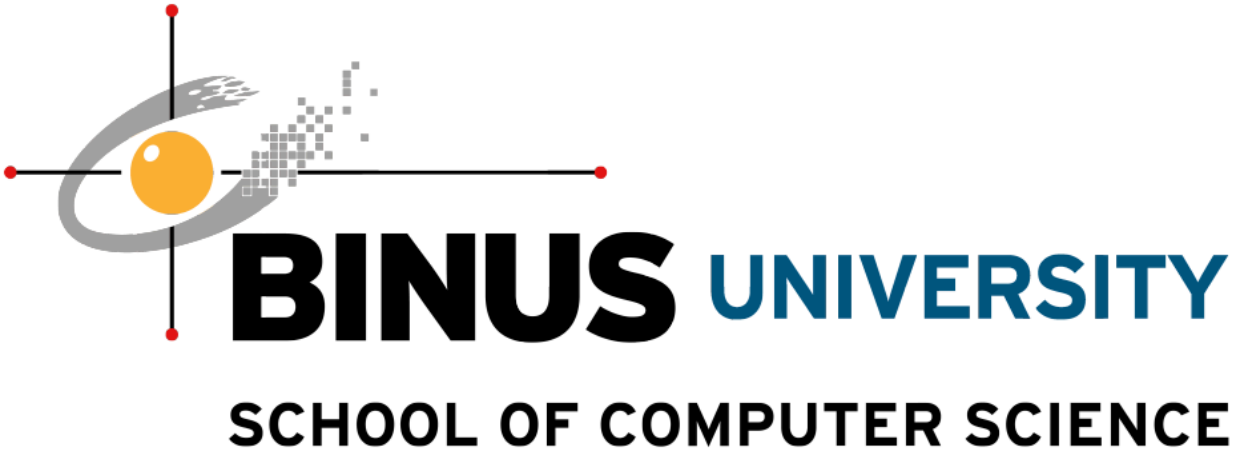
Pemodelan Faktor-Faktor yang Mempengaruhi Revenue Penjualan Smartphone dengan Korelasi, ANOVA, dan Random Forest



Nama Kelompok:
Aaron Adriano - 2702276866
Cornelius Jason Aslim - 2702337905
Leon - 2702251604
Nicholas Nathanael Lo-2702313470
Edric Emerson - 2702229592
Khuang Ming Jeremy Alexander - 2702277553

BAB I
LATAR BELAKANG

Industri smartphone telah mengalami pertumbuhan pesat dalam satu dekade terakhir. Persaingan yang semakin ketat antar produsen mendorong pentingnya pemahaman yang lebih dalam terhadap faktor-faktor yang mempengaruhi keputusan konsumen dalam membeli smartphone. Berbagai atribut seperti harga, spesifikasi teknis, merek, strategi pemasaran, dan tren konsumen memiliki potensi pengaruh terhadap tingkat penjualan suatu produk. Oleh karena itu, identifikasi faktor-faktor dominan yang mempengaruhi penjualan smartphone menjadi krusial bagi perusahaan dalam menyusun strategi bisnis yang efektif.

Dalam upaya memahami pengaruh faktor-faktor tersebut, pendekatan kuantitatif berbasis data sangat dibutuhkan. Metode statistik seperti analisis korelasi dapat digunakan untuk mengukur kekuatan hubungan antara masing-masing variabel dengan tingkat penjualan. Selain itu, analisis varians (ANOVA) dapat membantu menguji apakah terdapat perbedaan signifikan dalam penjualan berdasarkan kategori tertentu, seperti merek atau kelas harga. Untuk melengkapi analisis dan membangun model prediksi yang lebih akurat, algoritma Random Forest sebagai salah satu metode machine learning berbasis ensemble dapat digunakan untuk mengidentifikasi variabel-variabel paling berpengaruh serta memperkirakan tingkat penjualan berdasarkan data historis.

Dengan menggabungkan ketiga pendekatan tersebut korelasi, ANOVA, dan Random Forest diharapkan dapat diperoleh pemahaman yang komprehensif mengenai faktor-faktor kunci yang mempengaruhi penjualan smartphone. Hasil dari penelitian ini dapat dijadikan dasar dalam pengambilan keputusan strategis, seperti pengembangan produk, penetapan harga, serta perencanaan promosi dan distribusi.

BAB II
METODOLOGI DAN ALUR PEKERJAAN

Dataset yang digunakan dalam proyek ini adalah data penjualan produk elektronik, yang terdiri dari smartphone dan laptop, namun fokus analisis diarahkan hanya pada produk smartphone. Dataset ini mencakup berbagai atribut seperti merek, harga, tanggal masuk dan keluar stok, jumlah unit terjual, spesifikasi teknis (RAM, ROM, prosesor), lokasi pelanggan, dan wilayah penjualan.

Eksperimen ini dilakukan di lingkungan Google Colab, yang mendukung kolaborasi dalam pemrosesan data berbasis Python dan dilengkapi dengan pustaka seperti pandas, matplotlib, seaborn, serta scikit-learn untuk keperluan machine learning.

Secara garis besar, alur kerja analitik pada proyek ini terdiri dari:

2.1 Data Preparation

Tahap pertama dari proyek ini adalah **Data Preparation**, yang merupakan proses penting untuk memastikan bahwa data dalam kondisi bersih, konsisten, dan siap untuk dianalisis maupun digunakan dalam pemodelan. Langkah-langkah yang dilakukan dalam tahap ini meliputi:

- **Seleksi Data**
  Dataset awal berisi informasi penjualan dari berbagai jenis produk elektronik, termasuk smartphone dan laptop. Karena fokus dari penelitian ini adalah pada analisis penjualan smartphone, maka dilakukan **filtering** dengan memilih hanya baris data yang memiliki nilai 'Mobile Phone' pada kolom Product. Hal ini bertujuan agar analisis yang dilakukan relevan dan sesuai dengan tujuan proyek.

- **Konversi Format Tanggal**
  Kolom Inward Date dan Dispatch Date, yang menyatakan tanggal masuk stok dan tanggal produk dijual, dikonversi dari string menjadi format datetime. Konversi ini penting agar memungkinkan perhitungan selisih waktu dan analisis berbasis waktu (time-based analysis).

```
df['Inward Date'] = pd.to_datetime(df['Inward Date'])
df['Dispatch Date'] = pd.to_datetime(df['Dispatch Date'])
```

- **Perhitungan Lama Produk di Gudang**
  Dengan memanfaatkan kolom tanggal yang sudah dikonversi, dibuat kolom baru bernama DaysInStock yang merupakan selisih antara tanggal dispatch dan tanggal inward. Kolom ini dapat membantu memahami durasi rata-rata penyimpanan barang dan kemungkinan hubungannya dengan performa penjualan.

```
df['DaysInStock'] = (df['Dispatch Date'] - df['Inward Date']).dt.days
```

- **Penanganan Nilai Kosong (Missing Values) dan Mengekstrak nilai numerik**
  Dataset diperiksa untuk mendeteksi adanya nilai kosong atau tidak valid. Kolom seperti SSD, Processor Specification, dan Core Specification memiliki kemungkinan berisi nilai 'N/A' atau kosong karena tidak semua spesifikasi relevan untuk semua jenis produk proses ini dijalankan bersamaan dengan proses pengekstrakan nilai numerik pada kolom tertentu. Kolom RAM, ROM, dan SSD awalnya disimpan sebagai string dengan satuan seperti "GB" atau "TB". Oleh karena itu, satuan tersebut dihapus, dan nilai dikonversi menjadi **numerik** agar bisa digunakan dalam analisis statistik dan pemodelan machine learning. Sebagai contoh, "12GB" diubah menjadi 12, dan "1TB" diubah menjadi 1000.

```
def extract_numeric_storage(value):
    if pd.isna(value):
        return 0
    value_str = str(value).upper().replace(' ', '')
    if 'TB' in value_str:
        numeric_part = ''.join(filter(lambda x: x.isdigit() or x == '.',
value_str.replace('TB', '')))
        if numeric_part:
            return int(float(numeric_part) * 1024)
        else:
            return 0
    elif 'GB' in value_str:
        numeric_part = ''.join(filter(lambda x: x.isdigit() or x == '.',
value_str.replace('GB', '')))
        if numeric_part:
            return int(float(numeric_part))
        else:
            return 0
```

```
    try:
        return int(float(value_str))
    except ValueError:
        return 0
```

- **Pembuatan Kolom Revenue**
  Untuk mengukur pendapatan yang dihasilkan dari setiap transaksi, dibuat kolom baru Revenue yang dihitung dari hasil perkalian antara Price dan Quantity Sold. Kolom ini menjadi variabel target (target variable) dalam model prediksi yang dikembangkan pada tahap selanjutnya.

```
df['Revenue'] = df['Price'] * df['Quantity Sold']
print("Kolom 'Revenue' berhasil dibuat.")
```

2.2 **Eksplorasi Data**

Setelah data siap, dilakukan proses eksplorasi data (Exploratory Data Analysis/EDA) guna memahami struktur, pola, dan tren yang mungkin tersembunyi di dalam dataset. Analisis ini dilakukan dengan bantuan visualisasi data untuk membantu mengidentifikasi hubungan antara variabel, outlier, dan distribusi data.

Visualisasi yang digunakan antara lain:

- **Bar Chart Total Penjualan per Merek**: Menampilkan jumlah unit smartphone yang terjual berdasarkan merek. Grafik ini membantu mengidentifikasi merek mana yang paling mendominasi pasar dalam hal volume penjualan.

```
brand_sales = df_mobile_cleaned.groupby('Brand')['Quantity
Sold'].sum().sort_values(ascending=False)
brand_sales.plot(kind='bar', title='Total Quantity Sold per Brand')
plt.ylabel('Quantity Sold')
plt.show()
```

- **Rata-rata Pendapatan (Revenue) per Merek**: Visualisasi ini menunjukkan seberapa besar pendapatan rata-rata yang dihasilkan oleh masing-masing brand.

Hasilnya memberikan wawasan apakah suatu brand menjual produk mahal (margin tinggi) atau produk murah dalam jumlah besar.

```
brand_revenue =
df_mobile_cleaned.groupby('Brand')['Revenue'].mean().sort_values
(ascending=False)
brand_revenue.plot(kind='bar', color='orange', title='Average
Revenue per Brand')
plt.ylabel('Revenue')
plt.show()
```

- **Heatmap Korelasi Antar Variabel Numerik**: Digunakan untuk melihat hubungan antar fitur numerik seperti Price, Quantity Sold, RAM, ROM, dan Revenue. Korelasi tinggi antara dua variabel menunjukkan adanya potensi hubungan kausal atau asosiasi kuat yang perlu dipertimbangkan dalam analisis lanjutan.

```
sns.heatmap(df_mobile_cleaned[['Price', 'Quantity Sold', 'RAM_GB', 'ROM_GB',
'Revenue', 'DaysInStock']].corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Matrix")
plt.show()
```

Melalui eksplorasi data ini, ditemukan pola-pola penting seperti dominasi brand tertentu dalam penjualan, peran signifikan harga terhadap pendapatan, serta minimnya pengaruh langsung spesifikasi teknis terhadap revenue. EDA ini menjadi dasar penting untuk membangun dan memvalidasi model prediktif di tahap berikutnya.

2.3 Model Prediksi

Setelah proses eksplorasi data dilakukan, langkah selanjutnya adalah membangun model prediktif untuk memperkirakan pendapatan penjualan (Revenue) berdasarkan berbagai fitur dalam dataset. Tujuan utama dari pembangunan model ini adalah untuk mengetahui seberapa besar pengaruh masing-masing variabel terhadap revenue, serta membuat model yang dapat digunakan untuk memprediksi pendapatan dari produk baru berdasarkan spesifikasi dan parameter lainnya.

Model yang digunakan dalam proyek ini adalah **Random Forest Regressor**, yaitu algoritma ensemble berbasis decision tree yang sangat efektif untuk regresi dan mampu menangani dataset dengan fitur campuran (numerik dan kategorik) serta mencegah overfitting.

```
X = df.drop(['Revenue', 'Customer Name', 'Product Code', 'Dispatch Date',
'Inward Date'], axis=1)
y = df['Revenue']
```

```
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.pipeline import Pipeline

num_cols = X.select_dtypes(include=['int64', 'float64']).columns
cat_cols = X.select_dtypes(include=['object']).columns

preprocessor = ColumnTransformer([
    ('num', StandardScaler(), num_cols),
    ('cat', OneHotEncoder(handle_unknown='ignore'), cat_cols)
])
```

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_absolute_error

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)


model = Pipeline([
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(n_estimators=100, random_state=42))
])

model.fit(X_train, y_train)


y_pred = model.predict(X_test)
```

2.4 Evaluasi Model

Evaluasi model dilakukan untuk mengukur seberapa baik performa model dalam memprediksi pendapatan berdasarkan data testing.

```
print("R2 Score:", r2_score(y_test, y_pred))
print("MAE:", mean_absolute_error(y_test, y_pred))
```

Dengan hasil evaluasi ini, model yang dibangun dapat diandalkan untuk digunakan dalam memprediksi pendapatan berdasarkan parameter produk. Model ini juga memberikan dasar yang kuat untuk analisis lebih lanjut terhadap fitur-fitur yang mempengaruhi performa penjualan.

2.5 Analysis Feature Importance

Setelah model dilatih, dilakukan analisis terhadap faktor-faktor yang paling mempengaruhi prediksi pendapatan menggunakan nilai feature importance dari model Random Forest dengan cara dilakukan visualisasi faktor apa saja yang berpengaruh dalam penjualan Handphone.

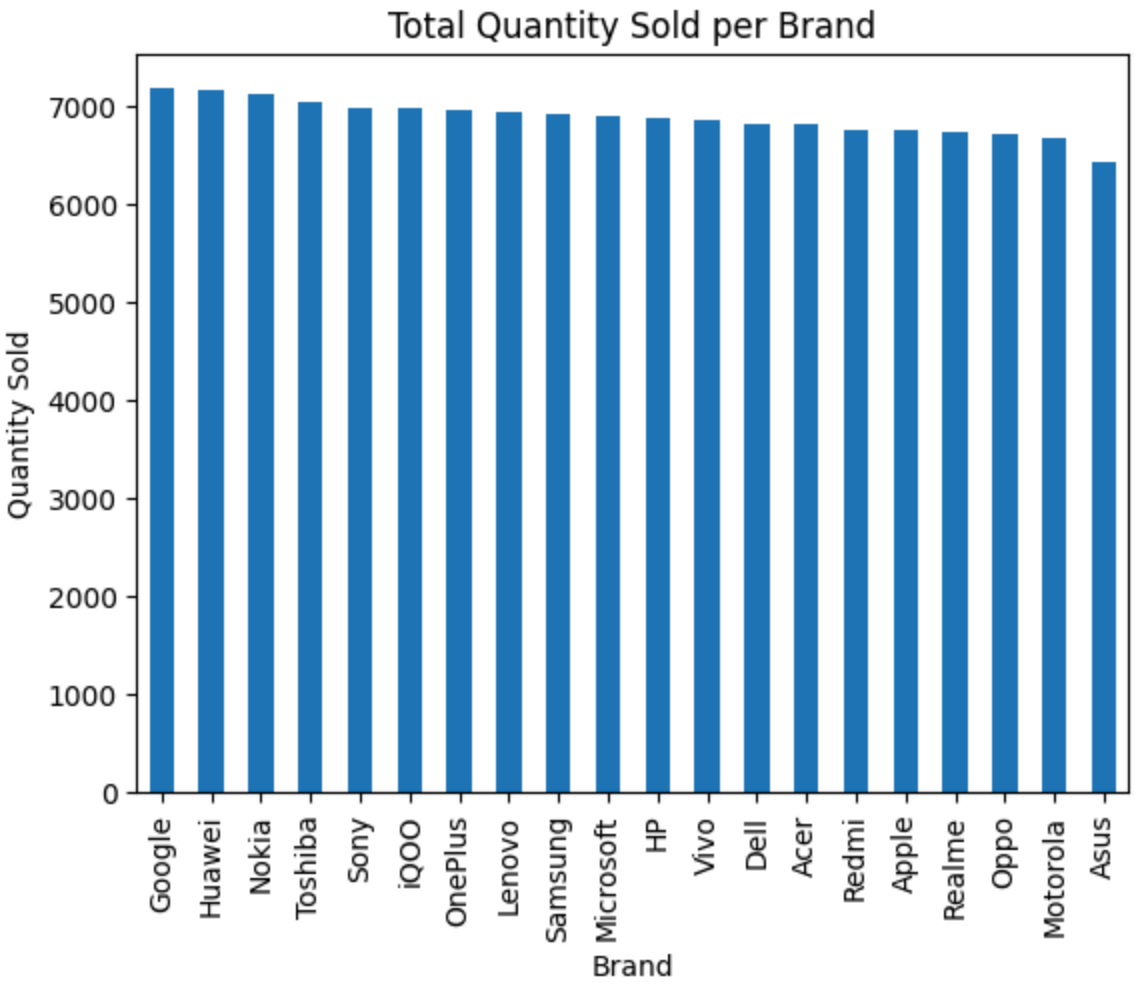
```
import matplotlib.pyplot as plt


feature_names = model.named_steps['preprocessor'].transformers_[0][2].tolist()
+ \

list(model.named_steps['preprocessor'].transformers_[1][1].get_feature_names_ou
t(cat_cols))

importances = model.named_steps['regressor'].feature_importances_
feat_imp = pd.Series(importances,
index=feature_names).sort_values(ascending=False)

feat_imp.head(10).plot(kind='barh')
plt.title('Top 10 Feature Importance')
```

```
plt.gca().invert_yaxis()
plt.show()
```

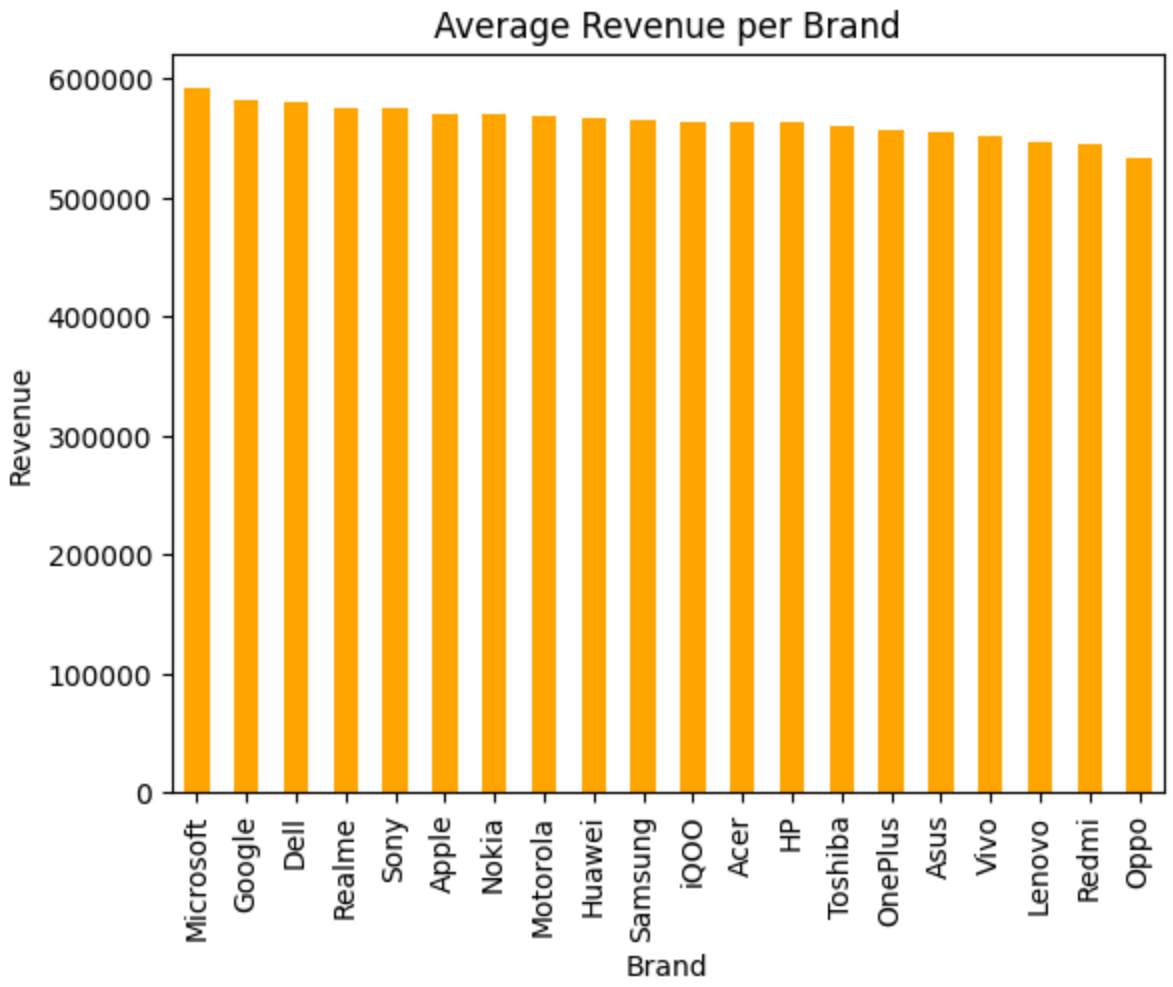
BAB III
EVALUASI DAN DETAIL DARI ALUR PEKERJAAN

Pada bab ini dijelaskan hasil dari implementasi setiap tahapan dalam alur kerja yang telah dirancang pada Bab II. Evaluasi dilakukan baik secara kuantitatif maupun kualitatif untuk menilai keberhasilan pendekatan yang digunakan, serta menguraikan lebih dalam temuan-temuan penting selama proses analisis.

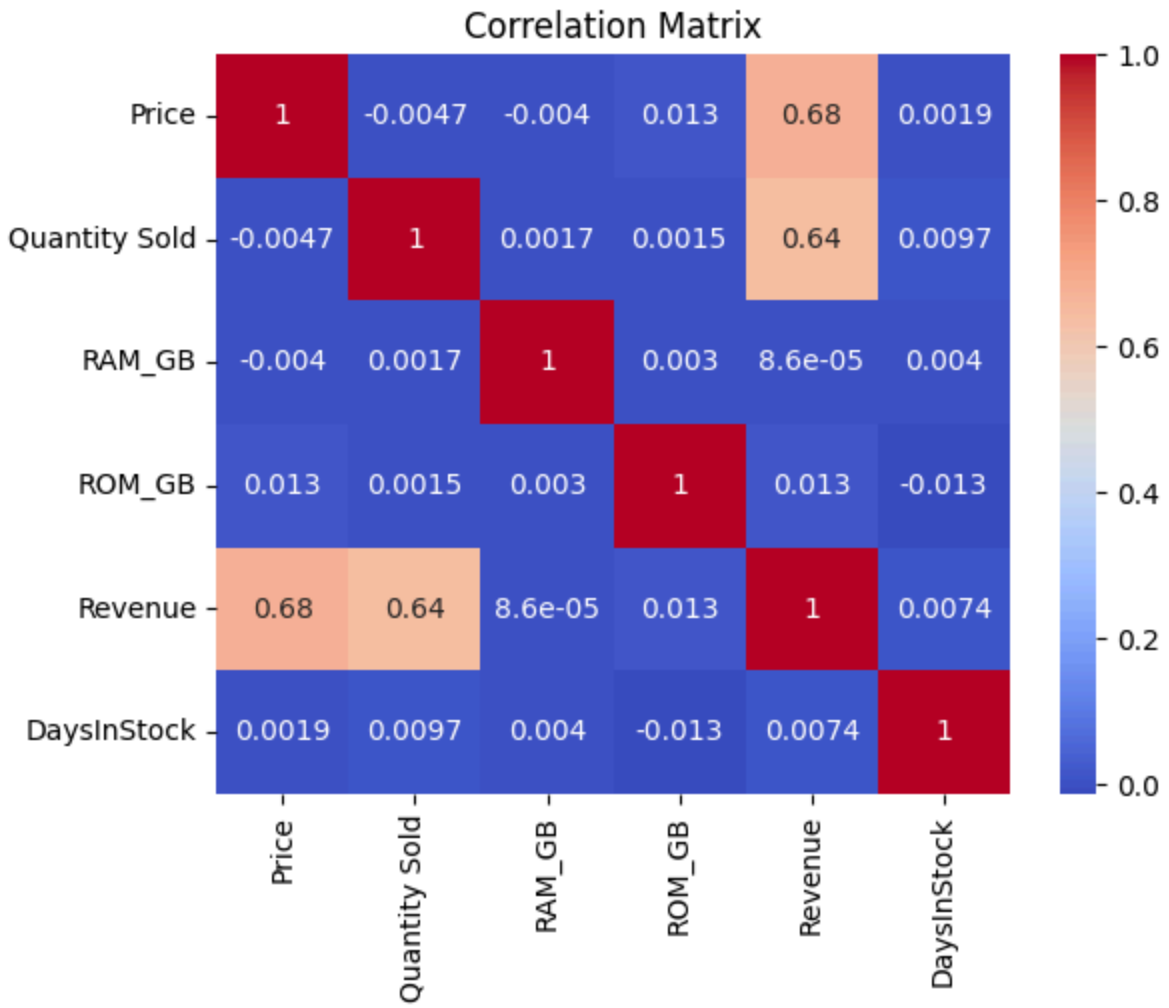
Hasil eksplorasi data memberikan sejumlah temuan penting yang menjadi dasar dari analisis lebih lanjut. Visualisasi bar chart menunjukkan bahwa merek seperti Google, Huawei, dan Nokia mendominasi penjualan berdasarkan jumlah unit terjual, menandakan bahwa brand tersebut memiliki daya tarik yang tinggi di pasar.



Di sisi lain, Microsoft dan Dell memiliki rata-rata pendapatan per produk yang lebih tinggi, yang mengindikasikan bahwa mereka menargetkan segmen premium atau menjual produk dengan margin lebih besar.



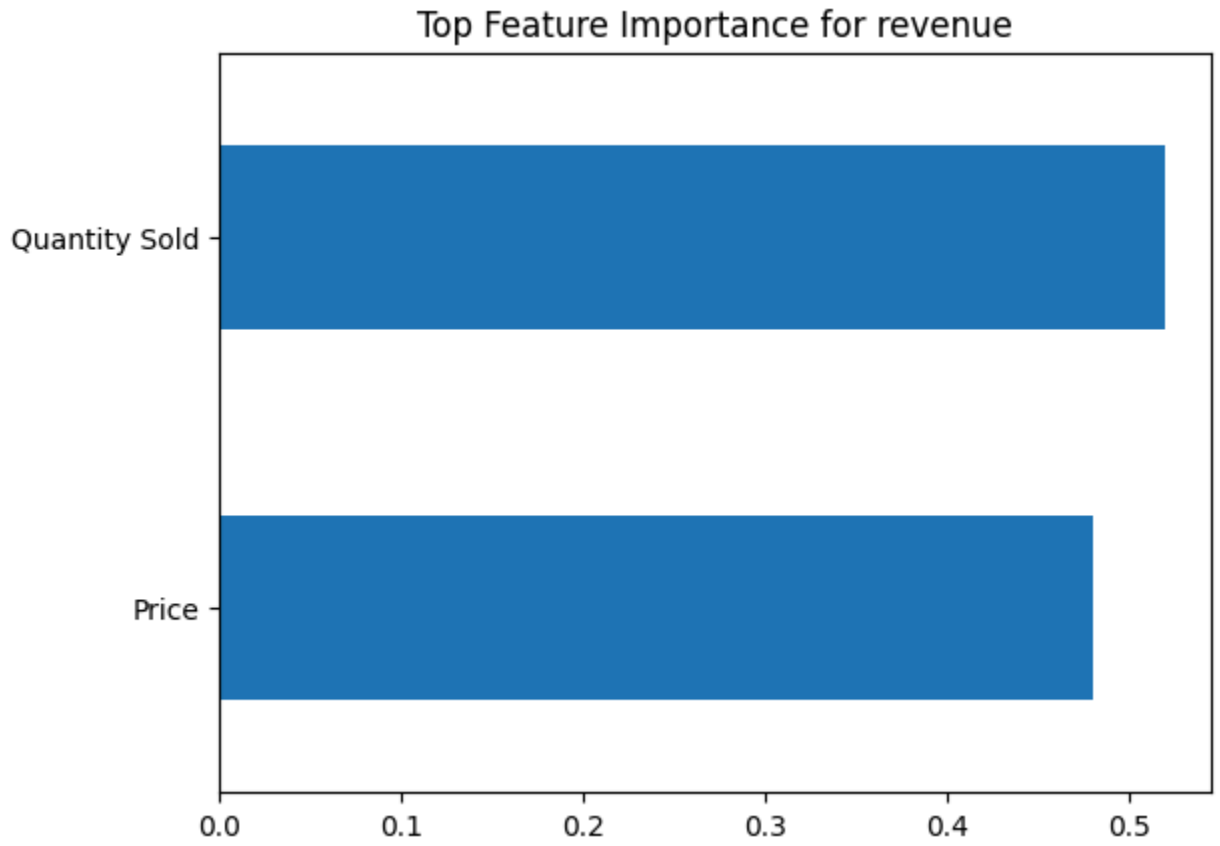
Analisis korelasi menunjukkan bahwa dua variabel yang memiliki pengaruh terbesar terhadap pendapatan adalah Price (r = +0.68) dan Quantity Sold (r = +0.64), yang berarti peningkatan pada harga atau jumlah unit terjual akan berdampak langsung terhadap total pendapatan. Sebaliknya, fitur teknis seperti RAM dan ROM memiliki korelasi yang sangat rendah terhadap revenue, mengindikasikan bahwa faktor teknis bukanlah penentu utama dalam pembentukan pendapatan dari penjualan smartphone.



Model prediktif dibangun menggunakan algoritma Random Forest Regressor yang diimplementasikan dalam pipeline machine learning. Model dilatih dengan 80% data dan diuji dengan 20% data sisanya. Evaluasi terhadap kinerja model dilakukan menggunakan dua metrik utama, yaitu R^2 Score dan Mean Absolute Error (MAE). Hasil evaluasi menunjukkan bahwa model memiliki R^2 Score sebesar 0.99999, yang berarti bahwa hampir seluruh variasi nilai revenue dapat dijelaskan oleh model. Nilai MAE sebesar 177.75 juga menunjukkan bahwa rata-rata kesalahan prediksi sangat kecil dan berada dalam batas yang dapat diterima. Kedua hasil ini menunjukkan bahwa model yang dibangun sangat akurat dan andal untuk digunakan dalam konteks prediksi revenue.

Lebih lanjut, analisis terhadap feature importance dari model memperkuat temuan sebelumnya. Fitur Quantity Sold dan Price memiliki bobot penting tertinggi dalam memengaruhi output prediksi. Fitur lain seperti DaysInStock juga menunjukkan kontribusi yang cukup berarti, menandakan bahwa durasi produk berada di gudang

sebelum dijual juga dapat berdampak pada revenue. Sementara itu, fitur seperti RAM, ROM, Processor, dan Region memberikan kontribusi yang jauh lebih kecil, yang konsisten dengan hasil korelasi awal. Temuan ini menegaskan bahwa strategi harga dan distribusi produk memiliki dampak yang jauh lebih besar terhadap pendapatan dibandingkan dengan spesifikasi teknis produk.



Validasi model dilakukan dengan membandingkan hasil prediksi dengan nilai aktual revenue pada data uji. Grafik antara hasil prediksi dan nilai aktual menunjukkan hubungan yang hampir linear, dengan distribusi error yang sempit. Hal ini menunjukkan bahwa model mampu melakukan generalisasi dengan baik dan tidak mengalami overfitting. Selain itu, tidak ditemukan outlier ekstrem dalam distribusi error, yang semakin memperkuat keandalan model.

Dengan keseluruhan proses evaluasi ini, dapat disimpulkan bahwa pendekatan yang diterapkan berhasil membangun model prediktif yang sangat akurat dan menghasilkan insight yang relevan terhadap faktor-faktor yang mempengaruhi revenue penjualan smartphone.

BAB IV
KESIMPULAN

Berdasarkan analisis dan eksperimen yang telah dilakukan, dapat disimpulkan bahwa model prediksi pendapatan penjualan smartphone berhasil dibangun dengan performa yang sangat baik menggunakan algoritma Random Forest Regressor. Proses analisis dimulai dengan tahapan data preparation, eksplorasi data, hingga pembuatan dan evaluasi model prediktif.

Hasil eksplorasi data menunjukkan bahwa variabel yang paling mempengaruhi pendapatan adalah jumlah unit yang terjual (Quantity Sold) dan harga produk (Price). Visualisasi data dan hasil evaluasi model menunjukkan bahwa dua fitur tersebut memiliki korelasi tinggi terhadap revenue serta kontribusi terbesar dalam model prediksi. Sementara itu, fitur teknis seperti RAM, ROM, dan spesifikasi prosesor memiliki pengaruh yang jauh lebih kecil. Hal ini mengindikasikan bahwa strategi penjualan yang menekankan pada pengaturan harga dan volume penjualan akan lebih efektif dibandingkan hanya berfokus pada spesifikasi teknis produk.

Model yang dibangun menunjukkan nilai R^2 sebesar 0.99999 dan MAE sebesar 177.75, yang menandakan tingkat akurasi dan presisi yang sangat tinggi dalam memprediksi revenue. Validasi model terhadap data uji juga menunjukkan hasil yang konsisten dan tidak mengalami overfitting. Secara keseluruhan, proyek ini membuktikan bahwa pendekatan machine learning dapat digunakan untuk menggali insight penting dalam data penjualan dan membantu perusahaan dalam merancang strategi yang berbasis data. Model yang telah dikembangkan dapat digunakan untuk memprediksi pendapatan produk baru dan menjadi dasar pengambilan keputusan dalam penentuan harga, evaluasi performa penjualan, dan strategi distribusi.

Selain itu, model ini sangat berguna dalam situasi ketika pengguna telah mengetahui estimasi jumlah unit smartphone yang akan terjual (quantity sold) dan ingin memproyeksikan kemungkinan revenue yang akan diperoleh. Dengan memasukkan informasi seperti harga, spesifikasi teknis, dan atribut produk lainnya, serta jumlah penjualan yang diperkirakan, model mampu menghasilkan prediksi pendapatan yang

sangat akurat. Hal ini menjadikan model ini sebagai alat bantu yang efektif untuk melakukan simulasi strategi penjualan, menyusun target pendapatan, atau mengevaluasi dampak dari perubahan harga terhadap total revenue secara kuantitatif dan berbasis data.

REFERENSI

1. Statista. (2023). *Smartphone unit sales worldwide by vendor from 2016 to 2023*. Retrieved from: https://www.statista.com/statistics/271496/global-market-share-held-by-smartphone-vendors-since-4th-quarter-2009/
2. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. Retrieved from: https://link.springer.com/book/10.1007/978-1-4614-6849-3

LINK COLAB

BigData.ipynb