# FINAL PROJECT

# ARTIFICIAL INTELEGENT

# “Predict Student Scores Based on the Number of Hours Studied”

Dosen Pengampu :

Yuyun Umaidah, S.Kom., M.Kom.

Disusun Oleh :


| | |
|---|---|
| Muhammad Rizqi Fadhilah | 2210631170089 |
| Juliana Widianti Dwi Putri | 2210631170025 |


## KELAS D

## INFORMATIKA

## FAKULTAS ILMU KOMPUTER

## UNIVERSITAS SINGAPERBANGSA KARAWANG

## 2024

# DAFTAR ISI

# BAB I

# PENDAHULUAN

## A. LATAR BELAKANG

Pendidikan merupakan salah satu aspek penting dalam pengembangan individu dan masyarakat. keberhasilan dalam pendidikan seringkali diukur melalui pencapaian akademik siswa, yang dinilai melalui berbagai bentuk evaluasi, termasuk nilai ujian. Salah satu faktor yang diyakini memiliki pengaruh signifikan terhadap pencapaian akademik adalah jumlah jam belajar yang dihabiskan oleh siswa. Pemahaman yang lebih baik tentang hubungan antara jumlah jam belajar dan nilai siswa dapat membantu guru, orang tua, dan siswa sendiri dalam merencanakan waktu belajar yang lebih efektif dan efisien.

Di era digital ini, teknologi kecerdasan buatan (Artificial Intelligence) dan pembelajaran mesin (Machine Learning) telah menjadi alat yang sangat berguna dalam menganalisis data dan membuat prediksi berdasarkan pola yang ditemukan dalam data tersebut. Dalam konteks pendidikan, teknologi ini dapat digunakan untuk memprediksi pencapaian akademik siswa berdasarkan berbagai faktor, termasuk jumlah jam belajar.

Regresi linear dan decision tree adalah dua metode yang umum digunakan dalam pembelajaran mesin untuk tugas prediksi. Regresi linear, sebagai metode statistik yang sederhana namun kuat, digunakan untuk memodelkan hubungan linier antara variabel input dan output. Di sisi lain, decision tree adalah algoritma yang lebih fleksibel yang dapat menangani hubungan non-linier antara variabel dan menghasilkan model yang mudah diinterpretasikan.

Proyek ini bertujuan untuk memprediksi nilai siswa berdasarkan jumlah jam belajar menggunakan kedua metode ini, yaitu regresi linear dan decision tree. Dengan membandingkan kinerja kedua model ini, kita dapat menentukan

metode mana yang lebih akurat dan tepat dalam konteks ini. Hasil dari proyek ini diharapkan dapat memberikan wawasan yang berguna bagi siswa, pendidik, dan orang tua dalam mengoptimalkan waktu belajar untuk mencapai hasil akademik yang lebih baik.

## B. TUJUAN

1. Mengembangkan model prediksi menggunakan regresi linear dan decision tree untuk memprediksi nilai siswa berdasarkan jumlah jam belajar.
2. Menganalisis hubungan antara jumlah jam belajar dan nilai siswa untuk memahami seberapa besar pengaruh jumlah jam belajar terhadap pencapaian akademik

## C. MANFAAT

1. Memberikan pemahaman yang lebih baik tentang hubungan antara jumlah jam belajar dan pencapaian akademik siswa, sehingga dapat digunakan untuk merencanakan strategi belajar yang lebih efektif
2. Meningkatkan pemahaman dan penerapan teknologi AI dan machine learning dalam bidang pendidikan, sehingga dapat mendorong inovasi lebih lanjut dalam pengajaran dan pembelajaran.

# BAB II

# DASAR TEORI

## 2.1 ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) adalah cabang ilmu komputer yang bertujuan untuk menciptakan sistem atau mesin yang dapat melakukan tugas-tugas yang biasanya memerlukan kecerdasan manusia. Ini mencakup kemampuan seperti memahami bahasa alami, mengenali gambar, menyelesaikan masalah, dan membuat keputusan. Menurut John McCarthy, yang merupakan salah satu pionir dalam bidang ini, AI adalah "ilmu dan teknik membuat mesin cerdas, terutama program komputer cerdas" .

Didalam Project kami ini AI digunakan untuk menganalisis data belajar siswa dan memprediksi nilai mereka berdasarkan jumlah jam belajar. AI memungkinkan kita untuk mengidentifikasi pola dalam data yang mungkin tidak terlihat oleh manusia dan membuat prediksi yang akurat berdasarkan pola tersebut.

## 2.2 REGRESI LINEAR

Regresi linear adalah teknik analisis data yang memprediksi nilai data yang tidak diketahui dengan menggunakan nilai data lain yang terkait dan diketahui. Secara matematis memodelkan variabel yang tidak diketahui atau tergantung dan variabel yang dikenal atau independen sebagai persamaan linier. Selain itu Regresi Linear merupakan metode statistik yang digunakan untuk memodelkan hubungan antara variabel dependen (respons) dan satu atau lebih variabel independen (prediktor). Model regresi linear sederhana dapat dinyatakan dengan persamaan:

$$y = \beta_0 + \beta_1 x$$

Dimana $y$ adalah variabel dependen, $x$ adalah variabel independen, $\beta_1$ adalah intercept, dan $\beta_0$ adalah koefisien regresi yang menunjukkan pengaruh variabel independen terhadap variabel dependen .

Pada Project ini kami menggunakan Regresi Linear Untuk memprediksi nilai siswa berdasarkan jumlah jam belajar. Dengan memodelkan hubungan linier antara jam belajar dan nilai siswa, kita dapat membuat prediksi yang membantu dalam merencanakan strategi belajar yang efektif.

## 2.3 DECISION TREE

Decision Tree Merupakan Sebuah Algoritma Flowchart yang berbentuk menyerupai struktur pohon yang digunakan untuk membantu membuat keputusan atau menyelesaikan tugas yang berkaitan dengan regresi dan klasifikasi. Struktur decision tree dimulai dari simpul akar (root node), cabang, simpul internal (internal node/decision node), dan terakhir simpul daun (leaf node/terminal node). Simpul akar (root node) mewakili pertanyaan atau masalah yang ingin dipecahkan. Kemudian cabang merupakan jalur keputusan, yang nantinya akan mengarah ke beberapa keputusan atau internal node. Setiap decision tree bisa memiliki beberapa internal node sebagai alternatif jawaban atau keputusan. Internal node juga bisa memiliki cabang node lain yaitu leaf node, yang akan mewakili keputusan akhir.

Dalam Project kami ini, Decision Tree digunakan untuk memprediksi nilai siswa berdasarkan jumlah jam belajar. Algoritma ini membantu dalam menangkap hubungan yang mungkin non-linear antara jam belajar dan nilai siswa, memberikan model yang mudah diinterpretasikan dan digunakan untuk membuat prediksi.

## 2.4 GOOGLE COLAB

Google Colab (Colaboratory) adalah layanan gratis dari Google yang memungkinkan pengguna menulis dan mengeksekusi kode Python langsung di browser mereka. Google Colab menyediakan akses ke sumber daya komputasi yang kuat, termasuk GPU dan TPU, serta integrasi dengan Google Drive, membuatnya menjadi alat yang sangat berguna untuk kolaborasi dalam proyek machine learning dan data science .

Pada Project ini Google Colab digunakan sebagai platform untuk menulis, mengeksekusi, dan berbagi kode Python. Kemampuan untuk menggunakan GPU dan TPU memungkinkan pelatihan model yang lebih cepat, dan integrasi dengan Google Drive memudahkan pengelolaan dan berbagi dataset serta notebook.

## 2.5 LIBRARY DAN PYTHON

Python adalah bahasa pemrograman tingkat tinggi yang dikenal karena sintaksisnya yang sederhana dan kemudahan penggunaannya. Python memiliki ekosistem library yang luas yang mendukung berbagai tugas dalam machine learning dan data science. Beberapa library yang relevan untuk proyek ini termasuk:

- NumPy: Untuk operasi numerik dan array.
- Pandas: Untuk manipulasi dan analisis data.
- Scikit-learn: Untuk algoritma machine learning seperti regresi linear dan decision tree.
- Matplotlib dan Seaborn: Untuk visualisasi data .

Library-library ini digunakan untuk mengolah data, membangun model prediksi, dan mengevaluasi kinerja model. Python, dengan dukungan komunitas dan ekosistem library yang kaya, merupakan bahasa pemrograman yang ideal untuk proyek machine learning dan data science.

# BAB III

# HASIL PEMBAHASAN DAN KODE

Pada bagian kali ini akan ditampilkan hasil dari run program pada Google Collabs beserta penjelasan mengenai setiap gambar dan data yang ditampilkan, lalu menampilkan langkah dari percobaan yang dilakukan, didapakanlah hasil diantaranya sebagai berikut :

1. Import dataset
2. Exploratory Data Analysis : statistical summary untuk mengetahui rata-rata, median, mean dkk, lalu lakukan scatter plot.
3. Melakukan analisis lainnya semisal correlation heatmap (opsional)
4. feature engineering (check duplikat data dan drop duplikat, check missing value, check outlier
5. splitting data ke X_train,X_test,y_train,y_test
6. Melakukan regresi modelling memakai linear regression, decision tree
7. Menggunakan model linear regression lalu keluarkan nilai intercept dan coef.
8. Plotting the actual and predicted values pada tiap model
9. Membandingkan 2 model dengan cara menampilkan 2 nilai r pada setiap model lalu mencari tau model dengan nilai r tertinggi.

**Gambar 1. Menunjukan Informasi data**

```
# Menampilkan beberapa informasi terkait kolom dan data dari student_scores
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25 entries, 0 to 24
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Hours   25 non-null     float64
 1   Scores  25 non-null     int64
dtypes: float64(1), int64(1)
memory usage: 528.0 bytes
```

Perlu diketahui untuk mencari pemodelan yang baik maka kita perlu mengetahui terlebih dahulu seperti apa data yang kita punya, baik dari kolom , tipe data , apakah ada nilai kosong, berapa banyak data yang ada. Untuk menampilkan hal tersebut kita bisa menggunakan syntax **data.info()** dimana fungsi ini akan menampilkan beberapa informasi terkait data sesuai gambar diatas. Data yang digunakan adalah **data student_scores.csv** yang dapat diakses pada link bagian akhir pembahasan.

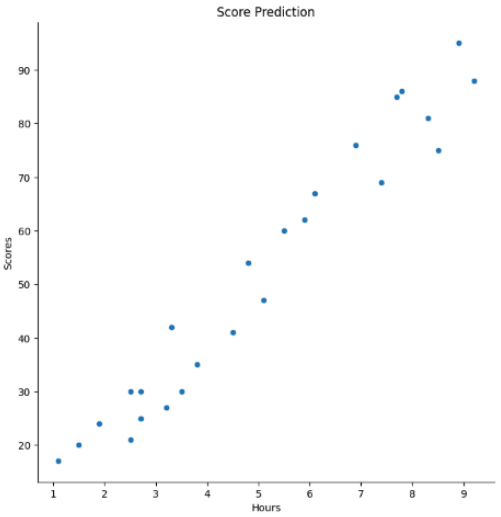
**Gambar 2. Menunjukan data Statistikal**

```
# Disini untuk menampilkan data statistikal dasar dari kumpulan data yang kita
data.describe()
```

| | Hours | Scores |
|---|---|---|
| **count** | 25.000000 | 25.000000 |
| **mean** | 5.012000 | 51.480000 |
| **std** | 2.525094 | 25.286887 |
| **min** | 1.100000 | 17.000000 |
| **25%** | 2.700000 | 30.000000 |
| **50%** | 4.800000 | 47.000000 |
| **75%** | 7.400000 | 75.000000 |
| **max** | 9.200000 | 95.000000 |

Tidak jauh berbeda dengan data.info selanjutnya kita perlu mengetahui nilai dari data statistikal untuk memahami data kita dengan lebih baik lagi seperti count untuk menghitung banyak data yang kita miliki, mean yang merupakan data rata - rata, lalu min max nilai terkecil dan terbesar dll.

**Gambar 3. Menampilkan Korelasi mengunakan Scatterplot**




Mencari korelasi menjadi salah satu contoh untuk melihat hubungan antara x dan y dimana **x sebagai Hours** dan **y sebagai Scores**, Menurut Arifa A (2022) analisis korelasi adalah metode evaluasi statistik yang dipergunakan untuk mempelajari kekuatan hubungan antara dua variabel yang diukur secara numerik. Korelasi positif adalah hubungan antara dua variabel di mana kedua variabel bergerak searah. Oleh karena itu, satu variabel meningkat seiring dengan peningkatan variabel lainnya, atau satu variabel menurun sedangkan variabel lainnya juga menurun.

Korelasi yang ditunjukan merupakan positif yang berarti dapat diambil kesimpulan bahwa dari data yang disajikan semakin tinggi jam atau waktu belajar murid maka akan semakin tinggi pula nilainya.

**Gambar 4. Membersihkan Duplikasi Data**

```
#Melakukan chek dan membuang jika terdapat data yang duplikat
duplicate_rows_before = df[df.duplicated()]
duplicate_rows_before
df = df.drop_duplicates()
print("Data Sesudah pemeriksaan Duplicate")
print(df.shape)
```

```
Data Sesudah pemeriksaan Duplicate
(25, 2)
```

Feature engineering dilakukan untuk memperkuat modeling dalam artian proses pembuatan, modifikasi, atau pemilihan fitur (variabel) dari data mentah untuk meningkatkan kinerja model machine learning. Fitur yang baik dapat

membantu model machine learning untuk memahami data lebih baik dan membuat prediksi yang lebih akurat. Oleh karena itu kita perlu mengecek apakah ada data duplikasi , jika ada lakukan drop pada data untuk membuang data yang duplikat.

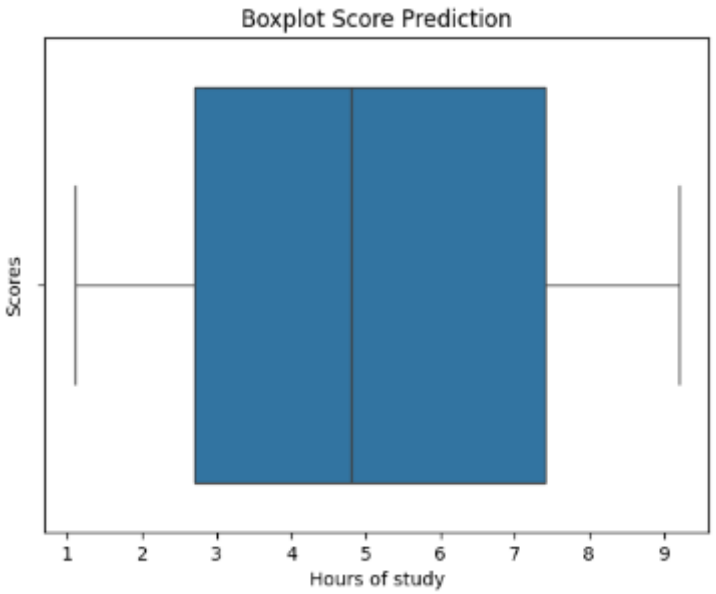
**Gambar 5. Melakukan Handling Missing Values**

```
#Memeriksa nilai yang null atau kosong
df.isna().sum()
```

```
Hours     0
Scores    0
dtype: int64
```

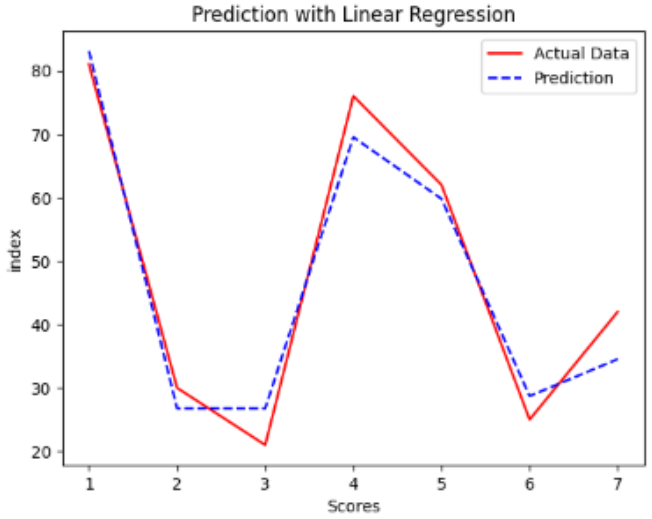
Langkah Selanjutnya setelah melakukan pembersihan duplikasi, maka kita perlu melakukan handling missing values, artinya kita perlu melakukan pembersihan baris terhadap kolom yang tidak memiliki nilai atau null. Dari data yang ditampilkan menampilkan angka nol yang berarti sudah tidak ada lagi nilai null atau sebuah kolom yang berisikan nilai null.

**Gambar 6. Analisa Outlier menggunakan Boxplot**




Dalam sebuah data, terkadang ada beberapa kolom yang diluar dari lingkup data yang tersedia dalam artian outlier.

**Gambar 7. Prediksi nilai Regresi Linear**




Dari grafik prediksi dimana garis merah merupakan data asli dari dataset dan garis biru putus - putus merupakan data prediksi menggunakan model dari Regresi Linear. Jika garis biru mengikuti garis merah dengan sangat dekat hal ini menandakan bahwa model regresi memiliki kinerja yang baik dalam memprediksi nilai ujian berdasarkan data pengujian begitu pula sebaliknya jika terdapat banyak perbedaaan atau jarak garis biru dan garis merah, menunjukan bahwa adanya kesalahan prediksi yang signifikan. Nilai y sebagai index dan x sebagai scores menjadi objek.

**Gambar 8. Nilai r square pada Model Regresi Linear**

```
print('r square Linear Regression:',rsq)
```

```
r square Linear Regression: 0.9553509219739938
```

Hasil r pada model regresi linear adalah 0.955 dimana hal ini sangat kuat untuk memprediksi data. karena semakin mendekati nilai 1 maka akan semakin kuat.

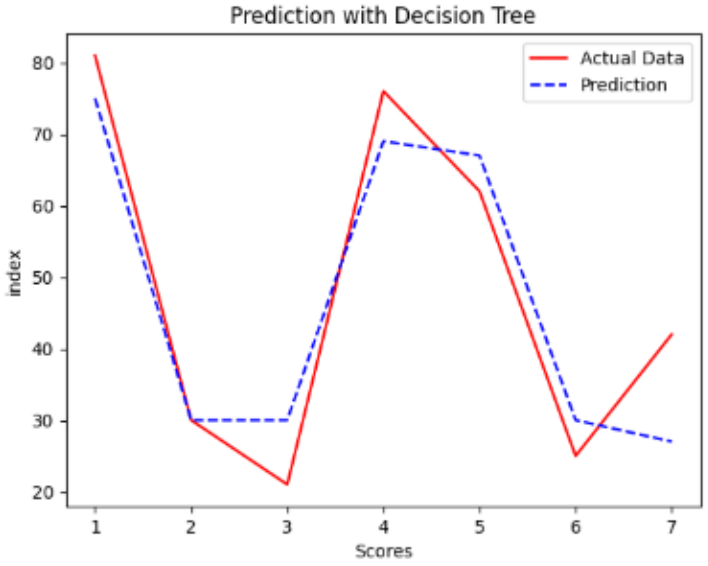
**Gambar 9. Nilai Alpha dan Betha pada Regresi Linear**

```
Intercept of the Linear Regression model: 2.4803670915057623
Coefficient of the line Linear Regression: [9.71409219]
```

$$y = 24803.67 + 9714.092\ x$$

Nilai intercept disini merupakan alpha sedangkan nilai coefficient merupakan betha dengan menggunakan model atau rumus dari regresi linear yaitu y = a + bx.

**Gambar 10. Prediksi nilai pada Model DecisionTree Regressor**




Sama seperti penjelasan pada model regresi sebelumnya. Semakin dekat garis merah dan biru maka model yang digunakan semakin baik dalam melakukan prediksi. Nilai y sebagai index dan x sebagai scores menjadi objek.

**Gambar 11. Nilai r square pada Model Decision Tree**

```
print('r square Decision Tree Results:',rsq_dt)
```

```
r square Decision Tree Results: 0.8803859268443893
```

Nilai r square yang didapat dari hasil modeling menggunakan Decision Tree Regresor adalah 0.88 tentu angka yang sangat kuat jika dilihat pada gambar 12 yang berisi tabel untuk melihat kekuatan korelasi.

**Gambar 12. Table nilai r pada modeling**

| Rentang Nilai $r$ | Kekuatan Korelasi | Interpretasi |
|---|---|---|
| 0.00 - 0.19 | Sangat Lemah | Tidak ada hubungan linear atau hampir tidak ada hubungan linear antara dua variabel. Hubungan sangat lemah atau tidak signifikan. |
| 0.20 - 0.39 | Lemah | Hubungan linear lemah antara dua variabel. Meskipun ada kecenderungan, hubungan tersebut tidak cukup kuat untuk diandalkan. |
| 0.40 - 0.59 | Sedang | Hubungan linear sedang antara dua variabel. Terdapat keterkaitan, namun tidak cukup kuat untuk digunakan sebagai prediktor utama. |
| 0.60 - 0.79 | Kuat | Hubungan linear kuat antara dua variabel. Terdapat keterkaitan yang signifikan dan dapat diandalkan dalam banyak kasus. |
| 0.80 - 1.00 | Sangat Kuat | Hubungan linear sangat kuat antara dua variabel. Keterkaitan sangat signifikan, mendekati hubungan sempurna. Model sangat dapat diandalkan untuk prediksi. |

Merupakan tabel yang berisi rentang nilai r , kekuatan korelasi dan interpretasi yang didapat dari nilai r yang kita ketahui.

**Gambar 13. Perbandingan nilai r , untuk pemodelan terbaik**

```
# Membandingkan nilai R-squared ( Yang tertinggi adalah yang terbaik)
if r_squared_linear_regression > r_squared_decision_tree:
    print("Model Linear Regression memiliki R-squared yang lebih tinggi:", r_squared_linear_regression)
    print("Kesimpulan: Model Linear Regression lebih baik dalam menjelaskan variasi dalam data.")
elif r_squared_decision_tree > r_squared_linear_regression:
    print("Model Decision Tree memiliki R-squared yang lebih tinggi:", r_squared_decision_tree)
    print("Kesimpulan: Model Decision Tree lebih baik dalam menjelaskan variasi dalam data.")
```

```
Model Linear Regression memiliki R-squared yang lebih tinggi: 0.9553509219739938
Kesimpulan: Model Linear Regression lebih baik dalam menjelaskan variasi dalam data.
```

Perbandingan dilakukan untuk melihat mana model yang paling baik dengan goals nilai r tertinggi lah yang memiliki kemampuan prediksi paling baik dalam memprediksi data dari dataset yang kita miliki. Dimana hasilnya yaitu Model Regresi menjadi model terbaik dengan nilai 0.95 dibanding dengan nilai r pada model decision tree yaitu 0.88.

**Link Google Collabs :**

**https://colab.research.google.com/drive/1p04GlQqb4SDlYy6bDUBq4v9vXD50XmKc#scrollTo=8csz3YxU5zm4**

**Link DataSet :**

**https://drive.google.com/drive/folders/11RYAMn2awK8TU7kY0_DhrQIZib1JqAyl?hl=id**

# BAB IV

# KESIMPULAN DAN SARAN

## 4.1 KESIMPULAN

Proyek "Memprediksi Nilai Siswa Berdasarkan Jumlah Jam Belajar" menggunakan teknologi kecerdasan buatan (AI) dan metode machine learning, seperti regresi linear dan decision tree, untuk memahami dan memprediksi hubungan antara jumlah jam belajar dan nilai siswa. AI memungkinkan analisis kompleks dan prediksi akurat, sementara regresi linear memberikan model sederhana untuk memahami pengaruh linier, dan decision tree menangkap hubungan non-linear dalam data belajar siswa. Dengan menggunakan Google Colab dan library Python seperti NumPy, Pandas, dan Scikit-learn, proyek ini berhasil mengembangkan dan menguji model prediksi secara efisien. Hasil dari model ini memberikan wawasan berharga yang dapat digunakan siswa untuk merencanakan waktu belajar lebih efektif, guru untuk mengidentifikasi siswa yang memerlukan perhatian lebih, dan pembuat kebijakan untuk merancang program pendidikan yang lebih baik. Proyek ini menunjukkan bahwa AI dan machine learning dapat memberikan kontribusi signifikan dalam meningkatkan hasil belajar siswa melalui pemahaman yang lebih mendalam tentang pengaruh jam belajar terhadap nilai akademik.

Dalam melakukan modeling kita perlu mengetahui data kita seperti apa, informasi yang dapat kita terima, lalu kita lakukan feature engineering untuk meningkatkan performa modeling kita , mencari duplikasi,mengecek nilai kosong, dan melakukan analisa untuk melihat outlier, sampai kepada melakukan modeling dari 2 model yaitu Regresi Linear dan Decision Tree yang kemudian diuji nilai r sebagai tolak ukur untuk menilai model mana yang terbaik berdasarkan nilai r tertinggi dari skala 0 sampai 1.

## 4.2 SARAN

Tentu proyek ini masih jauh dari kata sempurna karena kami adalah mahasiswa yang sedang belajar untuk mengembangkan AI yang lebih bermanfaat bagi masyarakat umum. Oleh karena itu, kami sangat mengharapkan saran dan masukan untuk membantu kami menyempurnakan dan mengembangkan proyek ini lebih lanjut. Saran dan masukan yang kami dapatkan akan sangat berarti bagi kemajuan proyek kami.

# DAFTAR PUSTAKA


1. **Draper, N. R., & Smith, H. (1998).** Applied Regression Analysis. John Wiley & Sons.
2. **Quinlan, J. R. (1986).** Induction of Decision Trees. Machine Learning, 1(1), 81-106.
3. **Van Rossum, G., & Drake, F. L. (2009).** Python 3 Reference Manual. CreateSpace Independent Publishing Platform