

**BUSSINES UNDERSTANDING, DATA UNDERSTANDING, DAN DATA
PREPARATION PREDIKSI PENJUALAN MINYAK DUNIA**

Disusun untuk Memenuhi Tugas Mata Kuliah Penambangan Data

Dosen Pengampu: Indra Waspada S.T, M.T.I



Disusun Oleh:

Naufal Daffa Pramudya

(30000323410019)

**MAGISTER SISTEM INFORMASI
SEKOLAH PASCASARJANA
UNIVERSITAS DIPONEGORO
SEMARANG
2024**

KATA PENGANTAR

Puji syukur kehadiran Allah SWT atas rahmat berkat dan karunia-Nya, penulis dapat menyelesaikan tugas Penambangan Data dengan judul “Prediksi penjualan Minyak Dunia“. Tugas akhir ini dibuat sebagai salah satu syarat untuk memenuhi Ujian Akhir Semeseter (UAS) Penambangan Data Bidang Magister Sistem Informasi, Fakultas Pascasarjana, Universitas Diponegoro. Penulis menyadari bahwa tugas ini tidak mungkin akan terselesaikan tanpa adanya bantuan dan peran serta dari berbagai pihak. Penulis juga terbuka akan adanya suatu diskusi dan saran yang membantu dalam penyempurnaan artikel ini.

DAFTAR ISI

COVER.....	<i>i</i>
KATA PENGANTAR.....	<i>ii</i>
DAFTAR ISI.....	<i>iii</i>
DAFTAR GAMBAR.....	<i>iv</i>
BAB I PENDAHULUAN.....	<i>1</i>
1.1 Latar Belakang.....	<i>1</i>
1.2 Tujuan	<i>2</i>
1.3 Manfaat	<i>2</i>
1.4 Batasan Masalah	<i>2</i>
BAB II PEMBAHASAN.....	<i>3</i>
2.1 <i>Business Understanding</i>	<i>3</i>
2.2 <i>Data Understanding</i>	<i>3</i>
2.3 <i>Data Preparation</i>	<i>4</i>
2.3.1 Pengumpulan Data	<i>4</i>
2.3.2 <i>Data Cleaning</i>	<i>5</i>
2.3.3 Normalisasi Data	<i>9</i>
DAFTAR PUSTAKA.....	<i>11</i>

DAFTAR GAMBAR

Gambar 2. 1 Atribut Data Understanding	3
Gambar 2. 2 Dataset Penjualan Minyak.....	4
Gambar 2. 3 Sampel Dataset	4
Gambar 2. 4 Mengecek Nilai Null pada Dataset	5
Gambar 2. 5 Menghapus duplikat data.....	6
Gambar 2. 6 Mencari Nilai Q1, Q3, IQR, Lower Bound, dan Upper Bound.....	7
Gambar 2. 7 Outlier Data Pada Kolom "Sales"	8
Gambar 2. 8 Informasi Dataset Setelah Pengecekan Outlier	8
Gambar 2. 9 Drop Kolom "Company ID" dan Konversi Tipe Data.....	9
Gambar 2. 10 Dataset Setelah Dinormalisasi	10

BAB I

PENDAHULUAN

1.1 Latar Belakang

Industri minyak adalah salah satu industri utama di dunia yang memiliki dampak besar pada perekonomian global (Purnamasari dkk., 2017). Minyak mentah tidak hanya digunakan sebagai sumber energi utama, tetapi juga sebagai bahan baku untuk berbagai industri, termasuk petrokimia, manufaktur, dan transportasi. Karena perannya yang vital dalam memenuhi kebutuhan energi dunia, fluktuasi dalam penjualan minyak dunia memiliki dampak yang signifikan pada stabilitas ekonomi global (Raudha Hanoeboen, 2017).

Dalam konteks prediksi untuk menentukan penjualan atau sales sangat dipengaruhi oleh berbagai atribut, Prediksi yang didapatkan akan dilakukan analisis yang memungkinkan perusahaan, pemerintah, investor, dan pemangku kepentingan lainnya untuk memahami penjualan, tren, serta insight dari penjualan minyak. Melalui prediksi yang akurat, pemodelan, deployment, evaluation serta implementasi yang baik dari metode yang digunakan akan dapat membantu suatu perusahaan dalam mengetahui insight dari penjualan minyak ini (Shimaa, 2024).

Pada prediksi penjualan minyak ini akan memberikan pemahaman yang mendalam tentang kinerja dari data penjualan, kondisi pasar global, serta dampaknya pada perekonomian secara keseluruhan. Dengan demikian prediksi berbasis data penjualan minyak dunia menjadi alat yang sangat penting bagi berbagai pihak untuk mengambil keputusan yang tepat dan mengelola risiko dengan lebih baik dalam lingkungan bisnis yang dinamis dan kompleks.

1.2 Tujuan

1. Memahami tren penjualan minyak dunia dalam beberapa periode waktu tertentu.
2. Mengetahui faktor-faktor yang mempengaruhi penjualan minyak dunia pada berbagai perusahaan.
3. Mengembangkan model prediksi yang akurat dalam memproyeksikan penjualan minyak di masa depan

1.3 Manfaat

Dengan melakukan berbagai metode prediksi yang dilakukan data penjualan minyak dunia memberikan manfaat yang signifikan dalam pemahaman mendalam tentang industri minyak, tren penjualan pada masa depan, pengambilan keputusan yang lebih tepat, identifikasi peluang dan risiko, dan perencanaan strategis yang lebih efektif. Dengan demikian, prediksi penjualan minyak dunia menjadi alat penting bagi perusahaan dan pemangku kepentingan lainnya dalam mengelola risiko, mengambil keputusan yang tepat, dan mencapai tujuan bisnis terkait dengan penjualan minyak dunia.

1.4 Batasan Masalah

Penelitian ini akan dibatasi pada tahapan awal analisis, yang meliputi *Business Understanding*, *Data Understanding*, dan *Data Preparation*. Batasan masalah ini akan mengarah pada pemahaman yang lebih mendalam tentang konteks bisnis, data yang digunakan, serta persiapan data untuk analisis lebih lanjut.

BAB II

PEMBAHASAN

2.1 *Business Understanding*

Terdapat beberapa aspek bisnis yang dapat dipahami dan dieksplorasi diantaranya:

1. Prediksi Sales

Tujuan dari adanya prediksi tren sales untuk melihat apakah penjualan sales dari berbagai perusahaan mengalami tren kenaikan atau penurunan dan dapat digunakan dalam melihat prediksi dari penjualan, dengan melihat pengaruh dari berbagai Dari prediksi sales ini, kita dapat mengetahui pergerakan sales pada maa mendatang. Hasil dari prediksi sales ini juga dapat digunakan sebagai evaluasi untuk penjualan, kenapa mengalami kenaikan dan ketika terdapat penurunan kenapa tren dapat turun yang dapat membantu dalam memprediksi Sales minyak dunia.

2.2 *Data Understanding*

Data understanding adalah tahap pengumpulan koleksi data awal, dan melakukan proses pengenalan terhadap data tersebut dengan tujuan untuk lebih mengenal *nature* dari data yang akan dipakai. Proses *data understanding* ini berguna untuk dapat mengetahui isi dari data tiap label, dalam proses termasuk eksplorasi data, pemahaman konteks bisnis, hingga pemahaman atribut pada dataset.

	Data	Variabel
INPUT	Company	1. Company name 2. Asset 3. Liabilities 4. Net Income
	Strategic or Product Development	1. CapEx 2. R&D
	Rasio Scoring	1. Tobin_q 2. SD of Tobin_q 3. HHI 4. AltmanZ
	Customer Metrics	1. NPS
	Transaction	1. Quarter 2. Year 3. Sales
Output	Prediksi Sales	Prediction Sales

Gambar 2. 1 Atribut *Data Understanding*

Dari hasil atribut *data understanding* dapat disimpulkan dataset memiliki 5 data *input* dengan 14 variabel, dan *output*-nya yaitu 3 analisis diantaranya analisis profit, alokasi investasi, dan tren sales, yang diperoleh dari hasil pengolahan dataset.

2.3 Data Preparation

2.3.1 Pengumpulan Data

Data yang dipakai diperoleh dari hasil observasi salah satu anggota kelompok yang merupakan proyek nyata yang pernah dikerjakan, dimana data yang diambil dari tahun 1990-2011.

```
import pandas as pd

data = pd.read_csv("data.csv", delimiter=';')

data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 388245 entries, 0 to 388244
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Company ID            388245 non-null  int64
1   Assets                388245 non-null  object
2   Liabilities           388245 non-null  object
3   Net Income            388245 non-null  object
4   CapEx                 388245 non-null  object
5   R&D                   388245 non-null  object
6   Quarter               388245 non-null  int64
7   Tobin_q               388245 non-null  object
8   NPS                   388245 non-null  object
9   SD of TobinQ          388245 non-null  object
10  HHI                   388245 non-null  object
11  Year                  388245 non-null  int64
12  AltmanZ               388245 non-null  object
13  Sales                 388245 non-null  object
dtypes: int64(3), object(11)
memory usage: 41.5+ MB
```

Gambar 2. 2 Dataset Penjualan Minyak

Dari gambar 2.2 diketahui dataset memiliki 14 atribut dengan total baris sebanyak 388.245.

data.head()													
	Company ID	Assets	Liabilities	Net Income	CapEx	R&D	Quarter	Tobin_q	NPS	SD of TobinQ	HHI	Year	Sales
0	105920	2769,69	2135,958	1,972	1,009	0,003111725	1	0,00549128	0,000711993	4,539607	0	2000	0,58036596 331,936
1	105920	2872,301	2225,633	15,743	2,722	0,024344794	2	0,005536071	0,005480972	4,539607	0	2000	0,58502036 333,341
2	105920	2962,789	2320,762	-3,313	3,896	-0,00516022	3	0,00580661	-0,001118203	4,539607	0	2000	0,54423416 342,813
3	105920	2875,184	2217,859	17,275	5,646	0,026280759	4	0,005726239	0,006008311	4,539607	0	2000	0,63241285 352,901
4	105920	2818,117	2182,371	-19,96	0,32	-0,031396188	1	0,00597723	-0,007082744	4,6744113	0,042366602	2001	0,57448918 345,999

Gambar 2. 3 Sampel Dataset

Gambar 2.3 menampilkan sampel dataset yang diambil dari 5 baris data teratas.

2.3.2 Data Cleaning

Cleansing Data merupakan suatu proses mengidentifikasi, mengoreksi dan menghapus data yang dianggap *null* atau hilang dari sebuah dataset yang ada. *Cleansing* ini memiliki berbagai tahapan seperti menangani nilai yang hilang, mendeteksi *duplicate* data, dan menangani *outlier* pada data.

Tahapan-tahapan *data cleaning*:

1. Menangani nilai yang hilang atau null

```
data_isnull = data.isnull()
print(data_isnull)
```

	Company ID	Assets	Liabilities	Net Income	CapEx	R&D	Quarter	\
0	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	
...	
388240	False	False	False	False	False	False	False	
388241	False	False	False	False	False	False	False	
388242	False	False	False	False	False	False	False	
388243	False	False	False	False	False	False	False	
388244	False	False	False	False	False	False	False	

	Tobin_q	NPS	SD of TobinQ	HHI	Year	AltmanZ	Sales
0	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False
...
388240	False	False	False	False	False	False	False
388241	False	False	False	False	False	False	False
388242	False	False	False	False	False	False	False
388243	False	False	False	False	False	False	False
388244	False	False	False	False	False	False	False

[388245 rows x 14 columns]

Gambar 2. 4 Mengecek Nilai *Null* pada Dataset

Gambar 2.4 menunjukkan data dari semua kolom bernilai *false* yang artinya tidak ada data yang bernilai *null*.

2. Menangani duplikasi data

```
# Menghapus duplikasi data
data.drop_duplicates(inplace=True)

# Informasi data setelah penghapusan duplikasi
print("\nInformasi data setelah penghapusan duplikasi:")
print(data.info())
```

```
Informasi data setelah penghapusan duplikasi:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 388245 entries, 0 to 388244
Data columns (total 14 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Company ID            388245 non-null  int64
1   Assets                388245 non-null  object
2   Liabilities            388245 non-null  object
3   Net Income             388245 non-null  object
4   CapEx                 388245 non-null  object
5   R&D                   388245 non-null  object
6   Quarter               388245 non-null  int64
7   Tobin_q               388245 non-null  object
8   NPS                   388245 non-null  object
9   SD of TobinQ          388245 non-null  object
10  HHI                   388245 non-null  object
11  Year                  388245 non-null  int64
12  AltmanZ               388245 non-null  object
13  Sales                 388245 non-null  object
dtypes: int64(3), object(11)
memory usage: 41.5+ MB
None
```

Gambar 2. 5 Menghapus duplikat data

Gambar 2.5 menambahkan kode untuk mengecek duplikasi data, jika ada duplikasi data, maka baris data tersebut akan dihapus dan menyisakan satu baris data saja. Dapat diketahui bahwa dataset tidak ditemukan duplikasi data karena jumlah baris tetap sama yaitu 388.245.

3. Menangani outlier data

Outlier Data adalah nilai yang jauh berbeda dari mayoritas data dalam kumpulan data, *outlier* ini memiliki fungsi sebagai indikator perubahan mendadak dalam anomali data. Namun pada proses ini memerlukan kehati-hatian dikarenakan dengan *outlier* maka bisa menghilangkan beberapa informasi penting yang terdapat pada dataset. Dalam studi kasus ini atribut yang dicek nilai *outlier*-nya adalah kolom “Sales”.

```

import pandas as pd

# Membaca data dari file CSV
data = pd.read_csv("data.csv", delimiter=';')

# Menghapus tanda koma dari nilai pada kolom "Sales"
data['Sales'] = data['Sales'].str.replace(',', '.')

# Mengubah nilai pada kolom "Sales" menjadi tipe data integer
data['Sales'] = data['Sales'].astype(float)

# Menghitung nilai batas atas dan batas bawah menggunakan IQR
Q1 = data['Sales'].quantile(0.25)
Q3 = data['Sales'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

print("Nilai Q1:", Q1)
print("Nilai Q3:", Q3)
print("Nilai IQR:", IQR)
print("Nilai lower_bound:", lower_bound)
print("Nilai upper_bound:", upper_bound)

# Menampilkan outlier pada kolom "Sales"
outliers = data[(data['Sales'] < lower_bound) | (data['Sales'] > upper_bound)]
print("\nOutlier pada kolom 'Sales':")
print(outliers)

Nilai Q1: 23.171
Nilai Q3: 291.017
Nilai IQR: 267.846
Nilai lower_bound: -378.598
Nilai upper_bound: 692.7860000000001

```

Gambar 2. 6 Mencari Nilai Q1, Q3, IQR, Lower Bound, dan Upper Bound

Gambar 2.6 menampilkan hasil dari nilai Q1, Q3, IQR, *Lower Bound*, dan *Upper Bound* pada atribut “Sales”, yaitu masing-masing bernilai 23.171, 291.017, 267.846, -378.598, dan 692.786.

Outlier pada kolom 'Sales':

Company ID	Assets	Liabilities	Net Income	CapEx	R&D
708	1010	4832,1	3084,3	262,2	225,2
739	186452	201,929	136,474	57,709	9,181
740	14438	8692	3526	239	269
741	14438	8739	3508	93	321
742	14438	8778	3500	93	86
...
388139	61034	10406,382	7383,567	-105,129	292,917
388140	61034	11257,262	7856,384	183,419	410,495
388141	61034	11700,259	8245,918	103,128	546,334
388142	61034	10215,001	7562,596	-650,873	716,765
388155	176259	2040,642	1367,154	213,772	170,63

Quarter	Tobin_q	NPS	SD of TobinQ	HHI	Year
708	4	0	0,065106265	4,9548321	0
739	4	0	0	4,5879631	0
740	3	0	0,027496548	5,0319152	0
741	3	0	0,01064195	5,0267253	0
742	4	0	0,010594669	4,0354877	0
...
388139	1	0,001095912	-0,010102359	4,852932	0,28011999
388140	2	0,001032533	0,016293393	4,852932	0,43360639
388141	3	0	0,008814164	4,852932	0,49260539
388142	4	0,000512463	0	4,852932	0,20557092
388155	4	0	0,10475723	5,509737	0

AltmanZ	Sales
708	1,411937
739	6,2528429
740	2,1085899
741	1,8667037
742	1,8795494
...	...
388139	0,71807152
388140	0,81540996
388141	0
388142	0
388155	1,6601719

[48527 rows x 14 columns]

Gambar 2. 7 Outlier Data Pada Kolom "Sales"

Gambar 2.7 menampilkan baris data yang terdapat *outlier* pada kolom “Sales” yaitu berjumlah 48.527 data.

```
# Menghapus outlier
data = data[(data['Sales'] >= lower_bound) & (data['Sales'] <= upper_bound)]

# Informasi data setelah pemrosesan outlier
print("\nInformasi data setelah pemrosesan outlier:\n")
print(data.info())
```

Informasi data setelah pemrosesan outlier:

```
<class 'pandas.core.frame.DataFrame'>
Index: 339718 entries, 0 to 388244
Data columns (total 14 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Company ID      339718 non-null  int64
1   Assets          339718 non-null  object
2   Liabilities     339718 non-null  object
3   Net Income      339718 non-null  object
4   CapEx          339718 non-null  object
5   R&D            339718 non-null  object
6   Quarter         339718 non-null  int64
7   Tobin_q         339718 non-null  object
8   NPS             339718 non-null  object
9   SD of TobinQ    339718 non-null  object
10  HHI             339718 non-null  object
11  Year            339718 non-null  int64
12  AltmanZ         339718 non-null  object
13  Sales           339718 non-null  float64
dtypes: float64(1), int64(3), object(10)
memory usage: 38.9+ MB
None
```

Gambar 2. 8 Informasi Dataset Setelah Pengecekan *Outlier*

Gambar 2.8 menampilkan informasi dataset setelah dilakukan pengecekan *outlier* pada kolom “Sales”, dimana jumlah data yang dihapus sebanyak 48.527, sehingga total akhir jumlah data adalah $388.245 - 48.527 = 339.718$.

2.3.3 Normalisasi Data

Normalisasi merupakan salah satu teknik yang umum digunakan dalam praproses data, terutama saat bekerja dengan algoritma pembelajaran mesin atau model statistik yang sensitif terhadap skala nilai fitur. Normalisasi memastikan bahwa semua fitur dalam dataset memiliki skala yang seragam. Ini mencegah fitur-fitur dengan skala yang besar mendominasi dalam proses pembelajaran, sehingga mencegah bias yang tidak diinginkan dalam model.

Sebelum melakukan normalisasi kolom “Company ID” akan dihapus karena kolom tersebut tidak dibutuhkan dalam tahap pengolahan data, dan juga kolom yang bertipe data *object* akan diubah tipe datanya menjadi *float*.

```
# Menghapus kolom "Company ID"
data.drop(columns=['Company ID'], inplace=True)

data['Assets'] = data['Assets'].str.replace(',', '.').astype(float)
data['Liabilities'] = data['Liabilities'].str.replace(',', '.').astype(float)
data['Net Income'] = data['Net Income'].str.replace(',', '.').astype(float)
data['CapEx'] = data['CapEx'].str.replace(',', '.').astype(float)
data['R&D'] = data['R&D'].str.replace(',', '.').astype(float)
data['NPS'] = data['NPS'].str.replace(',', '.').astype(float)
data['Tobin_q'] = data['Tobin_q'].str.replace(',', '.').astype(float)
data['SD of TobinQ'] = data['SD of TobinQ'].str.replace(',', '.').astype(float)
data['HHI'] = data['HHI'].str.replace(',', '.').astype(float)
data['AltmanZ'] = data['AltmanZ'].str.replace(',', '.').astype(float)

# Menampilkan informasi data setelah perubahan tipe data
print("\nInformasi data setelah perubahan tipe data:")
print(data.info())
```

```
Informasi data setelah perubahan tipe data:
<class 'pandas.core.frame.DataFrame'>
Index: 339718 entries, 0 to 388244
Data columns (total 13 columns):
#   Column          Non-Null Count  Dtype
---  ---
0   Assets           339718 non-null  float64
1   Liabilities       339718 non-null  float64
2   Net Income       339718 non-null  float64
3   CapEx            339718 non-null  float64
4   R&D              339718 non-null  float64
5   Quarter          339718 non-null  int64
6   Tobin_q          339718 non-null  float64
7   NPS              339718 non-null  float64
8   SD of TobinQ     339718 non-null  float64
9   HHI              339718 non-null  float64
10  Year             339718 non-null  int64
11  AltmanZ          339718 non-null  float64
12  Sales            339718 non-null  float64
dtypes: float64(11), int64(2)
memory usage: 36.3 MB
None
```

Gambar 2. 9 Drop Kolom "Company ID" dan Konversi Tipe Data

Gambar 2.9 menampilkan informasi dataset setelah menghapus atribut “Company ID” dan mengubah tipe data atribut yang masih *object* menjadi *float*. Diketahui jumlah atribut saat ini adalah **13** dan tipe data semuanya sudah bertipe numerik.

```

from sklearn.preprocessing import MinMaxScaler

# Normalisasi data ke dalam rentang [0,1]
scaler = MinMaxScaler()
data_normalized = scaler.fit_transform(data)

# Membuat DataFrame dari data yang sudah dinormalisasi
data_normalized = pd.DataFrame(data_normalized, columns=data.columns)

# Menampilkan informasi data setelah normalisasi
data_normalized.head()

```

	Assets	Liabilities	Net Income	CapEx	R&D	Quarter	Tobin_q	NPS	SD of TobinQ	HHI	Year	AltmanZ	Sales
0	0.002032	0.001764	0.250970	0.000139	0.518769	0.000000	0.000172	0.440434	0.591592	0.222295	0.47619	0.000030	0.479163
1	0.002110	0.001839	0.251358	0.000375	0.524128	0.333333	0.000173	0.461293	0.591592	0.222295	0.47619	0.000030	0.481191
2	0.002179	0.001917	0.250821	0.000537	0.516681	0.666667	0.000182	0.432429	0.591592	0.222295	0.47619	0.000028	0.494865
3	0.002112	0.001832	0.251401	0.000779	0.524616	1.000000	0.000179	0.463599	0.591592	0.222295	0.47619	0.000032	0.509427
4	0.002069	0.001803	0.250353	0.000044	0.510060	0.000000	0.000187	0.406341	0.639028	0.235750	0.52381	0.000029	0.499464

Gambar 2. 10 Dataset Setelah Dinormalisasi

Gambar 2.10 menampilkan dataset setelah dinormalisasi, dimana semua data dari masing-masing atribut bernilai dari rentang 0-1.

DAFTAR PUSTAKA

- Purnamasari, D., Program, M., Ekonomi, S., & Sukmana, R. (2017). *PENGARUH HARGA EMAS DUNIA, HARGA MINYAK MENTAH DUNIA DAN INDEKS PRODUKSI INDUSTRI TERHADAP INDEKS SAHAM DI JAKARTA ISLAMIC INDEX (JII) DALAM JANGKA PANJANG DAN JANGKA PENDEK (PERIODE JANUARI 2005-DESEMBER 2015)* 1). www.wikipedia.com
- Raudha Hanoebon. (2017). *ANALISIS PENGARUH HARGA MINYAK DUNIA, NILAI TUKAR RUPIAH, INFLASI DAN SUKU BUNGA SBI TERHADAP INDEKS HARGA SAHAM GABUNGAN (IHSG)*. XI(1).
- Shimaa Ouf1 (2024). *A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry*. Journal of Big Data (2024) 11:70