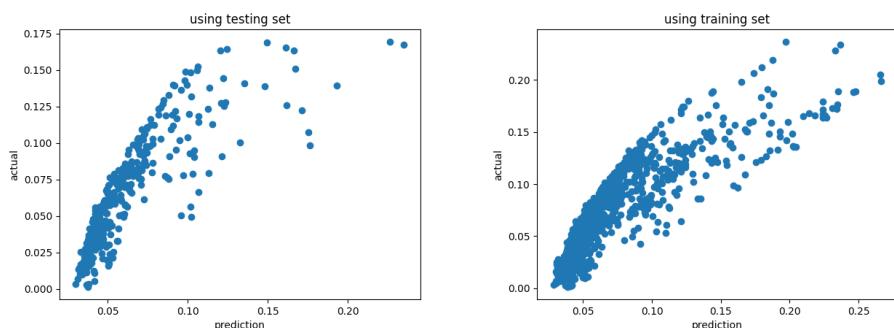
Namun, Random Search CV sering dipilih untuk model Random Forest karena model ini memiliki banyak *hyperparameter* yang perlu

dioptimalkan dan Random Search CV dapat mengeksplorasi ruang hyperparameter secara lebih efisien daripada Grid Search CV. Hal ini disebabkan karena Random Search CV secara acak memilih kombinasi hyperparameter untuk diuji, sehingga mempercepat proses pencarian hyperparameter yang optimal. Namun, pada dasarnya, teknik hyperparameter tuning yang digunakan akan tergantung pada model yang digunakan dan ukuran dataset yang diproses.

4.3.5 Plotting

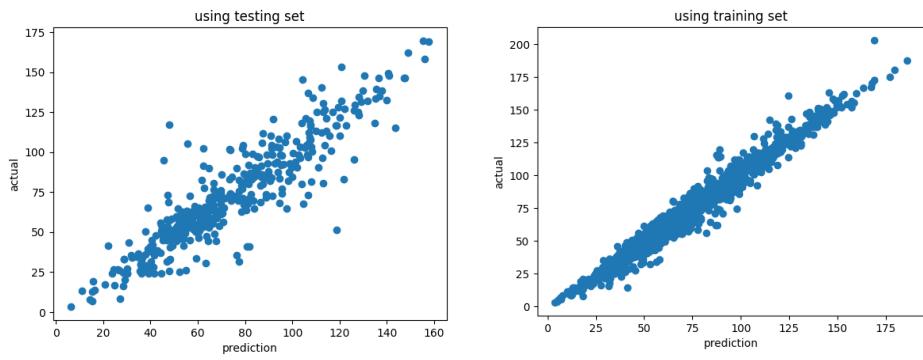
Tujuan dari *plotting* nilai aktual dan y prediksi adalah untuk melihat seberapa akurat model *machine learning* dalam memprediksi nilai target. Dengan membandingkan nilai aktual dan nilai prediksi, kita dapat melihat seberapa besar kesalahan prediksi model. *Plotting* tersebut juga dapat membantu kita untuk mengidentifikasi apakah terdapat bias atau pola yang berulang dalam kesalahan prediksi model.

Dengan menggunakan libraries `matplotlib.pyplot`, untuk membuat visualisasi grafik yang dapat membantu dalam melihat korelasi antara nilai y aktual dan y prediksi secara visual. Grafik yang dihasilkan dapat membantu dalam melihat pola dan sebaran data, serta melihat seberapa baik model dalam memprediksi nilai target.



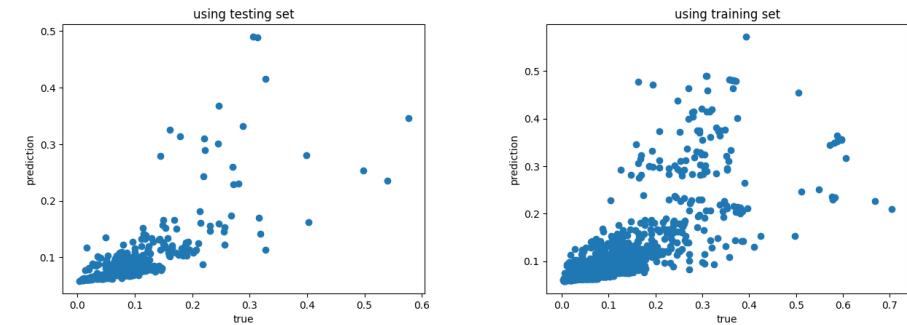
Plott on Testing Set Using Linear Regression Dataset 5

Plott on Training Set Using Linear Regression Dataset 5



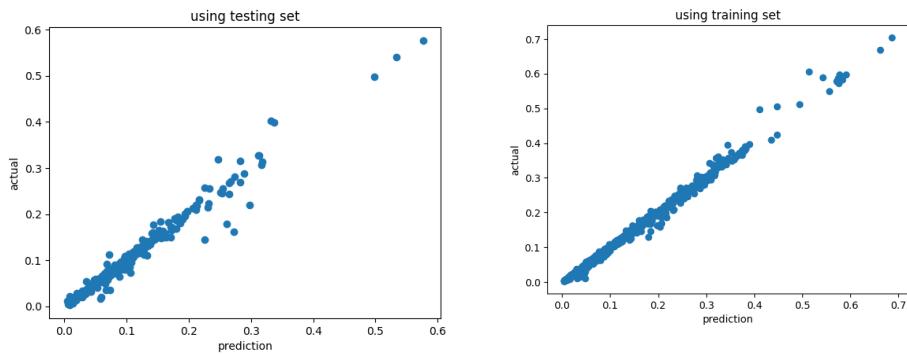
Plott on Testing Set Using Random Forest Dataset 5

Plott on Training Set Using Random Forest Dataset 5



Plott on Testing Set Using Linear Regression Dataset 6

Plott on Training Set Using Linear Regression Dataset 6



Plott on Testing Set Using Random Forest Dataset 6

Plott on Training Set Using Random Forest Dataset 6

4.4 Model Inspection

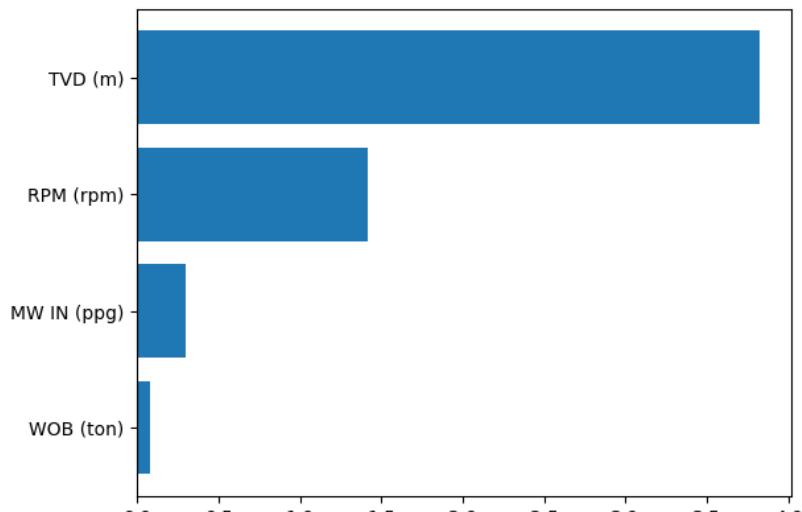
Model inspection berguna untuk melihat feature yang memiliki pengaruh paling besar dalam penentuan prediksi ROP. Dengan melakukan inspeksi model, kita dapat memeriksa hubungan antara setiap feature input dengan output target, sehingga kita dapat mengetahui mana yang feature yang paling signifikan atau paling berpengaruh terhadap hasil prediksi ROP.

Salah satu teknik yang dapat digunakan dalam *model inspection* adalah permutation feature importance. *Permutation feature importance* adalah teknik yang digunakan untuk mengevaluasi seberapa penting sebuah feature dalam model

dengan cara menghitung perubahan akurasi model ketika nilai feature tersebut diacak.

Dengan teknik *permutation feature importance*, kita dapat mengurutkan feature-input berdasarkan besarnya kontribusinya dalam menghasilkan prediksi yang akurat. Hal ini sangat berguna dalam memperbaiki atau meningkatkan model, karena kita dapat mengetahui feature-input mana yang perlu ditingkatkan atau dihilangkan dari model.

Sehingga, dengan melakukan model inspection kita dapat memperoleh informasi yang sangat penting dalam meningkatkan akurasi model dan juga memahami karakteristik data input dan output yang kita gunakan dalam prediksi ROP.

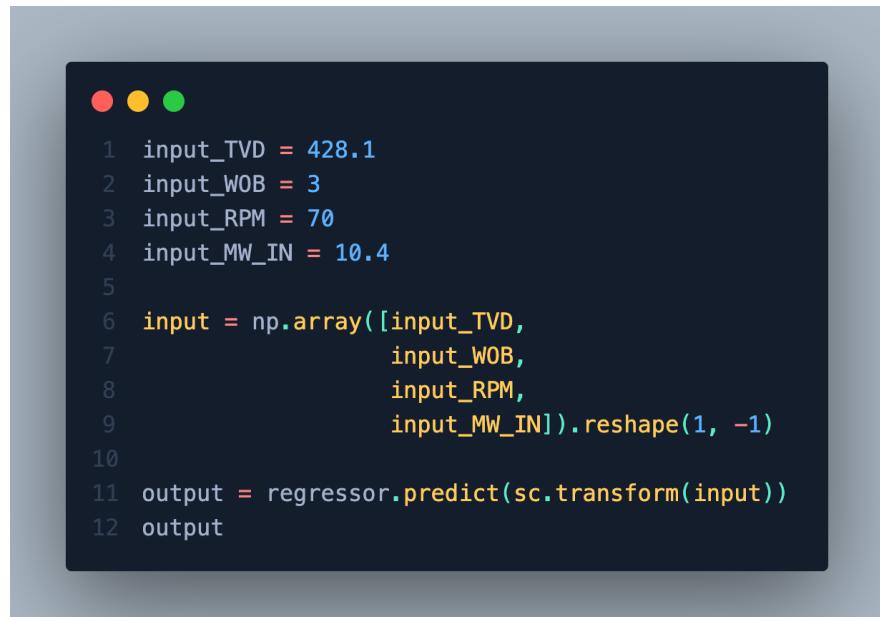


Gambar 19 Permutation Feature Importance

Dari hasil visual yang dihasilkan, didapatkan hasil bahwa feature yang memiliki pengaruh paling besar terhadap nilai ROP adalah TVD (m).

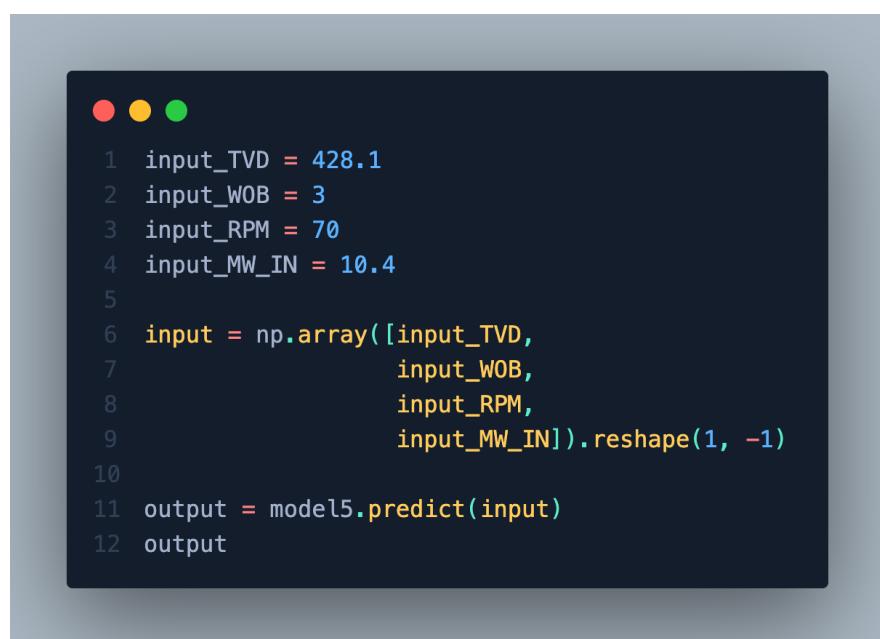
4.5 Prediction Using Model

Setelah didapatkan variable “model” untuk melakukan prediksi, maka prediksi ROP untuk seterusnya sudah dapat dilakukan. Untuk melakukan pengujian, digunakan data yang sesuai dari dataset original, sebagai berikut:



```
● ● ●
1 input_TVD = 428.1
2 input_WOB = 3
3 input_RPM = 70
4 input_MW_IN = 10.4
5
6 input = np.array([input_TVD,
7                 input_WOB,
8                 input_RPM,
9                 input_MW_IN]).reshape(1, -1)
10
11 output = regressor.predict(sc.transform(input))
12 output
```

Gambar 20 Prediction Using Model Linear Regression



```
● ● ●
1 input_TVD = 428.1
2 input_WOB = 3
3 input_RPM = 70
4 input_MW_IN = 10.4
5
6 input = np.array([input_TVD,
7                 input_WOB,
8                 input_RPM,
9                 input_MW_IN]).reshape(1, -1)
10
11 output = model5.predict(input)
12 output
```

Gambar 21 Prediction Using Model Random Forest

Setelah dilakukan running, maka didapatkan output ROP sebesar 34.079 (m/hr) untuk model *linear regression* sedangkan untuk model *random forest* sebesar 39.169 (m/hr).

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Model dari *rate of penetration* (ROP) yang telah dikembangkan menggunakan machine learning pada algoritma ini merupakan model *linear regression* dan *random forest regressor*.
2. Evaluasi dan investigasi dari hasil prediksi telah dilakukan dengan model dan didapatkan nilai dari score koefisien determinasi r^2 yang meningkat.
3. Pada model inspection, didapatkan parameter utama yang paling mempengaruhi dari hasil prediksi *rate of penetration* (ROP) yakni TVD (m).
4. Didapatkan dari hasil percobaan menggunakan model *linear regression* UP-5 mempunyai nilai akurasi yang lebih tinggi dibandingkan dengan dataset UP-6 sedangkan menggunakan model *random forest* didapatkan diantara kedua dataset UP-5 dan UP-6 hasil dengan akurasi terbaik didapatkan pada dataset UP-6.

5.2 Saran

Berdasarkan model dan algoritma yang telah dikembangkan dan dikonstruksi, terdapat nilai dari koefisien determinasi yang tinggi namun masih kurang cukup untuk dapat mengindikasikan bahwa seluruh *variabel independent* bersamaan memengaruhi variabel dependen secara sempurna. Maka dari itu, saran untuk pembuatan algoritma selanjutnya ialah dengan melakukan *data cleaning* untuk menghapus *outlier*, sehingga *fitting* terhadap data bisa dilakukan dengan lebih akurat.

DAFTAR PUSTAKA

- Barbosa, L. F. F. M., Nascimento, A., Mathias, M. H., & de Carvalho, J. A. (2019). Machine learning methods applied to drilling rate of penetration prediction and optimization - A review. *Journal of Petroleum Science and Engineering*, 183(March), 106332. <https://doi.org/10.1016/j.petrol.2019.106332>.
- Dupriest, F.E., Koederitz, W.L., 2005. Maximizing drill rates with real-time surveillance of mechanical specific energy. In: SPE/IADC Drilling Conference, Amsterdam, The Netherlands, 23-25 February. Society of Petroleum Engineers, Amsterdam, The Netherlands. <https://doi.org/10.2118/92194-MS>.
- Gandelman, R.A., 2012. Predição da ROP e Otimização em Tempo Real de Parâmetros Operacionais na Perfuração de Poços de Petróleo Offshore (Portuguese). Dissertação (metrado em tecnologia de processos químicos e bioquímicos). Escola de Química, Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ - Brasil. <http://epqb.eq. ufrj.br/download/predicao-da-rop-e-otimizacao-em-tempo-real.pdf>.
- Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8).
- Soares, C., Daigle, H., Gray, K., 2016. Evaluation of PDC bit ROP models and the effect of rock strength on model coefficients. *J. Nat. Gas Sci. Eng.* 34, 1225–1236. <https://doi.org/10.1016/j.jngse.2016.08.012>.
- Soares, C., Gray, K., 2019. Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models. *J. Pet. Sci. Eng.* 172, 934–959. <https://doi.org/10.1016/j.petrol.2018.08.083>.

LAMPIRAN

Link Google Collab:

https://colab.research.google.com/drive/1PyX5D_Y0YMpsgn7PuqZzmBtu-15Ntqw1?usp=sharing