CO2 Emission Prediction for Vehicles

Sultan Mehedi Masud(22101071), Susmita Biswas(22101380)

CSE422 - Artificial Intelligence

BRAC University,

Dhaka,

Bangladesh

January 03, 2025

Table of Contents

CO2 Emission Prediction for Vehicles

1. INTRODUCTION:

Our project aims to provide insight into the CO_2 emissions caused by various vehicles. This project utilizes a comprehensive dataset containing vehicle specifications and fuel consumption metrics to develop predictive models. We implemented and compared four different regression models: linear regression, Random Forest, K-nearest neighbors, and Decision Tree.

We took inspiration from the recent events happening in Dhaka regarding pollution and poor air quality. Dhaka is one of the most populated countries with a huge amount of traffic. So we wanted to explore how vehicles can negatively impact the environment by increasing CO_2 emissions.

2. Dataset Description:

- **Source**: Kaggle
  - **Link**: https://www.kaggle.com/datasets/tanishqdublish/vehcile-fuel-consumption
  - Reference: Kaggle, https://www.kaggle.com/.

- **Dataset Description**: This dataset provides an overview of fuel consumption in various types of vehicles as well as gives an insight into the CO_2 emission based on various technological features

- Initial Feature number: 81
- Our project is based on the feature **tailpipe_co2_ft1**, a continuous numerical value indicating it as a regression problem. The second reason is that the targeted (tailpipe_co2_ft1) feature is a numerical feature, not a categorical which means a classification problem is not possible.
- Number of Data Points (Rows): 38,113
- The dataset has both quantitative and qualitative features

```
print(f"dataset data type: ")

df.dtypes
```

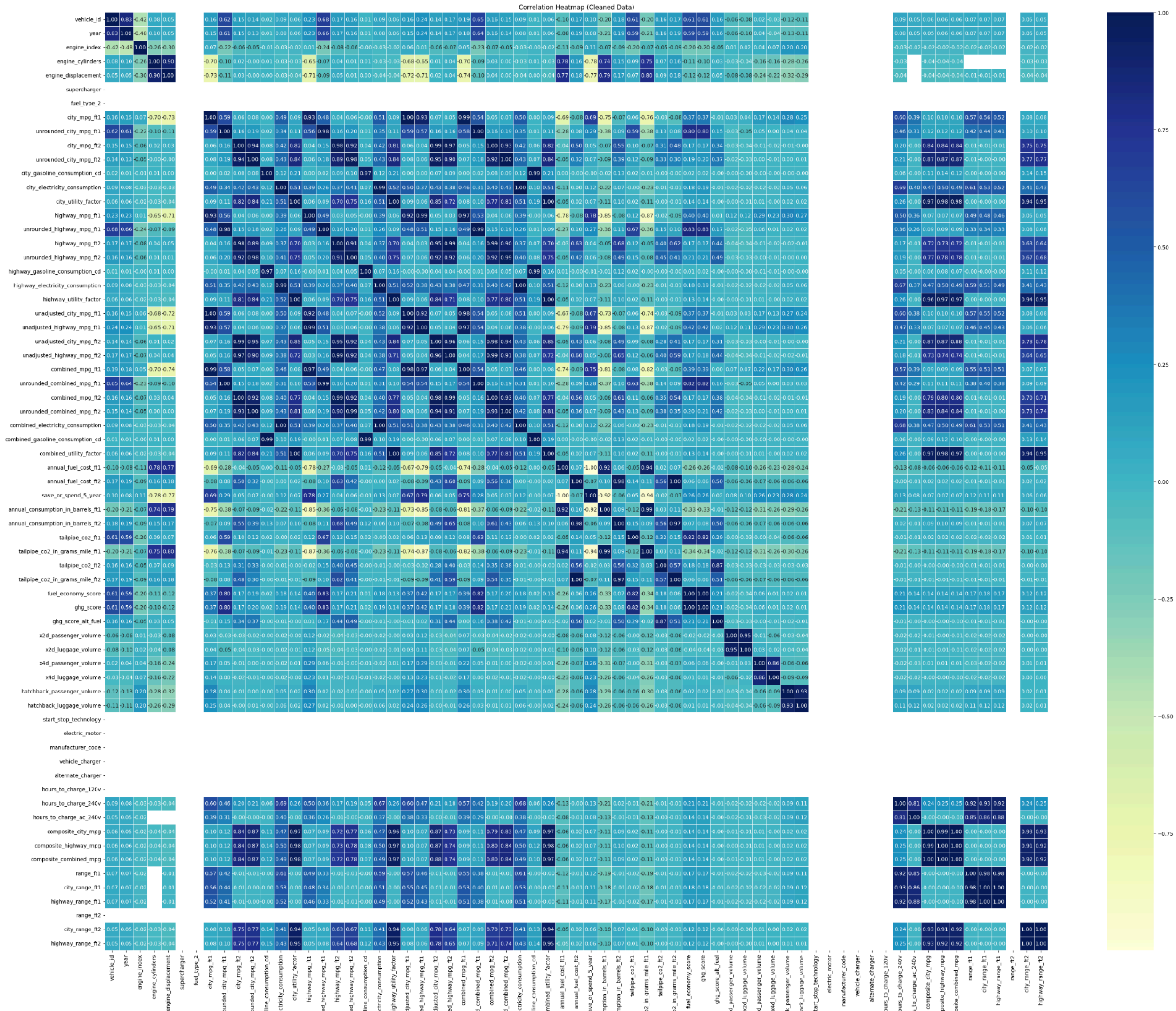
dataset data type:

	0
vehicle_id	int64
year	int64
make	object
model	object
class	object
...	...
city_range_ft1	float64
highway_range_ft1	float64
range_ft2	float64
city_range_ft2	float64
highway_range_ft2	float64

81 rows × 1 columns

dtype: object

Heatmap:



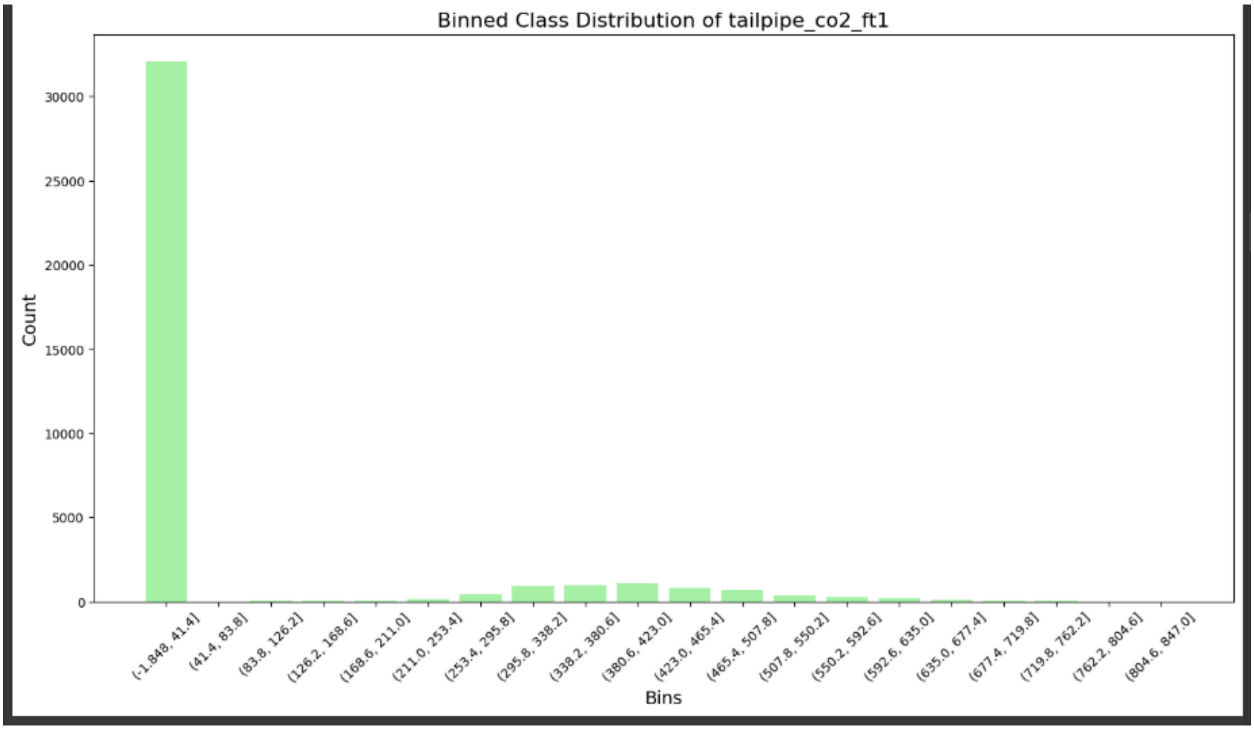
Imbalance dataset

For the output feature tailpipe_co2_ft1 we can see the data set is imbalance

With a large number of unique values and -1 dominating the data.

```
Unique Labels in the dependent column:  [ -1   0  84  87 169 133 322 361 228 500 387 401 385 421 442 349 394 373
 482 585 550 579 381 376 342 374 379 343 320 414 418 424 460 565 396 389
 395 429 337 499 480 486 413 439 440 475 345 432 646 633 488 580 517 604
 688 427 407 433 404 324 326 332 348 341 340 446 461 325 415 420 375 458
 327 371 454 510 476 473 459 402 403 584 317 346 350 562 563 552 544 516
 372 416 634 465 478 547 847 494 351 453 312 409 363 344 408 366 426 641
 411 554 462 428 431 570 581 612 430 448 524 520 399 436 487 463 533 496
 457 300 336 298 309 295 273 481 347 527 598 617 740 670 705 717 410 310
 423 531 618 573 625 290 305 276 321 301 284 265 714 724  81 378 383 477
 425 503 495 523 292 328 285 280 279 558 479 445 419 609 567 674 664 692
 679 293 129 229 642 630 666 645 755 782 687 370 356 357 364 606 450 471
 393 599 519 470 663 505 557 675 275 272 437 294 304 278 358 318 306 215
 386 438 490 392 532 564 623 307 365 353 288 354 200 218 244 266 384 406
 216 447 297 493 417 443 242 245 405 397 441 474 435 559 468 508 515 489
 368 360 391 464 589 616 504 502 296 314 247 291 283 388 333 334 380 676
 742 626 592 561 224 412 514 369 485 619 444 422 390 572 624 286 313 377
 456 367 277 311 359 398 400 576 574 498 506 614 469 680 696 569 587 615
 451 541 611 492 621 610 537 601 605 622 329 268 267 597 602 639 262 530
 352 513 549 578 525 637 238 607 299 316 330 355 467 220 534 593 186 179
 338 596 528 560 271 335 331 339 362 472 302 497 538 620 466  40 198 603
 556  91 571 526 483 568 672 711 449 536 270 264 521 230 792 511 659 522
 252 484 188 130 289 259 202 319 452 282 315 308 566 223 577 546 281 501
 382 507 243 323 274 636 632 255 671 455 206 638 796 263 178 217 303 434
 535 553 555 551 733 715 673 712 491 260 709 261 250 545 539 219 227 287
 588 529 185 613 600 158 138 652 184  37 518 104 196  51 654 691 542 214
 548 225 221 101 727 212 543 257 762 575 249 258 246 170 591 241 189  29
 199 248 512 540 269 704 716 194 106 662 697 661 683 122 222 254 112 210
 183 256 251 163 154 177 207  97 746 698 660 829 226 237 627 595 171 193
  78 594]
The number of occurences of the unique labels:
 tailpipe_co2_ft1
-1      31953
 0        133
 415       51
 347       51
 305       47
          ...
 241        1
 29         1
 199        1
 602        1
 594        1
Name: count, Length: 506, dtype: int64
```

Bar Chart Representation:



3. Dataset pre_processing:

- Faults:

Null Values:

```python
print("Datset Null checking: ")
df.isnull().sum()
```

Datset Null checking:

	0
vehicle_id	0
year	0
make	0
model	0
class	0
...	...
city_range_ft1	0
highway_range_ft1	0
range_ft2	38113
city_range_ft2	0
highway_range_ft2	0

81 rows × 1 columns

dtype: int64

Categorical Values:

```
1 df.select_dtypes(include=['object'])
```

	make	model	class	drive	transmission	transmission_type	engine_descriptor	turbocharger	fuel_type	fuel_type_1	gas_guzzler_tax	my_mpg_data	alternative_fuel_technology
0	Alfa Romeo	GT V6 2.5	Minicompact Cars	NaN	Manual 5-Speed	NaN	(FFS)	NaN	Regular	Regular Gasoline	NaN	N	NaN
1	Alfa Romeo	GT V6 2.5	Minicompact Cars	NaN	Manual 5-Speed	NaN	(FFS) CA model	NaN	Regular	Regular Gasoline	NaN	N	NaN
2	Alfa Romeo	Spider Veloce 2000	Two Seaters	NaN	Manual 5-Speed	NaN	(FFS)	NaN	Regular	Regular Gasoline	NaN	N	NaN
3	Alfa Romeo	Spider Veloce 2000	Two Seaters	NaN	Manual 5-Speed	NaN	(FFS) CA model	NaN	Regular	Regular Gasoline	NaN	N	NaN
4	AM General	DJ Po Vehicle 2WD	Special Purpose Vehicle 2WD	2-Wheel Drive	Automatic 3-Speed	NaN	(FFS)	NaN	Regular	Regular Gasoline	NaN	N	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
38108	Volvo	XC60 FWD	Small Sport Utility Vehicle 2WD	Front-Wheel Drive	Automatic (S8)	NaN	SIDI	True	Regular	Regular Gasoline	NaN	N	NaN
38109	Volvo	XC90 AWD	Standard Sport Utility Vehicle 4WD	All-Wheel Drive	Automatic (S8)	NaN	SIDI	True	Premium	Premium Gasoline	NaN	N	NaN
38110	Volvo	XC90 AWD	Standard Sport Utility Vehicle 4WD	All-Wheel Drive	Automatic (S8)	NaN	SIDI	True	Premium	Premium Gasoline	NaN	N	NaN
38111	Volvo	XC90 AWD PHEV	Standard Sport Utility Vehicle 4WD	All-Wheel Drive	Automatic (S8)	NaN	SIDI; PHEV	True	Premium and Electricity	Premium Gasoline	NaN	N	Plug-in Hybrid
38112	Volvo	XC90 FWD	Standard Sport Utility Vehicle 2WD	Front-Wheel Drive	Automatic (S8)	NaN	SIDI	True	Premium	Premium Gasoline	NaN	N	NaN

38113 rows × 13 columns

Solution:

Problem: Firstly by looking at the data set we can see we have a huge number of columns that have 0 or Nan value. To solve it we followed the drop rows and columns approach.

Delete Rows and Columns

- We dropped columns that have 50% or more than 50% 0 or Nan Values
- Also cleaned all the rows having all 0, we did not approach for more than 50% 0 because then we lost a significant amount of data. Also, 0 is a value we will need
- We cleared rows with Nan as well

Problem 2:

We had many columns that had similar data or similar type of data which made the prediction misleading to solve it we,

- we used correlation to find redundant columns and dropped them. The more the correlation, the greater the number of similar or redundant values. So we dropped them

  Encoding:

As we had categorical value, to make sure all data were aligned and numerically usable we did OneHotEncoding. This gave us the numerical representation of non numerical data.

4. Feature Scaling:

Features can have different ranges, like one feature can range up to 1-100 and another 1-2000. This can introduce bias for models like KNN where the model is dependent on distance calculation. To avoid that, we use standard scaling which balances the dataset by bringing the features to a common scale.

5. Dataset Splitting:

We did a stratified splitting where we used 70% to train and 30% to test.

6. Model Training and Testing:

We decided to run 4 models.

Linear Regression: As This is a regression or prediction problem, we decided to run Linear regression model.

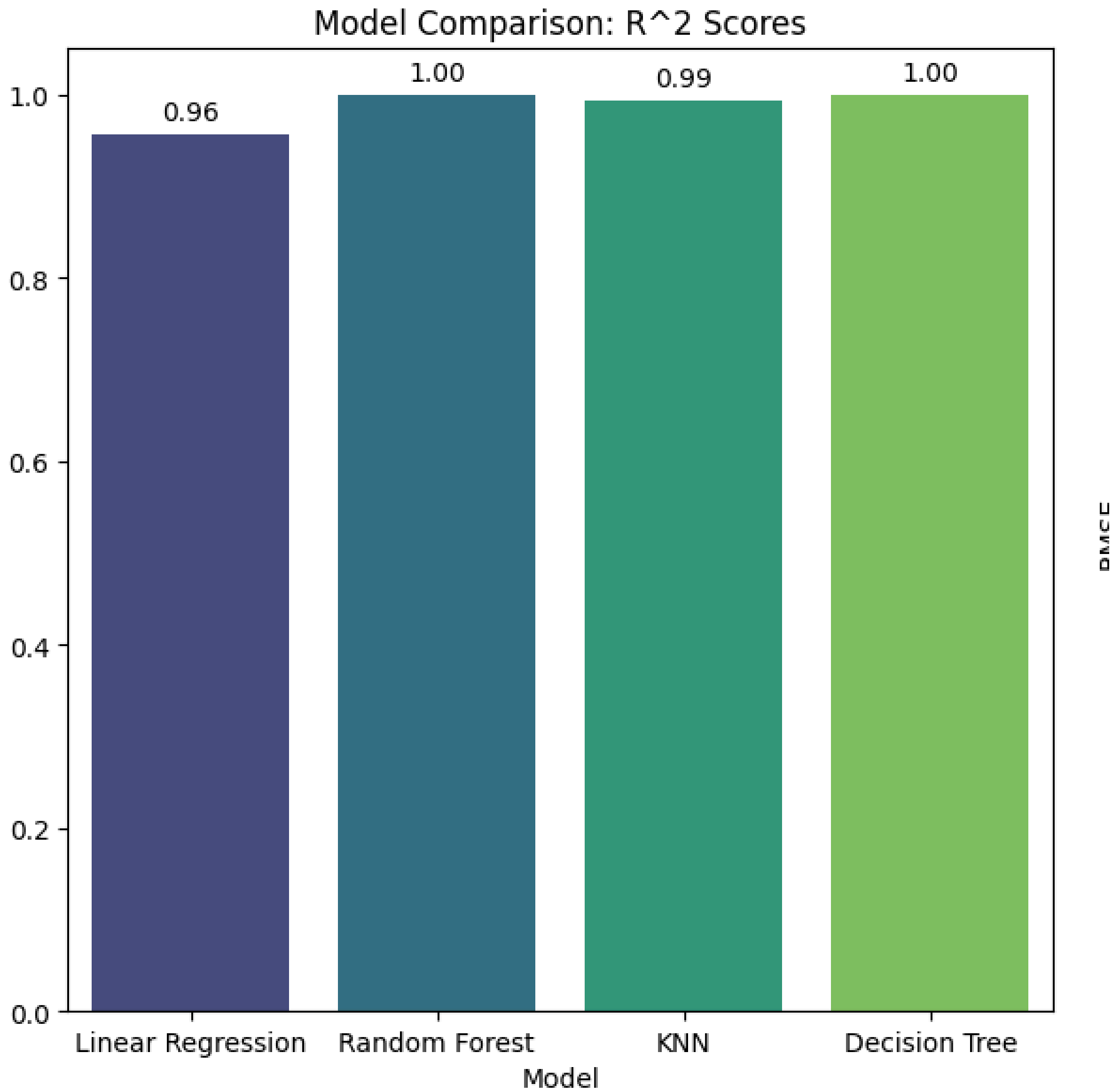
Random Forest Regressor Model:

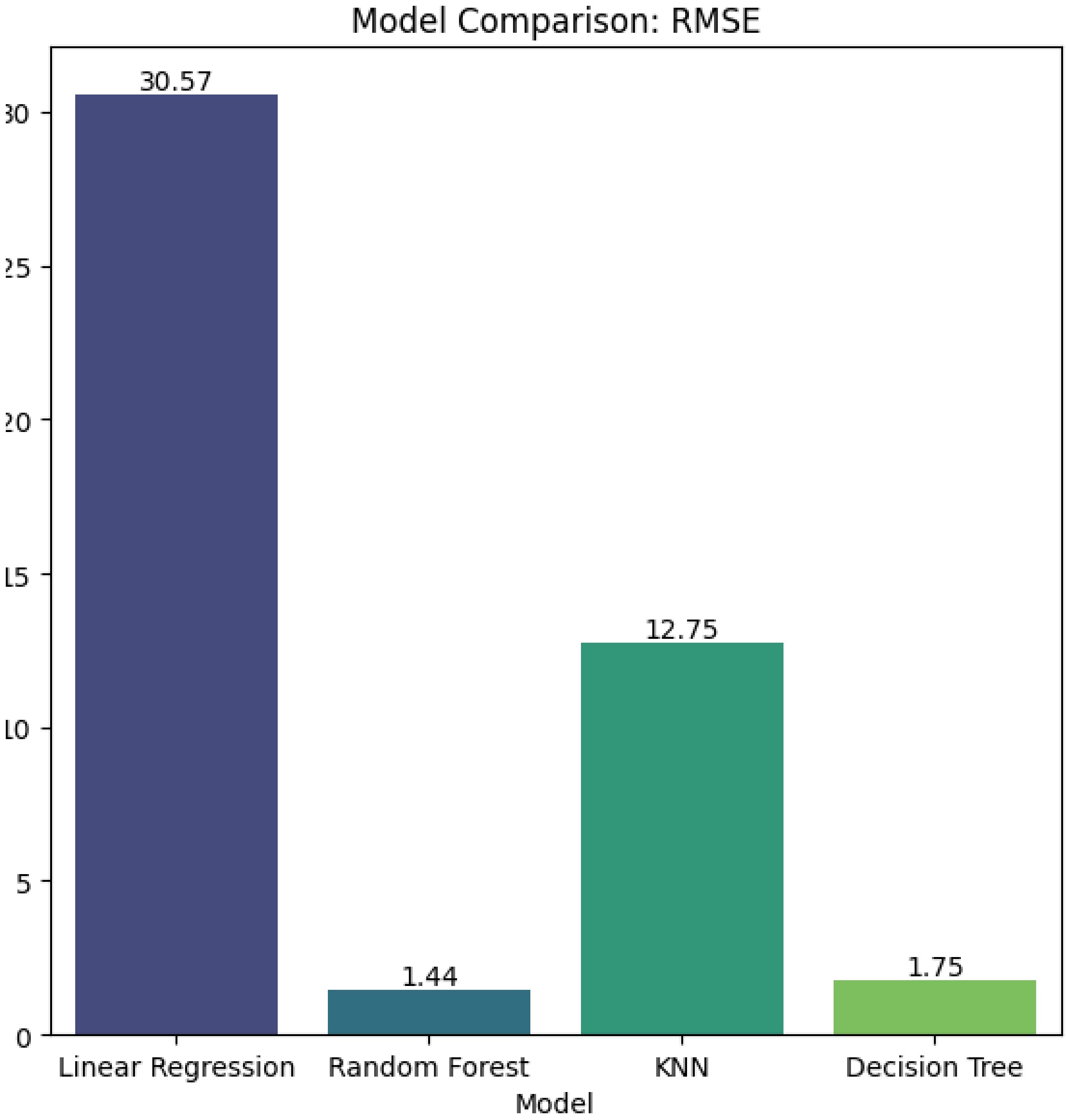
KNN regressor: Chosen for its simplicity and effective performance in predicting continuous values

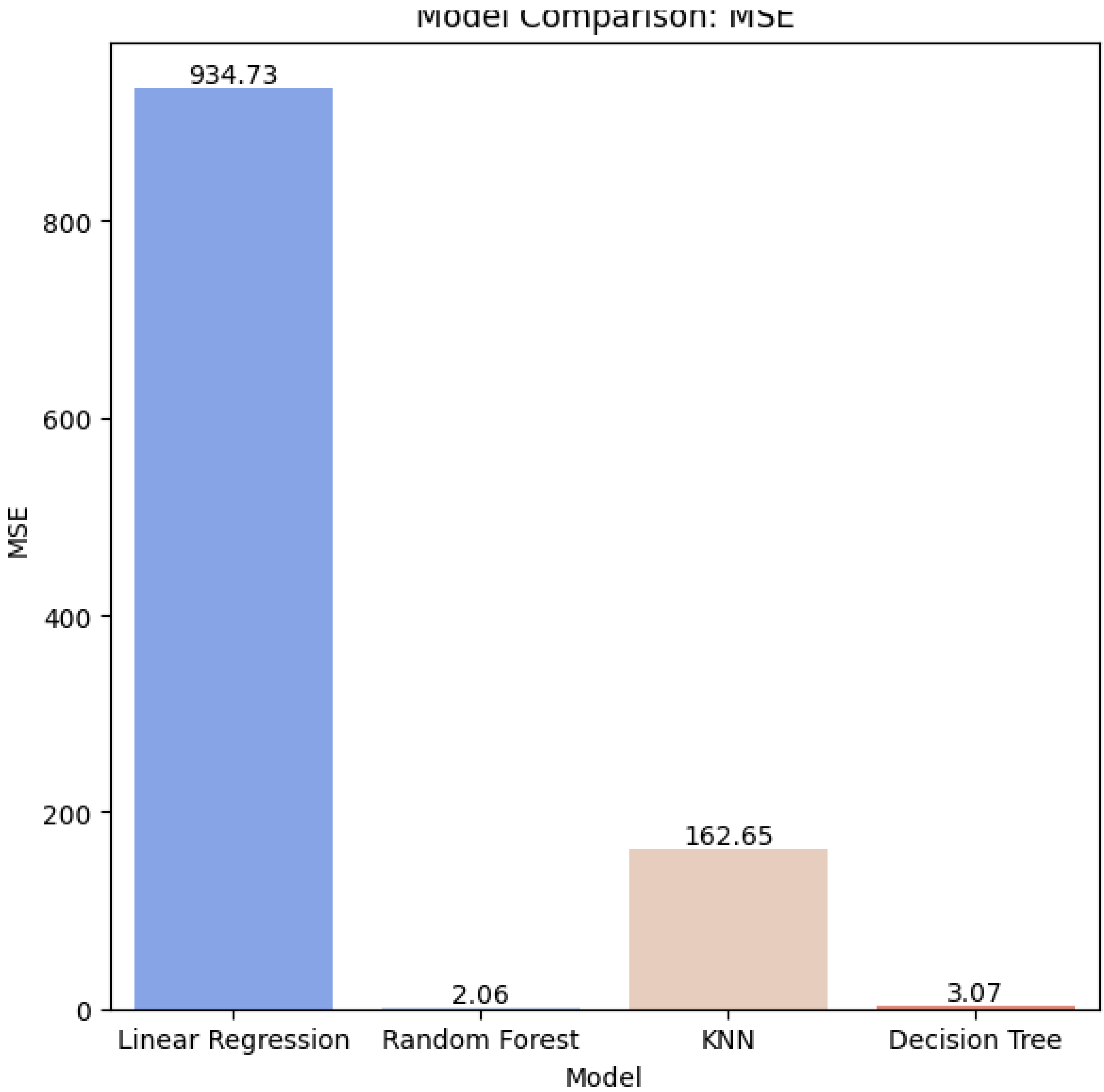
Decision Tree Regressor: Used for its flexibility in capturing non-linear relationships and ability to handle complex data interactions without needing feature scaling.

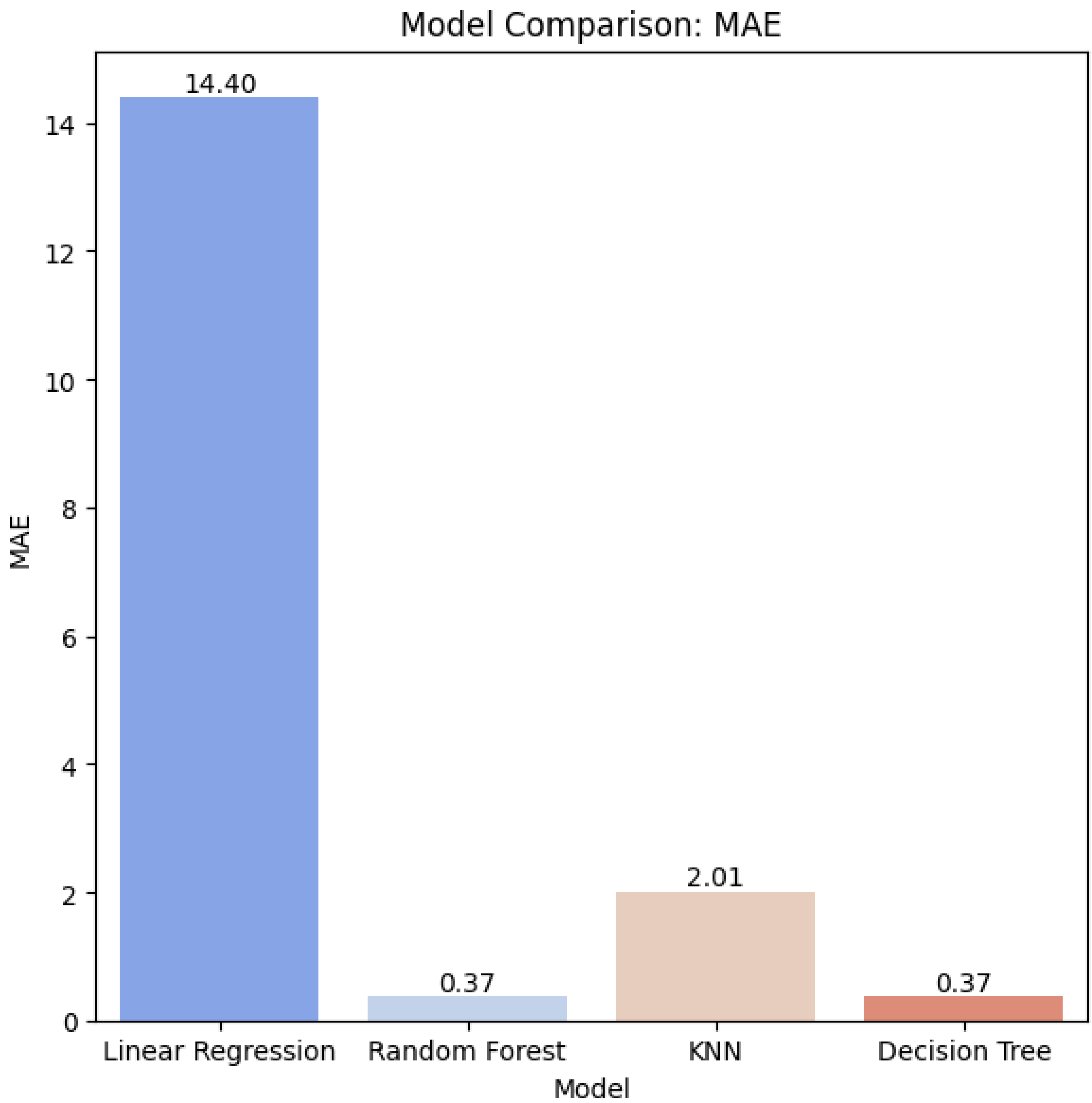
Random Forest Regressor: chosen for its ability to increase prediction accuracy and robustness

7. Comparison Analysis:




Model Comparison: RMSE
30.57
1.44
12.75
1.75
0
5
Linear Regression
Random Forest
KNN
Decision Tree
Model


Model Comparison: MSE
934.73
162.65
2.06
3.07
800
600
400
200
0
MSE
Linear Regression
Random Forest
KNN
Decision Tree
Model



This analysis revealed that the Random Forest model performed best with an R^2 score of [0.999], indicating strong predictive capability. It also has the lowest error values for MSE, RMSE, and MAE indicating it's high performance. The model comparison showed that models like Random Forest outperformed simpler models like Linear Regression and KNN in predicting CO2 emissions. This indicates a nonlinear and complex relationship between vehicle characteristics and CO2 emissions. These findings can help inform vehicle design decisions and emissions regulations

8. F1 scores

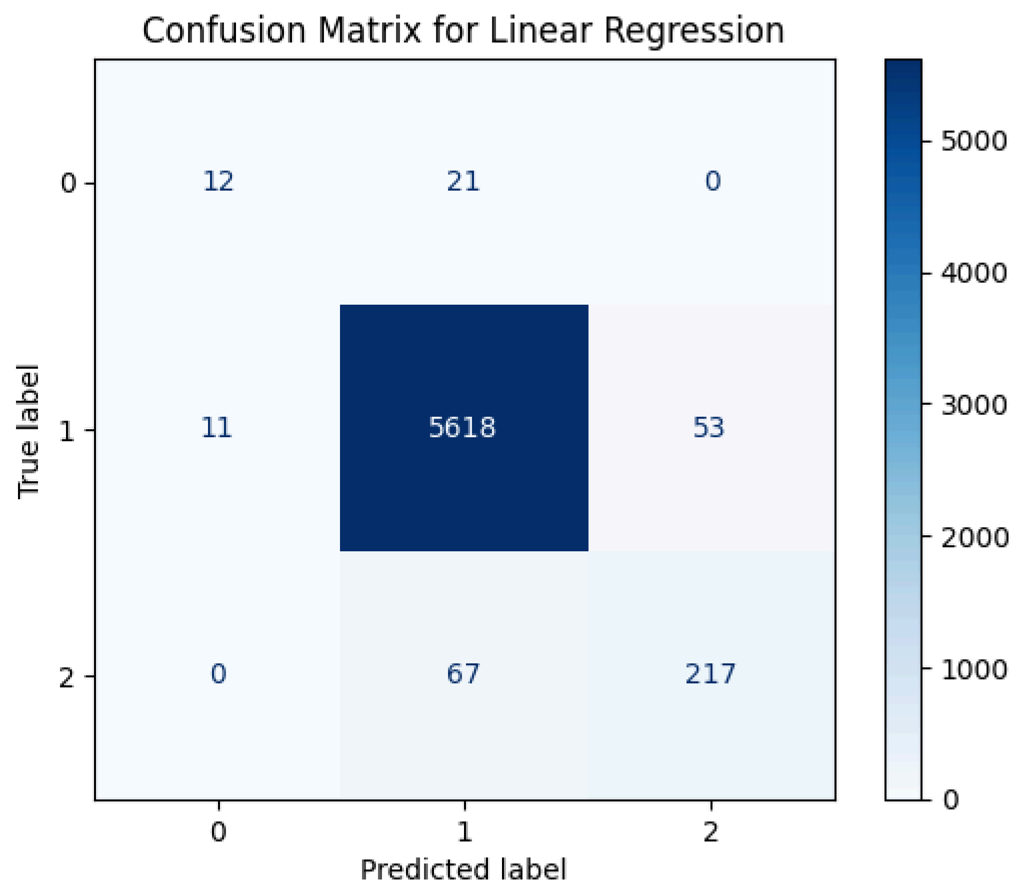
```
Linear Regression:
  Precision: 0.77
  Recall: 0.71
  F1 Score: 0.73

Random Forest:
  Precision: 0.99
  Recall: 0.94
  F1 Score: 0.96

KNN:
  Precision: 0.94
  Recall: 0.80
  F1 Score: 0.85

Decision Tree:
  Precision: 0.97
  Recall: 0.94
  F1 Score: 0.96
```

Random Forest and Decision Tree are the best-performing models with high precision, recall, and F1 Score.
KNN shows good precision but relatively lower recall, indicating it might miss some positives.
Linear Regression lags due to its linear assumptions, making it less suitable for this dataset.

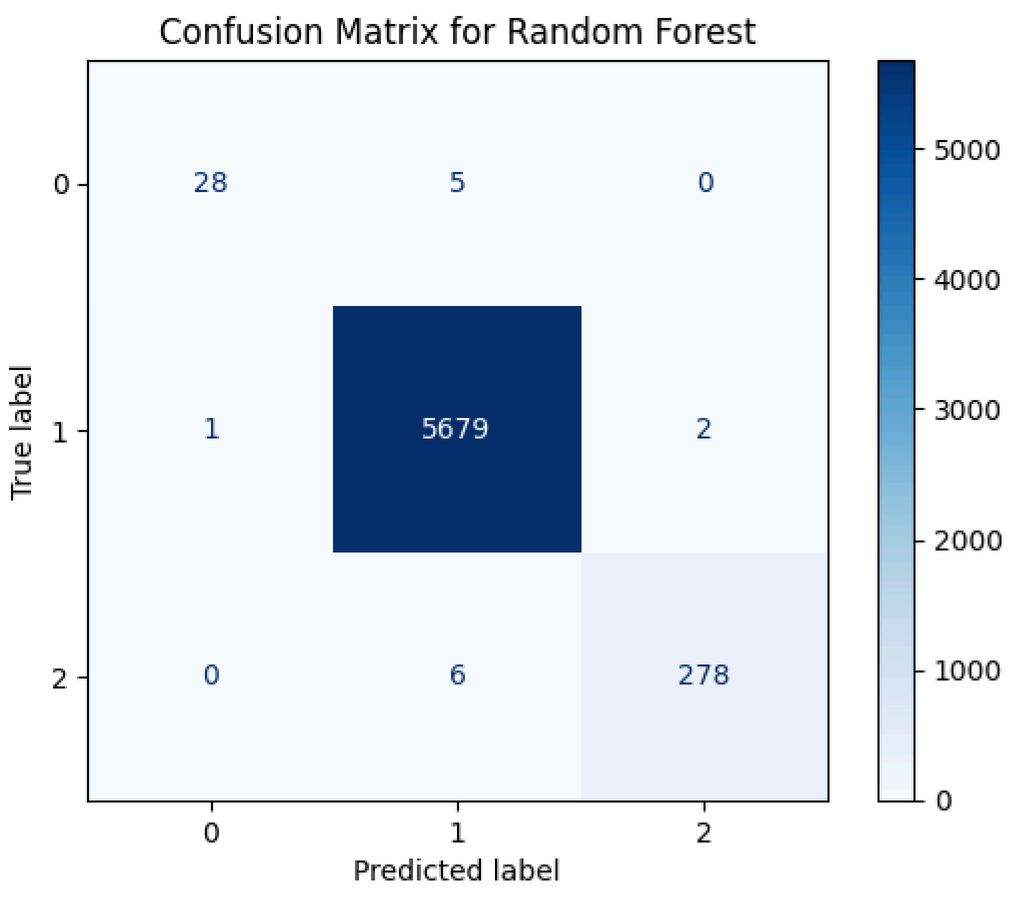
9. Confusion Matrix

```
Confusion Matrix for Linear Regression:
Actual \ Predicted | Class 0 | Class 1 | Class 2
-------------------------------------------
Class 0            | 12      | 21      | 0
Class 1            | 11      | 5618    | 53
Class 2            | 0       | 67      | 217


Confusion Matrix for Random Forest:
Actual \ Predicted | Class 0 | Class 1 | Class 2
-------------------------------------------
Class 0            | 28      | 5       | 0
Class 1            | 1       | 5679    | 2
Class 2            | 0       | 6       | 278


Confusion Matrix for KNN:
Actual \ Predicted | Class 0 | Class 1 | Class 2
-------------------------------------------
Class 0            | 16      | 17      | 0
Class 1            | 2       | 5661    | 19
Class 2            | 0       | 20      | 264


Confusion Matrix for Decision Tree:
Actual \ Predicted | Class 0 | Class 1 | Class 2
-------------------------------------------
Class 0            | 28      | 5       | 0
Class 1            | 2       | 5678    | 2
Class 2            | 0       | 7       | 277
```

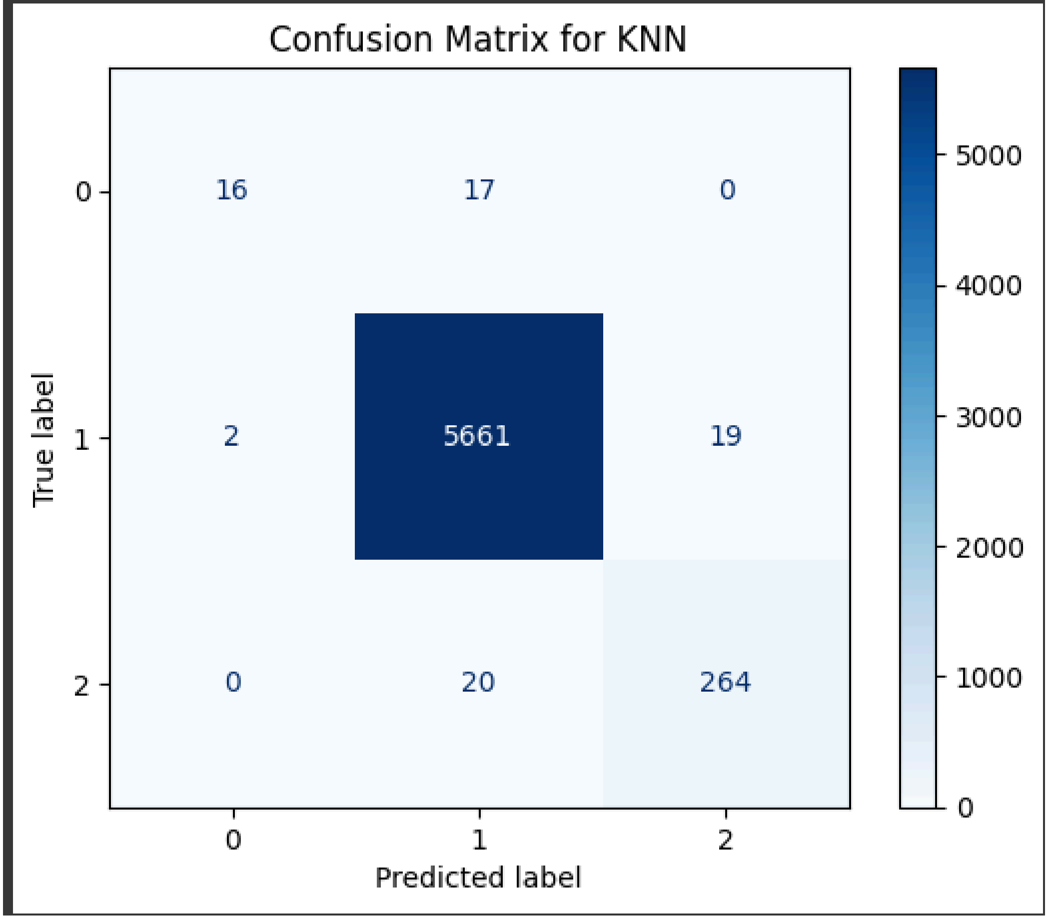
```
Confusion Matrix for Linear Regression:
```

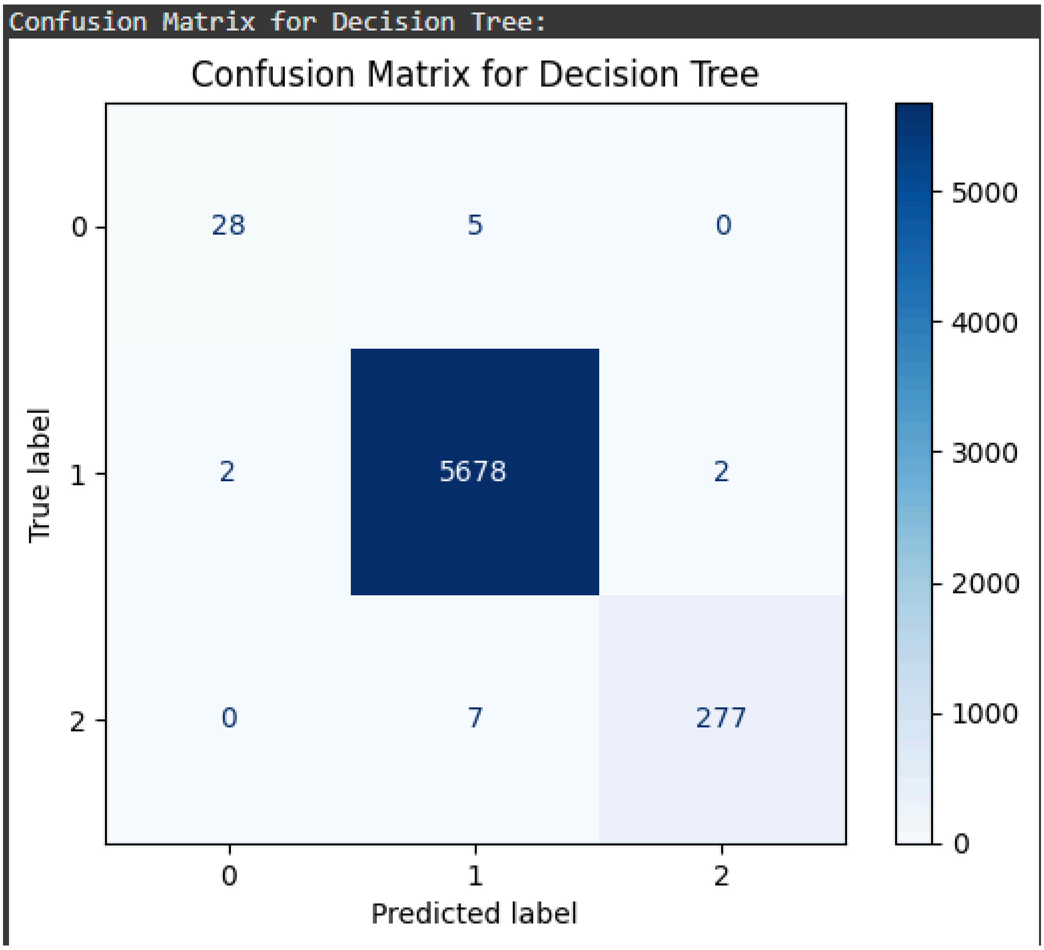


Confusion Matrix for Random Forest:



Confusion Matrix for KNN:





10. Conclusion:

We aimed to explore the data that gave us an understanding of how carbon emissions and vehicles are connected. This project helped us understand those factors and gave us a new scope for future work.