TEAM : INFINITY

Exploration and Analysis of Electric Vehicle Type Using Random Forest Classifier

Final Project DSA
COMPFEST 15

Meet Our Team

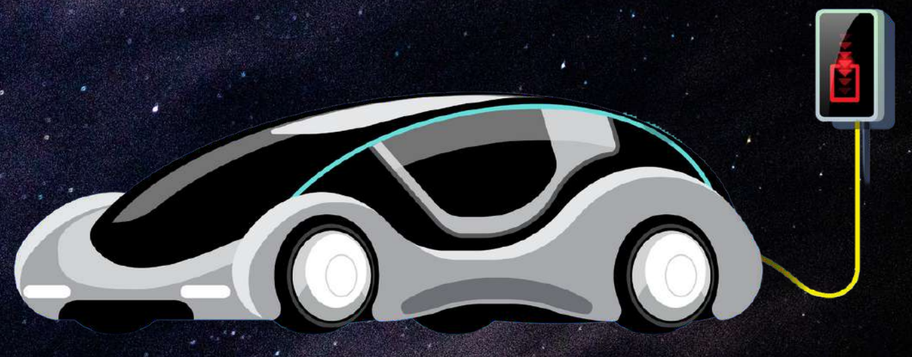
Khoirul Amar Sidik

Balqis Dwian Fitri Zamzami

Mujadid Choirus Surya

Outline

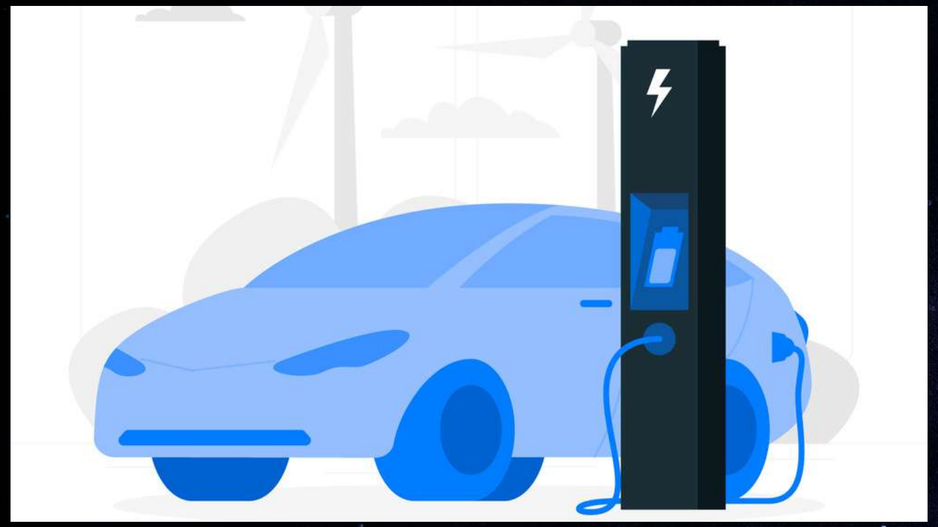
1. Introduction and Objective
2. About Data
3. Data Pre-processing & cleaning
4. Exploratory Data Analysis (EDA)
5. Feature Engineering
6. Modeling & Evaluation
7. Hyperparameter Tuning
8. Conclusions and Recommendations



Introduction

Dalam era mobilitas berkelanjutan, kendaraan listrik (Electric Vehicle/EV) telah menjadi solusi potensial untuk mengurangi dampak polusi udara dan emisi gas rumah kaca. Pemahaman dan wawasan terkait perkembangan electric vehicle menjadi menarik untuk dianalisis guna menentukan langkah-langkah menuju mobilitas yang lebih ramah lingkungan dan pembangunan infrastuktur yang berkelanjutan.

Objective



Proyek ini bertujuan untuk menganalisis data terkait electric vehicles untuk mendapatkan insight menarik dari data dan mengembangkan model prediktif dalam mengklasifikasikan Battery Electric Vehicles (BEV) dan Plug-in Hybrid Electric Vehicles (PHEV). Dengan demikian, proyek ini akan membantu konsumen dalam memilih kendaraan yang sesuai di daerahnya, mendukung pengembangan infrastruktur pengisian daya yang efisien, mengukur dampak lingkungan yang lebih tepat dan mendukung perancangan kebijakan publik yang lebih baik.

About Data

Data utama yang kami gunakan adalah **Electric Vehicle Population Data.** Data ini menunjukkan Battery Electric Vehicle (BEV) dan Plug-in Hybrid Electric Vehicle (PHEV) yang saat ini terdaftar melalui Departemen Perizinan (DOL) Negara Bagian Washington.

Kami juga menggunakan data : **Electric Vehicle Title and Registration Activity** sebagai data tambahan untuk melakukan eksplorasi.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 143596 entries, 0 to 143595
Data columns (total 17 columns):
 #   Column                                             Non-Null Count   Dtype
---  ------                                             --------------   -----
 0   VIN (1-10)                                         143596 non-null  object
 1   County                                             143574 non-null  object
 2   City                                               143574 non-null  object
 3   State                                              143596 non-null  object
 4   Postal Code                                        143574 non-null  float64
 5   Model Year                                         143596 non-null  int64
 6   Make                                               143596 non-null  object
 7   Model                                              143596 non-null  object
 8   Electric Vehicle Type                              143596 non-null  object
 9   Clean Alternative Fuel Vehicle (CAFV) Eligibility  143596 non-null  object
 10  Electric Range                                     143596 non-null  int64
 11  Base MSRP                                          143596 non-null  int64
 12  Legislative District                               143269 non-null  float64
 13  DOL Vehicle ID                                     143596 non-null  int64
 14  Vehicle Location                                   143571 non-null  object
 15  Electric Utility                                   143574 non-null  object
 16  2020 Census Tract                                  143574 non-null  float64
dtypes: float64(3), int64(4), object(10)
memory usage: 18.6+ MB
```

Data Pre-processing & cleaning

Terdapat missing data dalam dataset, kemudian kami lakukan drop untuk setiap missing data tersebut

Handling Missing Data

↓

Cek Duplicated Data

Tidak terdapat data duplikat dari dataset

↓

Rename Columns

Rename Columns bertujuan untuk mempermudah keterbacaan data

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) sangat berguna untuk mendapatkan pemahaman tentang data dan memperoleh pengetahuan berharga melalui visualisasi data. Dalam konteks proyek ini, EDA melibatkan analisis statistika deskriptif, analisis univariat, serta analisis korelasi.

Analisis Statistika Deskriptif

Statistika Deskriptif Data Numerik

	count	mean	std	min	25%	50%	75%	max
PostalCode	143266.0	9.825842e+04	3.020566e+02	9.800100e+04	9.805200e+04	9.812200e+04	9.837000e+04	9.940300e+04
ModelYear	143266.0	2.019867e+03	3.016107e+00	1.997000e+03	2.018000e+03	2.021000e+03	2.022000e+03	2.024000e+03
Electric_Range	143266.0	7.048542e+01	9.712199e+01	0.000000e+00	0.000000e+00	1.900000e+01	1.110000e+02	3.370000e+02
Base_MSRP	143266.0	1.372484e+03	9.445067e+03	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	8.450000e+05
Legislative_District	143266.0	2.937197e+01	1.482398e+01	1.000000e+00	1.800000e+01	3.300000e+01	4.300000e+01	4.900000e+01
DOL_Vehicle_ID	143266.0	2.092281e+08	8.353968e+07	4.385000e+03	1.668815e+08	2.111311e+08	2.364567e+08	4.792548e+08
2020_Census_Tract	143266.0	5.303966e+10	1.616041e+07	5.300195e+10	5.303301e+10	5.303303e+10	5.305307e+10	5.307794e+10
Longitude	143266.0	-1.220929e+02	1.005144e+00	-1.246252e+02	-1.223942e+02	-1.222918e+02	-1.221517e+02	-1.170444e+02
Latitude	143266.0	4.746846e+01	6.133369e-01	4.558386e+01	4.735799e+01	4.761385e+01	4.771558e+01	4.899634e+01

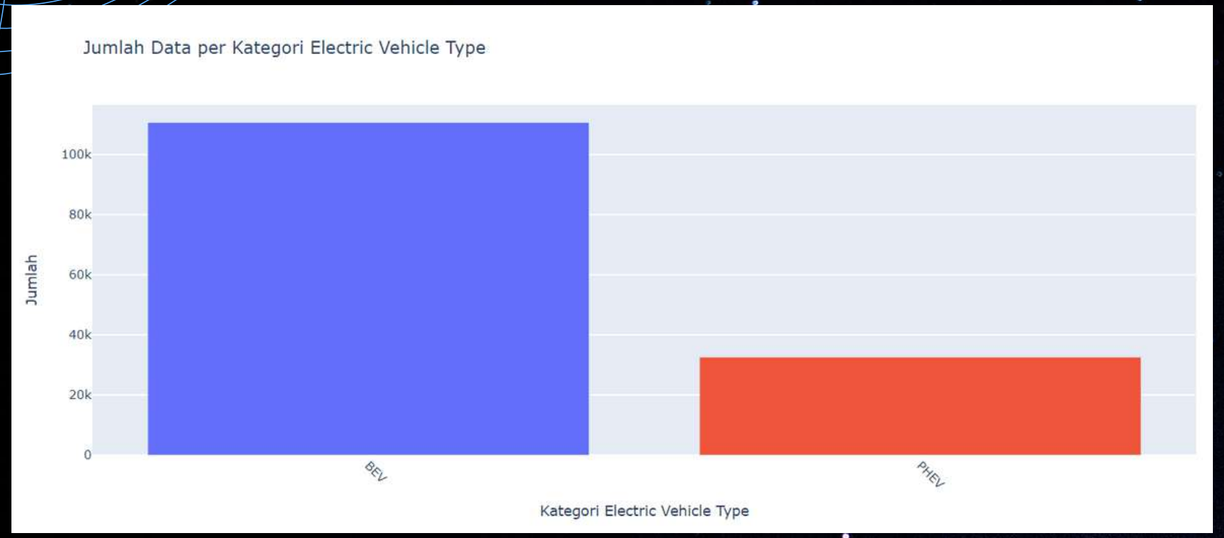
Statistika Deskriptif Data Kategorik

	count	unique	top	freq
VIN	143266	9303	7SAYGDEE7P	652
County	143266	39	King	75383
City	143266	456	Seattle	24662
State	143266	1	Washington	143266
Make	143266	37	TESLA	65396
Model	143266	127	MODEL 3	26684
Electric_Vehicle_Type	143266	2	BEV	110651
CAFV_Eligibility	143266	3	Eligibility unknown as battery range has not b...	63840
Electric_Utility_Category	143266	3	Multi type Utilities	84353

Kita bisa melihat nilai-nilai statistika deskriptif dari dataset seperti count, mean, std, min, max dan persentil untuk data numerik serta count, unique, top dan frekuensi untuk data kategorik. Dengan demikian kita dapat melihat persebaran data.

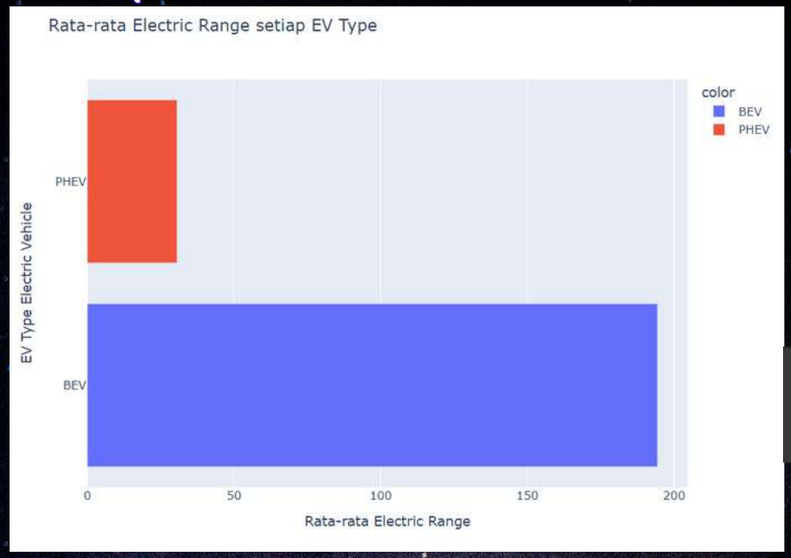
Analisis Univariat

Jumlah Data per Kategori Electric Vehicle Type



Grafik di atas menunjukkan sebaran variabel target, dimana 77,23% data merupakan data dengan kategori BEV. Ini menunjukan ada ketidak seimbangan dalam kolom EV Type (variabel target) . Hal ini perlu di perhatikan dalam membuat model nanti.

Rata-rata Electric Range berdasarkan EV Type

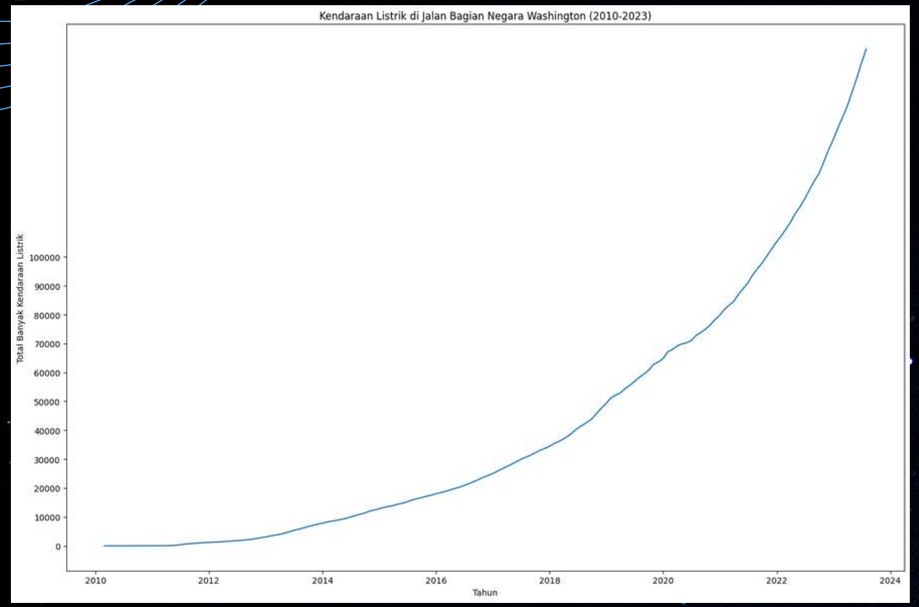


Electric_Vehicle_Type	
BEV	194.372626
PHEV	30.641944

Grafik diatas menunjukan sebaran nilai rata-rata dari EV range berdasarkan EV type. Diperoleh nilai rata-rata EV range untuk BEV adalah 194,37 dan PHEV 30,64.

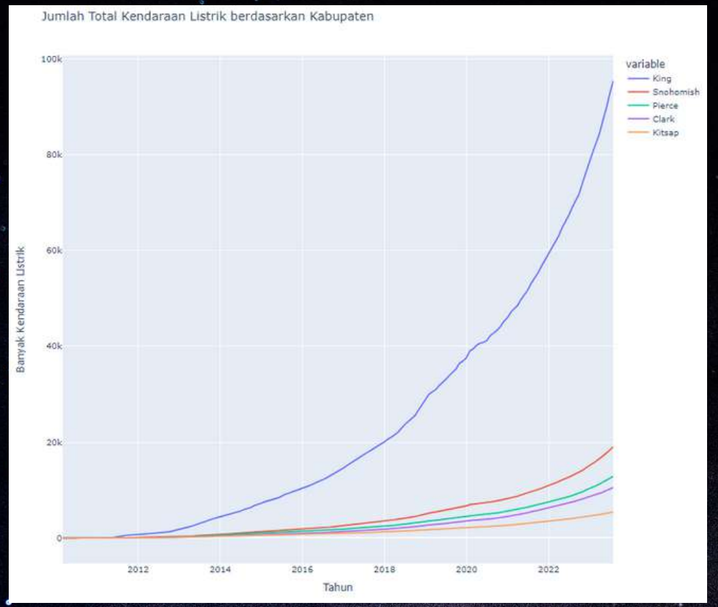
Analisis Univariat

Kendaraan Listrik di Jalan Bagian Negara Washington (2010-2023)



Grafik di atas menunjukkan pertumbuhan banyaknya kendaraan listrik yang melintas di jalan Wahington dalam periode tahun. Kita bisa melihat grafik yang eksponensial dari pertumbuhan EV.

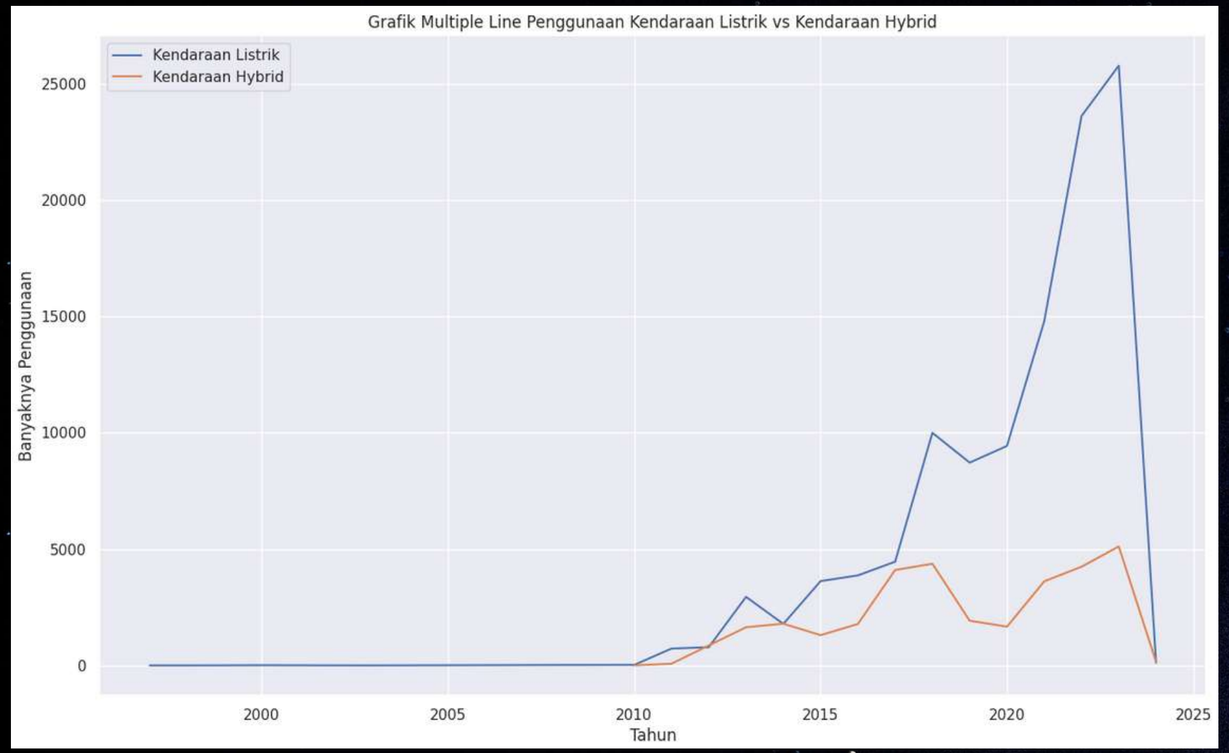
Jumlah Total Kendaraan Listrik berdasarkan Kabupaten



Grafik diatas menunjukan pertumbuhan kendaraan listrik berdasarkan wilayah top 5 kabupaten dengan jumlah kendaran listrik terbanyak. Diperoleh informasi pertumbuahan EV eksponensial serta king menjadi wilayah dengan pertumbuhan paling tinggi.

Analisis Univariat

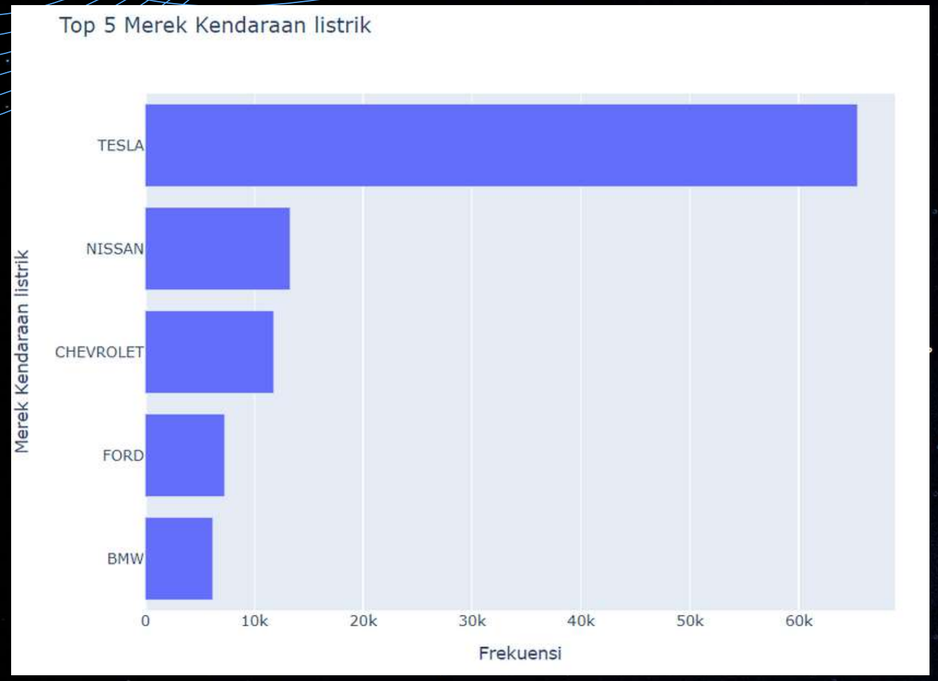
Pertumbuhan Penggunaan Kendaraan Listrik vs Kendaraan Hybrid Per Tahun



Grafik ini menunjukan pertumbuhan EV berdasarkan EV Type (BEV dan PHEV) dari tahun 2010 sampe 2023 terdapat peningkatan diantara keduannya namun data menunjukan EV type BEV lebih banyak dan lebih eksponensial dibandingakn PHEV.

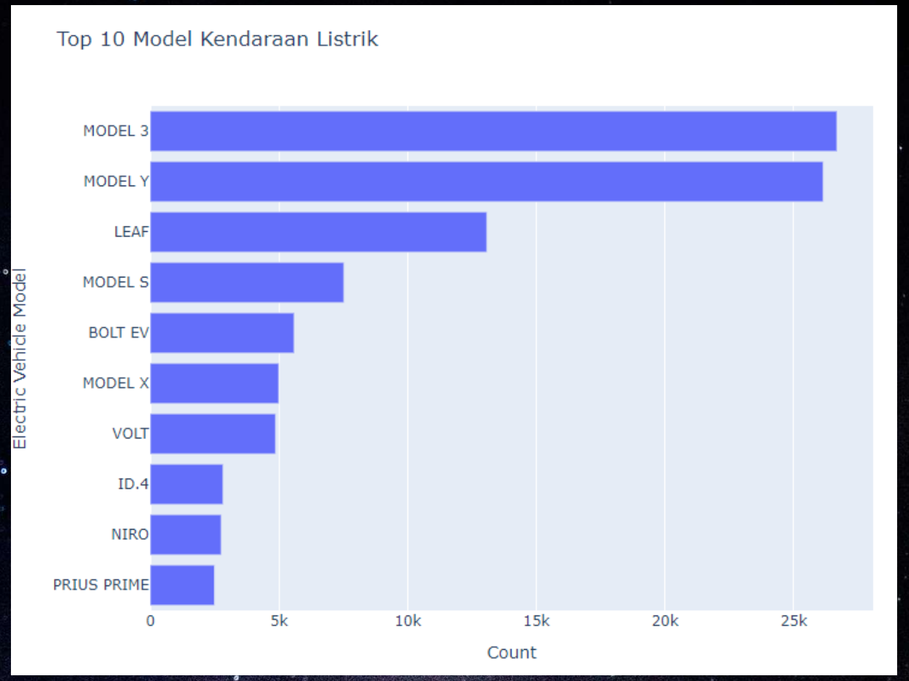
Analisis Univariat

Top 5 Kendaraan Listrik



Grafik di atas menunjukkan top 5 merek kendaraan listrik, dimana urutan pertama adalah Tesla diikuti Nissan, Chevrolet, Ford, dan BMW.

Top 10 Model Kendaraan Listrik



Grafik diatas menunjukkan Top 10 model EV yang paling sering ditemukan. Tiga besar diantaranya adalah Model 3, Model Y, dan LEAF.

Analisis Korelasi

Korelasi menunjukkan bahwa tidak banyak fitur yang berkorelasi kuat secara linier dengan target. Artinya sebagian besar korelasi yang terdapat dalam dataset bersifat non-linier. Berdasarkan hal tersebut, model yang tepat digunakan adalah model yang memiliki performa baik pada dataset yang memiliki banyak korelasi non-linier seperti Decision Trees, Random Forests dan lainnya. Pada Project ini kami menggunakan Random Forest.

Feature Engineering

Electric Utility Handling

Dari kolom Electric Utility kita mengubah menjadi 3 kategori yaitu Single Type Utilities, Multi type Utilities, dan Only One Option

Update coloumn 'Electric Vehicle Type'

Mengganti nama kategori dalam kolom 'Electric Vehicle Type' menjadi BEV dan PHEV

Menambah kolom latitude dan longitude berdasarkan kolom 'Vehicle Location'

Menambahkan kolom koordinat latitude dan longitude berdasarkan kolom 'Vehicle Location'

Feature Engineering

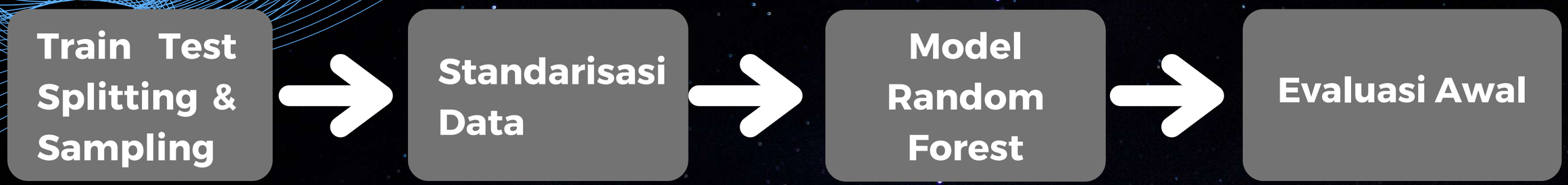
Handling State Code

Mengubah kolom state menjadi nama wilayah sebenarnya

Encoding

Mengubah data kategorik menjadi numerik menggunakan ordinal encoding dan frekuensi encoding.

Modeling


Train Test Splitting & Sampling
Standarisasi Data
Model Random Forest
Evaluasi Awal


Hyperparameter Tuning
Model Terbaik
Evaluasi Akhir

Train Test Splitting & Sampling

```
# Split Data
X = train.drop('Electric_Vehicle_Type', axis=1).values
y = train['Electric_Vehicle_Type'].values

# membagi data menjadi set data latih dan uji dengan ukuran uji 30%.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, stratify=y, random_state=42)
```

```
0    110651
1     32615
Name: Electric_Vehicle_Type, dtype: int64
```

Seperti yang kita ketahui saat eksplorasi kelas target tidak seimbang, hal ini perlu dihindari agar model tidak menghasilkan bias. Untuk mengatasi hal ini, diperlukan pengambilan sampel sehingga jumlah kelas target mendekati kondisi seimbang.

```
Original dataset shape: Counter({0: 77456, 1: 22830})
Resample dataset shape: Counter({0: 22830, 1: 22830})
```

Modeling

```
# Lakukan standarisasi data

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_resampled)
X_test_scaled = scaler.transform(X_test)
```

```
[ ]  # Tentukan model

     model = RandomForestClassifier(n_estimators=100)
     model.fit(X_train_resampled, y_train_resampled)
```

▾ RandomForestClassifier

RandomForestClassifier()

```
[ ]  y_pred = model.predict(X_test)
```

Modeling

Model Evaluation

Confusion Matrix:

```
[[33149    46]
 [    1  9784]]
```

Evaluation Metrics:

	Metrics	Score
0	Accuracy	0.998906
1	Precision	0.995320
2	Recall	0.999898
3	F1 Score	0.997604

Akurasi = (Jumlah Prediksi Benar) / (Jumlah Prediksi Benar + Jumlah Prediksi Salah)

Presisi = (True Positives) / (True Positives + False Positives)

Recall = (True Positives) / (True Positives + False Negatives)

F1-Score = 2 * (Presisi * Recall) / (Presisi + Recall)

Modeling

Hyperparameter Tuning

```
[ ]  # Hyperparameter tuning menggunakan GridSearchCV

     params = {
         'n_estimators': [100, 150],
         'max_depth': [None, 5, 10],
         'bootstrap': [True, False],
     }
     grid_search = GridSearchCV(model, params, cv=5, scoring='accuracy')
     grid_search.fit(X_train_scaled, y_train_resampled)
```


```
[ ]  # Hasil terbaik dari GridSearchCV

     best_params = grid_search.best_params_
     best_model = grid_search.best_estimator_
```

```
[ ]  # Evaluasi model

     y_pred_hyper = best_model.predict(X_test_scaled)
```

Modeling

Best Model Evaluation

Evaluasi model setelah dilakukan proses tuning

```
Best Parameters:  {'bootstrap': False, 'max_depth': None, 'n_estimators': 100}
Confusion Matrix:
 [[33157    38]
 [    0  9785]]

Evaluation Metrics:
```

	Metrics	Score
0	Accuracy	0.999116
1	Precision	0.996132
2	Recall	1.000000
3	F1 Score	0.998062

Signifikan Fitur

	Feature	Importance
0	Electric_Range	0.419801
1	Model	0.184926
2	CAFV_Eligibility	0.180007
3	Make	0.079735
4	VIN	0.047785

Modeling

Konklusi

- Model Random Forest Terbaik mempunyai akurasi 0.99, yang berarti model mempunyai akurasi yang tinggi dalam memprediksi kelas.
- Kendaraan listrik dengan type BEV memiliki presentase lebih besar dari PHEV dalam data, yaitu sebesar 77,23%.
- Kendaraan listrik jenis BEV mempunyai rata-rata electric range lebih besar dibandingan jenis PHEV. Diperoleh nilai rata-rata EV range untuk BEV adalah 194,37 dan PHEV 30,64.
- Pertumbuhan kendaraan listrik cukup eksponensial dari tahun ke tahun. Pertumbuhan paling signifikan ada di wilayah King.

Rekomendasi

- Pemahaman yang baik tentang perbedaan antara BEV dan PHEV membantu konsumen memilih kendaraan yang sesuai dengan kebutuhan mereka, termasuk pertimbangan terkait lingkungan, jangkauan perjalanan, dan efisiensi bahan bakar. Rekomendasi ini dapat memandu konsumen untuk membuat keputusan yang lebih bijak dalam pembelian kendaraan listrik.

- Pengembangan Infrastruktur yang Efisien. Mengetahui jenis kendaraan yang paling umum digunakan di suatu wilayah membantu dalam merencanakan dan mengembangkan infrastruktur pengisian daya yang sesuai. Ini memastikan bahwa sumber daya diarahkan secara efisien ke pengisian daya yang diperlukan oleh mayoritas pemilik kendaraan listrik di wilayah tersebut.

- Produsen mobil memerlukan pemahaman yang baik tentang tren dalam penjualan BEV dan PHEV untuk merencanakan produksi dan inovasi produk yang lebih efisien.

- PHEV memiliki emisi lebih tinggi ketika menggunakan mesin pembakaran internalnya. Oleh karena itu, pemahaman tentang seberapa banyak BEV dan PHEV yang beroperasi di jalan-jalan memungkinkan pengukuran dampak lingkungan yang lebih akurat dari penggunaan kendaraan listrik.

Thank You