



Máster en Big Data y Data Science

Asignatura: Prácticas en empresa

Memoria final de prácticas

Alumno: Gómez Ramírez, Daniel

Dirección: Valencia

Tutor prácticas: Rosell Tornel, Ricardo

Edición 2021-2022

Periodo de prácticas 27-09-2021 a 09-11-2021

Índice

1.	Introducción	7
2.	Objetivos.....	8
3.	Desarrollo o metodología.....	9
3.1.	Descripción del dataset.....	9
3.2.	Variables del dataset	9
3.3.	Eliminación de factores no posibles/outliers	10
3.3.1.	Agrupación de las operaciones	11
3.3.2.	Observaciones por subárea:.....	17
3.3.1.	Creación de la variable objetivo (target).....	18
3.3.2.	Codificación de características y normalización.....	22
3.3.3.	Tablas para mantener la interpretabilidad	23
3.3.4.	Balanceo del dataset	24
3.4.	Análisis de correlaciones	27
3.5.	Análisis de componentes principales (PCA).....	31
3.5.1.	Clustering jerárquico.....	40
3.6.	Reducción de variables PCA	46
3.7.	Clusterización mediante Kmeans	54
3.8.	Modelado Random Forest	60
4.	Resultados.....	62
4.1.	Correas	63
4.2.	Fuga de aire.....	66
4.3.	Motor	69
4.4.	Sistema de Frenos.....	72
4.5.	Sistema de Refrigeración.....	75
5.	Conclusiones	78
5.1.	Futuras líneas de mejora	79
6.	Bibliografía.....	80

Índice de ilustraciones

Ilustración 1. Código, Eliminación Outliers. Ilustración propia	11
Ilustración 2. Código de agrupación. Ilustración propia	12
Ilustración 3. Código de agrupación. Ilustración propia	13
Ilustración 4. Código de agrupación. Ilustración propia	14
Ilustración 5. Código de agrupación. Ilustración propia	15
Ilustración 6. Código de agrupación. Ilustración propia	16
Ilustración 7. Número de observaciones por subárea. Ilustración propia	17
Ilustración 8. Dataset balanceado de correas. Ilustración propia.....	24
Ilustración 9. Dataset balanceado de Fuga de aire. Ilustración propia	25
Ilustración 10. Dataset balanceado de Motor. Ilustración propia	26
Ilustración 11. Dataset balanceado de Sistema de frenos. Ilustración propia...26	26
Ilustración 12. Dataset balanceado de Sistema de refrigeración. Ilustración propia	27
Ilustración 13. Heatmap correlaciones dataset correas	28
Ilustración 14. Heatmap correlaciones dataset fuga de aire	29
Ilustración 15. Heatmap correlaciones dataset motor	29
Ilustración 16. Heatmap correlaciones dataset Sistema de frenos	30
Ilustración 17. Heatmap correlaciones dataset sistema de refrigeración.....30	30
Ilustración 18. Número variables sintéticas que representan el 99% de la varianza para correas. Ilustración propia.....32	32
Ilustración 19. Número variables sintéticas que representan el 99% de la varianza para fuga de aire. Ilustración propia.....32	32
Ilustración 20. Número variables sintéticas que representan el 99% de la varianza para motor. Ilustración propia.....33	33
Ilustración 21. Número variables sintéticas que representan el 99% de la varianza para sistema de frenos. Ilustración propia	33
Ilustración 22. Número variables sintéticas que representan el 99% de la varianza para Sistema de refrigeración. Ilustración propia.....34	34
Ilustración 23. Representación gráfica de las dos primeras variables sintéticas para correar. Ilustración propia	35
Ilustración 24. Representación gráfica de las dos primeras variables sintéticas para fuga de aire. Ilustración propia.....36	36
Ilustración 25. Representación gráfica de las dos primeras variables sintéticas para motor. Ilustración propia.....37	37
Ilustración 26. Representación gráfica de las dos primeras variables sintéticas para sistema de frenos. Ilustración propia	38
Ilustración 27. Representación gráfica de las dos primeras variables sintéticas para sistema de refrigeración. Ilustración propia	39
Ilustración 28. Distancia entre variables en función de la varianza explicada en correas. Ilustración propia	40
Ilustración 29. Distancia entre variables en función de la varianza explicada en Fuga de aire. Ilustración propia.....41	41

Ilustración 30. Distancia entre variables en función de la varianza explicada en motor. Ilustración propia	41
Ilustración 31. Distancia entre variables en función de la varianza explicada en sistema de frenos. Ilustración propia.....	42
Ilustración 32. Distancia entre variables en función de la varianza explicada en sistemas de refrigeración. Ilustración propia	42
Ilustración 33. Dendrograma características de correas. Ilustración propia	43
Ilustración 34. Dendrograma características de fuga de aire. Ilustración propia	44
Ilustración 35. Dendrograma características de motor. Ilustración propia	44
Ilustración 36. Dendrograma características de sistema de frenos. Ilustración propia	45
Ilustración 37. Dendrograma características sistema de refigreación. Ilustración propia	45
Ilustración 38. Numero variables sintéticas que representan el 95% de la varianza para correas. Ilustración propia.....	46
Ilustración 39. Numero variables sintéticas que representan el 95% de la varianza para fuga de aire. Ilustración propia.....	47
Ilustración 40. Numero variables sintéticas que representan el 95% de la varianza para motor. Ilustración propia.....	47
Ilustración 41.Numero variables sintéticas que representan el 95% de la varianza para sistema de frenos2. Ilustración propia	48
Ilustración 42. Numero variables sintéticas que representan el 95% de la varianza para sistema de refrigeración. Ilustración propia	48
Ilustración 43. Distribución PCA para correas. Ilustración propia	49
Ilustración 44. Distribución PCA para fuga de aire. Ilustración propia	50
Ilustración 45. Distribución PCA para motor. Ilustración propia	51
Ilustración 46. Distribución PCA para sistema de frenos. Ilustración propia....	52
Ilustración 47.Distribución PCA para sistema de refrigeración. Ilustración propia	53
Ilustración 48. Clusters para correas. Ilustración propia	54
Ilustración 49. Clusters para fuga de aire. Ilustración propia	55
Ilustración 50. Clusters para motor. Ilustración propia	55
Ilustración 51. Clusters para sistema de frenos. Ilustración propia.....	56
Ilustración 52. Clusters para sistema de refrigeración. Ilustración propia.....	56
Ilustración 53. Correlaciones con kmeans para correas. Ilustración propia.....	57
Ilustración 54. Correlaciones con kmeans para fuga de aire. Ilustración propia	58
Ilustración 55. Correlaciones con kmeans para motor. Ilustración propia.....	58
Ilustración 56. Correlaciones con kmeans para sistema de frenos. Ilustración propia	59
Ilustración 57.Correlaciones con kmeans para sistema de refrigeración. Ilustración propia	59
Ilustración 58. Resultados clúster 0 para correas. Ilustración propia.....	63

Ilustración 59. Resultados clúster 1 para correas. Ilustración propia.....	64
Ilustración 60. Resultados combinados para correas. Ilustración propia	65
Ilustración 61.Resultados clúster 0 para fuga de aire. Ilustración propia.....	66
Ilustración 62. Resultados clúster 1 para fuga de aire. Ilustración propia.....	67
Ilustración 63. Resultados combinados para fuga de aire. Ilustración propia ...	68
Ilustración 64. Resultados clúster 0 para motor. Ilustración propia.....	69
Ilustración 65. Resultados clúster 1 para motor. Ilustración propia.....	70
Ilustración 66. Resultados combinados para motor. Ilustración propia.....	71
Ilustración 67. Resultados clúster 0 para sistema de frenos. Ilustración propia	72
Ilustración 68. Resultados clúster 1 para sistema de frenos. Ilustración propia	73
Ilustración 69. Resultados combinados para sistema de frenos. Ilustración propia	74
Ilustración 70. Resultados clúster 0 para sistema de refrigeración. Ilustración propia	76
Ilustración 71. Resultados clúster 1 para sistema de refrigeración. Ilustración propia	76
Ilustración 72. Resultados combinados para sistema de refrigeración. Ilustración propia	77

Índice de tablas

Tabla 1. Muestra database. Tabla propia	10
Tabla 2. muestra tabla sin operaciones agrupadas. Tabla propia	16
Tabla 3. Muestra target 1 para correas. Tabla propia	18
Tabla 4. Muestra target 1 para fuga de aire. Tabla propia	18
Tabla 5. Muestra target 1 para motor. Tabla propia	19
Tabla 6. Muestra target 1 para sistema de frenos. Tabla propia	19
Tabla 7. Muestra target 1 para sistema de refrigeración. Tabla propia	21
Tabla 8. Missings por característica. Tabla propia	21
Tabla 9. codificaciones variables categóricas. Tabla propia	22
Tabla 10. Normalización de variables numéricas. Tabla propia	22
Tabla 11. Características categoricas y codificacion. Tabla propia	23

1. Introducción

En la presente memoria de prácticas, se explicará en detalle los pasos seguidos para llevar a cabo la predicción de si un autobús ha de ir al taller por una causa en concreto.

Disponemos de una base de datos de autobuses, con diferentes características cada uno de ellos, diferente edad, diferente tipología, etc. También disponemos de información sobre las últimas revisiones en taller y el problema que tuvieron para tener que ir a taller.

Con esta información, se ha procedido a realizar un estudio para predecir la probabilidad de que un autobús x tenga que ir al taller por una razón y.

Para la realización de este estudio, se han utilizado diferentes técnicas vistas a lo largo del Máster de Big Data y Data Science. Las principales asignaturas que se han aplicado han sido visualización, Machine Learning, Minería de datos y Estadística Avanzada.

Por otro lado, el lenguaje de programación usado a lo largo de las prácticas ha sido Python, aunque la extracción de datos desde la base de datos ha sido mediante el lenguaje SQL.

2. Objetivos

Los objetivos principales del presente trabajo de prácticas en empresa es el predecir si un autobús necesitará ir al taller por una avería en concreto, que analizaremos usando los datos extraídos de la base de datos, provista por parte de la empresa.

Los objetivos marcados para alcanzar dicho objetivo se pueden resumir en los siguientes:

- 1) Realizar un análisis descriptivo de los datos de los que disponemos para ganar una mayor interpretabilidad y conocimiento de ellos, para posteriormente sacar conclusiones significativas.
- 2) Realizar la limpieza de datos, para que podamos realizar predicciones y usar esos datos con mayor fiabilidad.
- 3) Encontrar las variables más importantes, que nos ayudarán a tener una mejor predicción sobre la causa de la avería.
- 4) Consensuar con la empresa dichas variables, para enfocarnos en las variables que más valor aportan a la empresa.
- 5) Analizar e investigar diferentes modelos de Machine Learning y ver con cuál de ellos obtenemos unos mejores resultados y por qué. Además, considerar la interpretabilidad de los resultados y si los resultados son entendibles a la hora de explicarlos al cliente.
- 6) Utilizar métodos de visualización vistas a lo largo del Máster para una mejor interpretabilidad para el lector.
- 7) Previsión de las piezas necesarias en los talleres, para pedirlas con el tiempo suficiente, si próximamente un autobús va a ir a taller para cambiar una pieza en concreto o por un motivo en específico.
- 8) Poder ahorrar costes por pedir a grandes volúmenes o volúmenes más pequeños, cuando predecimos el volumen aproximado de piezas que se necesitará para reparar los diferentes autobuses.

3. Desarrollo o metodología

3.1. Descripción del dataset

La empresa nos ha facilitado los datos, extraídos desde una base de datos SQL, en formato .csv. Es con este .csv con el que hemos llevado a cabo nuestro análisis.

El .csv analizado consta de las siguientes variables:

- Código de autobús: se dispone de 342 autobuses diferentes.
- Operaciones realizadas a cada uno de los autobuses.
- Edad del autobús
- Número de días desde la anterior reparación
- Capacidad de cada autobús
- Marca de cada autobús: se dispone de 12 marcas diferentes
- Modelo de cada autobús
- Potencia de cada autobús
- Tipo de autobús: Interurbano, discrecional, urbano y turismo
- Subtipo de cada uno de los autobuses: normal, microbús, minibús, doble piso, VTC y articulado

3.2. Variables del dataset

A continuación, se detalla el nombre de cada una de las variables disponibles en nuestro dataset para un mayor entendimiento de la base de datos.

nomopera: nombre de la operación realizada en el taller

v_codibus: código del autobús

v_edad: edad del autobús

ndias_ant: número de días desde la anterior reparación

capacidad: capacidad del autobús

marca: marca del autobús

modelo: modelo del autobús

potencia: potencia del autobús

tipo: depende del uso y del modelo del vehículo

subtipo: depende de la capacidad y la longitud del vehículo

A continuación, podemos visualizar una pequeña muestra de cómo es nuestro dataset.

Podemos ver que aparecen las variables mencionadas anteriormente y que hay mucha diversidad de variables; es decir, nos encontramos con variables numéricas y variables no numéricas.

Además, también disponemos de variables a las que posteriormente se les deberá llevar a cabo una transformación para mejorar su uso e interpretabilidad.

Tabla 1. Muestra database. Tabla propia.

	nomopera	v_codigibus	v_edad	ndias_ant	capacidad	marca	modelo	potencia	tipo	subtipo
0	Flexible escape	B117	6	1866.0	74.0	MAN	TOURING INTERCITY	228.0	Interurbano	Normal
1	Cambio filtro retarde	U198	24	8471.0	73.0	MAN	SIN ASIGNAR	NaN	BUS TURISTICO	Turistic Valencia
2	Fuga refrigerante	U198	24	8471.0	73.0	MAN	SIN ASIGNAR	NaN	BUS TURISTICO	Turistic Valencia
3	Fuga refrigerante	B117	6	3.0	74.0	MAN	TOURING INTERCITY	228.0	Interurbano	Normal
4	Alternador	X110	2	703.0	16.0	FORD	TRANSIT	92.0	Discrecional	Microbus
...
49403	Fuga de aire	B158	14	44.0	53.0	MAN	LION'S COACH	324.0	Discrecional	Normal
49404	Fallo en AVS	J266	17	1.0	55.0	SCANIA	ATLANTIS	280.0	Discrecional	Normal
49405	Fuga refrigerante	E652	4	4.0	5.0	FORD	MONDEO BA7	NaN	TURISMO	NaN
49406	Plataforma elevadora	T252	14	7.0	27.0	IVECO	DIVO II INTERCITY	259.0	Discrecional	Normal
49407	Correas	A109	13	1.0	36.0	TEMSA	OPALIN 8.4	NaN	MICROBUS	Micro 35 Plazas

49408 rows × 10 columns

Se puede observar en la imagen superior, que disponemos de un total de 49808 observaciones y 10 variables diferentes.

3.3. Eliminación de factores no posibles/outliers

Tras diversas reuniones con el equipo de la empresa de prácticas, se ha consensuado que hay ciertas operaciones que no pueden ser posible y se derivan de errores. Es por ello que antes de realizar nuestro estudio y análisis en detalle, vamos a proceder a su eliminación para, así, poder centrarnos en las situaciones que pueden darse de una manera real y que nuestros datos no estén alterados.

Por lo comentado anteriormente, se consensua con el equipo de la empresa, que nos vamos a focalizar en el análisis de los autobuses en los que el número de días desde la revisión anterior oscila entre 0 y 150. Se procede, por lo tanto, a eliminar los registros cuyo número de días desde la anterior reparación es negativo o superior a 150 ya que estos casos son outliers y no son posibles en la vida real. Es decir, pueden estar derivados de error de medición.

A continuación, podemos ver el código para la eliminación de dicho outliers.

```
df_drop=df[df['ndias_ant']>150].index  
df=df.drop(df_drop)  
df_drop=df[df['ndias_ant']<=0].index  
df=df.drop(df_drop)  
df=df.reset_index(drop = True)  
df
```

Ilustración 1. Código, Eliminación Outliers. Ilustración propia

3.3.1. Agrupación de las operaciones

Para obtener nuestra variable objetivo (target), se crea una columna **grupopera** que se decide, mediante reunión con el equipo, agrupar las diferentes observaciones por similitud en el tipo de reparación o sistema al que pertenezcan.

Para el sistema de refrigeración, se ha decidido que se agrupan las operaciones que contengan: Bomba de agua, intercooler, manguito calefacción, manguito motor, radiador, termostato, fuga refrigerante y presión en el circuito de refrigeración. Se puede visualizar esta agrupación en la ilustración número 2.

Para el sistema de dirección se agrupan las operaciones que contengan: caja de dirección, fuga aceite servodirección, rótula dirección, alineado dirección, timonería dirección. Se puede ver como se ha hecho esta agrupación en la ilustración 2.

Para el sistema de frenos se agrupan las operaciones que contengan: Frenos de cualquier eje, disco y bomba de freno. Se puede ver como se ha hecho esta agrupación en la ilustración 2.

SISTEMA DE REFRIGERACIÓN

```
df['grupopera'] = np.where(
    (df['nomopera'].str.contains('Bomba agua'))|
    (df['nomopera'].str.contains('Intercooler'))|
    (df['nomopera'].str.contains('Manguito/s calefaccion'))|
    (df['nomopera'].str.contains('Manguito/s motor'))|
    (df['nomopera'].str.contains('Radiador'))|
    (df['nomopera'].str.contains('Termostatos'))|
    (df['nomopera'].str.contains('Fuga refrigerante'))|
    (df['nomopera'].str.contains('Presion en circuito refrigeracion'))|
    , 'Sistema refrigeración',np.nan)
```

SISTEMA DE DIRECCIÓN

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Caja direccion'))|
    (df['nomopera'].str.contains('Fuga aceite servodireccion'))|
    (df['nomopera'].str.contains('Rotulas direccion'))|
    (df['nomopera'].str.contains('Alineado de direccion'))|
    (df['nomopera'].str.contains('Timoneria direccion'))|
    , 'Sistema dirección',df['grupopera'])
```

SISTEMA FRENO

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Frenos 1º eje'))|
    (df['nomopera'].str.contains('Frenos 2º eje'))|
    (df['nomopera'].str.contains('Frenos 3º eje'))|
    (df['nomopera'].str.contains('Freno mano'))|
    (df['nomopera'].str.contains('Disco'))|
    (df['nomopera'].str.contains('Bomba freno'))|
    , 'Sistema frenos',df['grupopera'])
```

Ilustración 2. Código de agrupación. Ilustración propia

Para el sistema de escape se agrupan las operaciones que contengan, flexible escape figa escape, catalizador, silencioso. Se puede ver como se ha hecho esta agrupación en la ilustración 3.

Para el sistema de amortiguación se agrupan las operaciones que contienen: regular suspensión, amortiguador o amortiguadores en cualquier eje, tirantes y miembros de cualquier eje. Se puede ver como se ha hecho esta agrupación en la ilustración 3.

Para el sistema de alimentación se agrupan las operaciones que contengan: inyectores o el reglaje de los mismo. Se puede ver como se ha hecho esta agrupación en la ilustración 3.

SISTEMA DE ESCAPE

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Flexible escape'))|
    (df['nomopera'].str.contains('Fuga escape'))|
    (df['nomopera'].str.contains('Catalizador'))|
    (df['nomopera'].str.contains('Silencioso'))|
    , 'Sistema de escape', df['grupopera'])
```

SISTEMA AMORTIGUACIÓN

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Regular suspension'))|
    (df['nomopera'].str.contains('Amortiguador 2º eje derecho'))|
    (df['nomopera'].str.contains('Amortiguadores 2º eje'))|
    (df['nomopera'].str.contains('Amortiguadores 3º eje'))|
    (df['nomopera'].str.contains('Tirantes/silembrocs 1º eje'))|
    (df['nomopera'].str.contains('Tirantes/silembrocs 2º eje'))|
    , 'Sistema de amortiguación', df['grupopera'])
```

SISTEMA DE ALIMENTACIÓN

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Inyectores'))|
    (df['nomopera'].str.contains('Reglaje inyectores'))|
    , 'Sistema de alimentación', df['grupopera'])
```

Ilustración 3. Código de agrupación. Ilustración propia

Para el Motor se agrupan las operaciones que contengan, motor, árbol de levas, casquillos de biela, turbo y manguito de admisión, menos turbina motora que irá a parte y se describe como se agrupa posteriormente. Se puede ver como se ha hecho esta agrupación en la ilustración 4.

Para el precalentador se agrupan las operaciones que contengan, precalentador. Se puede ver como se ha hecho esta agrupación en la ilustración 4.

Para el alternador se agrupan las operaciones que contengan, alternador. Se puede ver como se ha hecho esta agrupación en la ilustración 4.

Para el ABS/EBS se agrupan las operaciones que contengan, ABS/EBS. Se puede ver como se ha hecho esta agrupación en la ilustración 4.

Para el Adblue se agrupan las operaciones que contengan, adblue. Se puede ver como se ha hecho esta agrupación en la ilustración 4.

MOTOR

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Cambiar turbina/motor condensadora'))|
    (df['nomopera'].str.contains('Culata del compresor del motor'))|
    (df['nomopera'].str.contains('Fuga aceite motor'))|
    (df['nomopera'].str.contains('Limpiar motor'))|
    (df['nomopera'].str.contains('Motor'))|
    (df['nomopera'].str.contains('Poleas motor'))|
    (df['nomopera'].str.contains('Reten trasero motor'))|
    (df['nomopera'].str.contains('Ruido en motor'))|
    (df['nomopera'].str.contains('Arbol de levas'))|
    (df['nomopera'].str.contains('Casquillos de biela'))|
    (df['nomopera'].str.contains('Potencia motor'))|
    (df['nomopera'].str.contains('Turbo'))|
    (df['nomopera'].str.contains('Manguito/s admision'))|
    (df['nomopera'].str.contains('Tacos motor'))|
    , 'Motor', df['grupopera'])
```

PRECALENTADOR

```
df['grupopera'] = np.where((df['nomopera'].str.contains('recalentador'))|
    , 'Precalentador', df['grupopera'])
```

ALTERNADOR

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Alternador'))|
    # (df['nomopera'].str.contains('Soporte alternador'))|
    (df['nomopera'].str.contains('Alternador auxiliar'))|
    , 'Alternador', df['grupopera'])
```

ABS/EBS

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Fallo en ABS/EBS'))|
    , 'ABS/EBS', df['grupopera'])
```

ADBLUE

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Fallo adblue'))|
    (df['nomopera'].str.contains('Fuga adblue'))|
    , 'Adblue', df['grupopera'])
```

Ilustración 4. Código de agrupación. Ilustración propia

Para las baterías se agrupan las operaciones que contengan, Batería. Se puede ver como se ha hecho esta agrupación en la ilustración 5.

Para correas se agrupan las operaciones que contengan, correa o correas. Se puede ver como se ha hecho esta agrupación en la ilustración 5.

Para reglaje de válvulas se agrupan las operaciones que contengan, válvula o válvulas. Se puede ver como se ha hecho esta agrupación en la ilustración 5.

Para filtro de partículas se agrupan las operaciones que contengan, filtro de partículas. Se puede ver como se ha hecho esta agrupación en la ilustración 5.

Para EDC se agrupan las operaciones que contengan, EDC. Se puede ver como se ha hecho esta agrupación en la ilustración 5.

Para Fuga de aire se agrupan las operaciones que contengan, Fuga de aire. Se puede ver como se ha hecho esta agrupación en la ilustración 5.

Para turbina motor de aire se agrupan las operaciones que contengan, turbina de motor.

Se puede ver como se ha hecho esta agrupación en la ilustración 5.

BATERIAS

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Baterias'))|
                            (df['nomopera'].str.contains('Carga de baterias'))|
                            , 'Baterias',df['grupopera'])
```

CORREAS

```
df['grupopera'] = np.where((df['nomopera'].str.contains('orrea'))|
                            , 'Correas',df['grupopera'])
```

REGLAJE VÁLVULAS

```
df['grupopera'] = np.where((df['nomopera'].str.contains('valvula'))|
                            , 'Reglaje válvulas',df['grupopera'])
```

FILTRO PARTICULAS

```
df['grupopera'] = np.where((df['nomopera'].str.contains('iltro part'))|
                            , 'Filtro de Particulas',df['grupopera'])
```

EDC

```
df['grupopera'] = np.where((df['nomopera'].str.contains('EDC'))|
                            , 'Fallo en EDC',df['grupopera'])
```

FUGA DE AIRE

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Fuga de aire'))|
                            , 'Fuga de aire',df['grupopera'])
```

TURBINA MOTOR

```
df['grupopera'] = np.where((df['nomopera'].str.contains('turbina'))|
                            , 'Cambiar turbina/motor condensadora',df['grupopera'])
```

Ilustración 5. Código de agrupación. Ilustración propia

Para gasoil de aire se agrupan las operaciones que contengan, Fuga de aire. Se puede ver como se ha hecho esta agrupación en la ilustración 6.

GASOIL

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Deposito de gasoil'))|
#(df['nomopera'].str.contains('Cambio filtro gasoil'))|
(df['nomopera'].str.contains('Fuga gasoil'))|
,'Gasoil',df['grupopera'])
```

FALLO EN AVS

```
df['grupopera'] = np.where((df['nomopera'].str.contains('AVS'))|
,'Fallo en AVS',df['grupopera'])
```

FUELLE

```
df['grupopera'] = np.where((df['nomopera'].str.contains('Fuelle 1º eje derecho'))|
(df['nomopera'].str.contains('Fuelle 1º eje izquierdo'))|
(df['nomopera'].str.contains('Fuelles 2º eje'))|
(df['nomopera'].str.contains('Fuelles1º eje'))|
,'Fuelle',df['grupopera'])
```

PINCHAZO

```
df['grupopera'] = np.where((df['nomopera'].str.contains('pinchazo'))|
,'Pinchazo',df['grupopera'])
```

Ilustración 6. Código de agrupación. Ilustración propia

Se eliminan de operaciones no agrupadas del dataset. Como se muestra en la Tabla 2.

Tabla 2. Muestra tabla sin operaciones agrupadas. Tabla propia

	grupopera	marca	modelo	v_codigbus	tipo	v_edad	capacidad	ndias_ant	subtipo	potencia
0	Sistema refrigeración	MAN	TOURING INTERCITY	B117	Interurbano	6	74.0	3.0	Normal	228.0
1	Sistema frenos	MERCEDES BENZ	TOURING	T227	Discrecional	4	38.0	3.0	Normal	310.0
4	Sistema de alimentación	MAN	SIN ASIGNAR	U208	BUS TURISTICO	24	71.0	22.0	Turistic Valencia	NaN
6	Baterias	SCANIA	CS40 CITY II	H086	Urbano	1	100.0	18.0	Normal	191.0
8	Correas	MAN	SIN ASIGNAR	U238	BUS TURISTICO	25	81.0	12.0	Turistic Valencia	NaN
...
45946	Sistema frenos	MAN	i6 13.37	C099	Discrecional	7	71.0	147.0	Normal	353.0
45947	Fuga de aire	MAN	LION'S COACH	B158	Discrecional	14	53.0	44.0	Normal	324.0
45948	Fallo en AVS	SCANIA	ATLANTIS	J266	Discrecional	17	55.0	1.0	Normal	280.0
45949	Sistema refrigeración	FORD	MONDEO BA7	E652	TURISMO	4	5.0	4.0	NaN	NaN
45951	Correas	TEMSA	OPALIN 8.4	A109	MICROBUS	13	36.0	1.0	Micro 35 Plazas	NaN

18298 rows × 10 columns

Vemos que nos quedan 18298 observaciones.

3.3.2. Observaciones por subárea:

Podemos ver de forma gráfica como el número de observaciones de cada categoría, es muy diferente, lo que nos indica que no están balanceadas las categorías, dando el modelo predictivo más importancia a las categorías que tienen más observaciones.

En la ilustración 7 se nos indica el número de observaciones por subáreas. Este fue el motivo principal por el que no se intentó predecir la categoría.

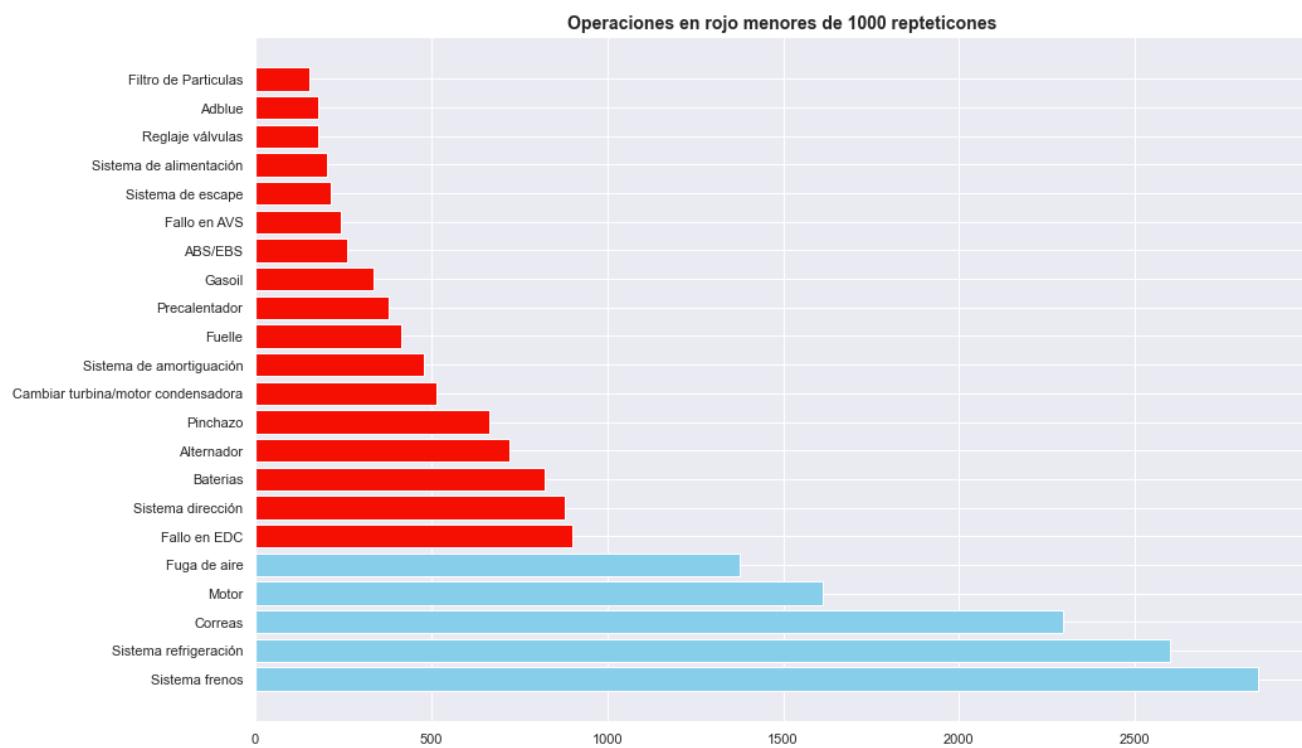


Ilustración 7. Número de observaciones por subárea. Ilustración propia

Para proseguir, se decidió analizar las subáreas con más de 1000 observaciones. Ya que tenían un número suficiente como para poder implementar un modelo.

Se decidió aplicar modelos predictivos, sobre, si sucedía una Fuga de aire, un problema en el motor, un problema por correas, si es sistema de refrigeración se averiaba y si el sistema de frenos tenía una avería.

Se realizarán 5 scrips diferentes en Jupyter uno por cada subárea. Para poder implementar un modelo para cada uno de ellos, los cuales se van a presentar de forma paralela, ya que el proceso fue muy similar.

3.3.1. Creación de la variable objetivo (target)

Correas

Para lograr lo descrito en el apartado anterior, si la reparación es por el grupo correas se añade a la columna target un 1 si no un 0, creando de esta forma una variable auxiliar, también llamada *Dummy* que nos permita luego seleccionar un subconjunto de datos que solo contenga, en el primer caso las averías de correas como se muestra en la tabla 3 a continuación.

Tabla 3. Muestra target 1 para correas. Tabla propia

	grupopera	marca	modelo	v_codigibus	tipo	v_edad	capacidad	ndias_ant	subtipo	potencia	target
0	Sistema refrigeración	MAN	TOURING INTERCITY	B117	Interurbano	6	74.0	3.0	Normal	228.0	0
1	Sistema frenos	MERCEDES BENZ	TOURING	T227	Discrecional	4	38.0	3.0	Normal	310.0	0
4	Sistema de alimentación	MAN	SIN ASIGNAR	U208	BUS TURISTICO	24	71.0	22.0	Turistic Valencia	NaN	0
6	Baterías	SCANIA	CS40 CITY II	H086	Urbano	1	100.0	18.0	Normal	191.0	0
8	Correas	MAN	SIN ASIGNAR	U238	BUS TURISTICO	25	81.0	12.0	Turistic Valencia	NaN	1
...
45946	Sistema frenos	MAN	16 13.37	C099	Discrecional	7	71.0	147.0	Normal	353.0	0
45947	Fuga de aire	MAN	LION'S COACH	B158	Discrecional	14	53.0	44.0	Normal	324.0	0
45948	Fallo en AVS	SCANIA	ATLANTIS	J266	Discrecional	17	55.0	1.0	Normal	280.0	0
45949	Sistema refrigeración	FORD	MONDEO BA7	E652	TURISMO	4	5.0	4.0	NaN	NaN	0
45951	Correas	TEMSA	OPALIN 8.4	A109	MICROBUS	13	36.0	1.0	Micro 35 Plazas	NaN	1

18298 rows × 11 columns

Fuga de aire

En el segundo script, si la reparación es por el grupo fuga de aire se añade a la columna target un 1 si no un 0 como se muestra en la tabla 4

Tabla 4. Muestra target 1 para fuga de aire. Tabla propia

	grupopera	marca	modelo	v_codigbus	tipo	v_edad	capacidad	ndias_ant	subtipo	potencia	target
0	Sistema refrigeración	MAN	TOURING INTERCITY	B117	Interurbano	6	74.0	3.0	Normal	228.0	0
1	Sistema frenos	MERCEDES BENZ	TOURING	T227	Discrecional	4	38.0	3.0	Normal	310.0	0
4	Sistema de alimentación	MAN	SIN ASIGNAR	U208	BUS TURISTICO	24	71.0	22.0	Turistic Valencia	NaN	0
6	Baterías	SCANIA	CS40 CITY II	H086	Urbano	1	100.0	18.0	Normal	191.0	0
8	Correas	MAN	SIN ASIGNAR	U238	BUS TURISTICO	25	81.0	12.0	Turistic Valencia	NaN	0
...
45946	Sistema frenos	MAN	i6 13.37	C099	Discrecional	7	71.0	147.0	Normal	353.0	0
45947	Fuga de aire	MAN	LION'S COACH	B158	Discrecional	14	53.0	44.0	Normal	324.0	1
45948	Fallo en AVS	SCANIA	ATLANTIS	J266	Discrecional	17	55.0	1.0	Normal	280.0	0
45949	Sistema refrigeración	FORD	MONDEO BA7	E652	TURISMO	4	5.0	4.0	NaN	NaN	0
45951	Correas	TEMSA	OPALIN 8.4	A109	MICROBUS	13	36.0	1.0	Micro 35 Plazas	NaN	0

18298 rows x 11 columns

Motor

En el tercer script, si la reparación es por el grupo de motor se añade a la columna target un 1, si no un 0 como se muestra en la tabla 5

Tabla 5. Muestra target 1 para motor. Tabla propia

	grupopera	marca	modelo	v_codigbus	tipo	v_edad	capacidad	ndias_ant	subtipo	potencia	target
0	Sistema refrigeración	MAN	TOURING INTERCITY	B117	Interurbano	6	74.0	3.0	Normal	228.0	0
1	Sistema frenos	MERCEDES BENZ	TOURING	T227	Discrecional	4	38.0	3.0	Normal	310.0	0
4	Sistema de alimentación	MAN	SIN ASIGNAR	U208	BUS TURISTICO	24	71.0	22.0	Turistic Valencia	NaN	0
6	Baterías	SCANIA	CS40 CITY II	H086	Urbano	1	100.0	18.0	Normal	191.0	0
8	Correas	MAN	SIN ASIGNAR	U238	BUS TURISTICO	25	81.0	12.0	Turistic Valencia	NaN	0
...
45946	Sistema frenos	MAN	i6 13.37	C099	Discrecional	7	71.0	147.0	Normal	353.0	0
45947	Fuga de aire	MAN	LION'S COACH	B158	Discrecional	14	53.0	44.0	Normal	324.0	0
45948	Fallo en AVS	SCANIA	ATLANTIS	J266	Discrecional	17	55.0	1.0	Normal	280.0	0
45949	Sistema refrigeración	FORD	MONDEO BA7	E652	TURISMO	4	5.0	4.0	NaN	NaN	0
45951	Correas	TEMSA	OPALIN 8.4	A109	MICROBUS	13	36.0	1.0	Micro 35 Plazas	NaN	0

18298 rows x 11 columns

Sistema de Frenos

En el cuarto script, si la reparación es por el grupo de sistema de frenos, se añade a la columna target, un 1, si no, un 0 como se muestra en la tabla 6

Tabla 6. Muestra target 1 para sistema de frenos. Tabla propia

	grupopera	marca	modelo	v_codigbus	tipo	v_edad	capacidad	ndias_ant	subtipo	potencia	target
0	Sistema refrigeración	MAN	TOURING INTERCITY	B117	Interurbano	6	74.0	3.0	Normal	228.0	0
1	Sistema frenos	MERCEDES BENZ	TOURING	T227	Discrecional	4	38.0	3.0	Normal	310.0	1
4	Sistema de alimentación	MAN	SIN ASIGNAR	U208	BUS TURISTICO	24	71.0	22.0	Turistic Valencia	NaN	0
6	Baterias	SCANIA	CS40 CITY II	H086	Urbano	1	100.0	18.0	Normal	191.0	0
8	Correas	MAN	SIN ASIGNAR	U238	BUS TURISTICO	25	81.0	12.0	Turistic Valencia	NaN	0
...
45946	Sistema frenos	MAN	i6 13.37	C099	Discrecional	7	71.0	147.0	Normal	353.0	1
45947	Fuga de aire	MAN	LION'S COACH	B158	Discrecional	14	53.0	44.0	Normal	324.0	0
45948	Fallo en AVS	SCANIA	ATLANTIS	J266	Discrecional	17	55.0	1.0	Normal	280.0	0
45949	Sistema refrigeración	FORD	MONDEO BA7	E652	TURISMO	4	5.0	4.0	NaN	NaN	0
45951	Correas	TEMSA	OPALIN 8.4	A109	MICROBUS	13	36.0	1.0	Micro 35 Plazas	NaN	0

18298 rows × 11 columns

Sistema de refrigeración

En el quinto script, si la reparación es por el grupo de sistema de refrigeración, se añade a la columna target, un 1, si no, un 0 como se muestra en la tabla 7

Tabla 7. Muestra target 1 para sistema de refrigeración. Tabla propia

	grupopera	marca	modelo	v_codigbus	tipo	v_edad	capacidad	ndias_ant	subtipo	potencia	target
0	Sistema refrigeración	MAN	TOURING INTERCITY	B117	Interurbano	6	74.0	3.0	Normal	228.0	1
1	Sistema frenos	MERCEDES BENZ	TOURING	T227	Discrecional	4	38.0	3.0	Normal	310.0	0
4	Sistema de alimentación	MAN	SIN ASIGNAR	U208	BUS TURISTICO	24	71.0	22.0	Turistic Valencia	NaN	0
6	Baterías	SCANIA	CS40 CITY II	H086	Urbano	1	100.0	18.0	Normal	191.0	0
8	Correas	MAN	SIN ASIGNAR	U238	BUS TURISTICO	25	81.0	12.0	Turistic Valencia	NaN	0
...
45946	Sistema frenos	MAN	i6 13.37	C099	Discrecional	7	71.0	147.0	Normal	353.0	0
45947	Fuga de aire	MAN	LION'S COACH	B158	Discrecional	14	53.0	44.0	Normal	324.0	0
45948	Fallo en AVS	SCANIA	ATLANTIS	J266	Discrecional	17	55.0	1.0	Normal	280.0	0
45949	Sistema refrigeración	FORD	MONDEO BA7	E652	TURISMO	4	5.0	4.0	NaN	NaN	1
45951	Correas	TEMSA	OPALIN 8.4	A109	MICROBUS	13	36.0	1.0	Micro 35 Plazas	NaN	0

18298 rows x 11 columns

Se comprueban los missings por cada característica, después de analizarlos con el equipo, se decide eliminarlos, en vez de imputarlos, para que no generen datos erróneos, que puedan confundir al modelo. Podemos ver la cantidad de missings en la tabla 8.

En los 5 scripts obtenemos valores muy similares.

Tabla 8. Missings por característica. Tabla propia

```

grupopera      0.00
marca          1.88
modelo          2.13
v_codigbus     0.00
tipo            2.32
v_edad          0.00
capacidad       2.72
ndias_ant       1.88
subtipo          6.23
potencia        19.90
target          0.00
dtype: float64

```

3.3.2. Codificación de características y normalización

Se codifican numéricamente las características para poderlas utilizar en los modelos, ya que la mayoría de los modelos requieren valores numéricos para poder incluirlos en los algoritmos de predicción. Como se muestra a continuación en la Tabla 9

Tabla 9. Codificaciones variables categóricas. Tabla propia

	grupopera	marca	modelo	v_codigbus	tipo	v_edad	subtipo
0	Sistema refrigeración	1	1		1	6	1
1	Sistema frenos	2	2		2	4	1
2	Baterias	3	3		3	1	1
3	Sistema frenos	1	4		4	1	1
4	Sistema frenos	1	4		4	1	1
...
14535	Sistema frenos	2	14		23	2	15
14536	Sistema frenos	1	41		78	2	7
14537	Sistema frenos	1	41		78	2	7
14538	Fuga de aire	1	21		85	2	14
14539	Fallo en AVS	3	16		28	2	17

Posteriormente normalizamos los valores numéricos para que ninguna característica tome más relevancia que otra, y se une los valores normalizados con los codificados, se elimina la columna **grupopera** ya que ya tenemos nuestro objetivo a predecir que es la columna **target**, como se muestra en la Tabla 10.

Tabla 10. Normalización de variables numéricas. Tabla propia

	marca	modelo	v_codigbus	tipo	v_edad	subtipo	capacidad	ndias_ant	potencia	target
10334	2	19	136	2	17	1	0.310127	0.033784	0.832215	0.0
5329	6	37	61	2	11	2	0.025316	0.013514	0.201342	0.0
6370	3	16	28	2	13	1	0.316456	0.027027	0.731544	0.0
10705	2	26	211	2	10	2	0.069620	0.141892	0.244966	0.0
225	1	21	36	2	1	1	0.303797	0.141892	0.926174	0.0
...
6126	1	2	6	2	14	1	0.310127	0.128378	0.802013	0.0
14101	3	13	25	2	12	1	0.316456	0.175676	0.728188	0.0
1663	1	2	6	2	9	1	0.310127	0.047297	0.802013	0.0
13431	1	24	40	2	11	3	0.189873	0.013514	0.382550	0.0
4259	1	17	30	2	7	1	0.354430	0.040541	0.879195	0.0

1959 rows x 10 columns

3.3.3. Tablas para mantener la interpretabilidad

Para poder mantener la interpretabilidad de las variables se han creado las siguientes tablas que se han juntado en la Tabla 11.

Mediante estas tablas podemos saber qué código pertenece a qué código de autobús, marca, modelo, tipo y subtípico.

Tabla 11. Características categóricas y codificación. Tabla propia

v_codigbus	codigo	Marca	codigo	modelo	codigo
0	B117	1	0	MAN	1
1	T227	2	1	MERCEDES BENZ	2
2	H086	3	2	SCANIA	3
3	B129	4	3	FORD	4
4	H076	5	4	VOLVO	5
...	5	IVECO	6
338	82HC	339	6	RENAULT	7
339	F193	340	7	DAF	8
340	63HC	341	8	CUMMINS	9
341	THP0313	342	9	IVECO/FIAT	10
342	ALS0314	343	10	FIAT	11
			11	PEUGEOT	12
			12	VOLKSWAGEN	13
tipo	codigo	subtipo	codigo		
0	Interurbano	1	0	Normal	1
1	Discrecional	2	1	Microbus	2
2	Urbano	3	2	Midibus	3
3	Turismo	4	3	Doble Piso	4
			4	VTC	5
			5	Articulado	6

3.3.4. Balanceo del dataset

Para crear un data set balanceado entre 0 y 1 en la variable objetivo se han separado las observaciones con target igual a 1, y posteriormente seleccionando de forma aleatoria un número parecido de observaciones con target igual 0.

Para cada uno de los datasets se ha seleccionado un numero distinto de observaciones con target igual a 0, dependiendo de las observaciones con target igual a 1 que tengamos en cada subárea.

Correas

En la Ilustración 8 podemos ver como se ha balanceado correctamente el dataset para correas como target igual a 1

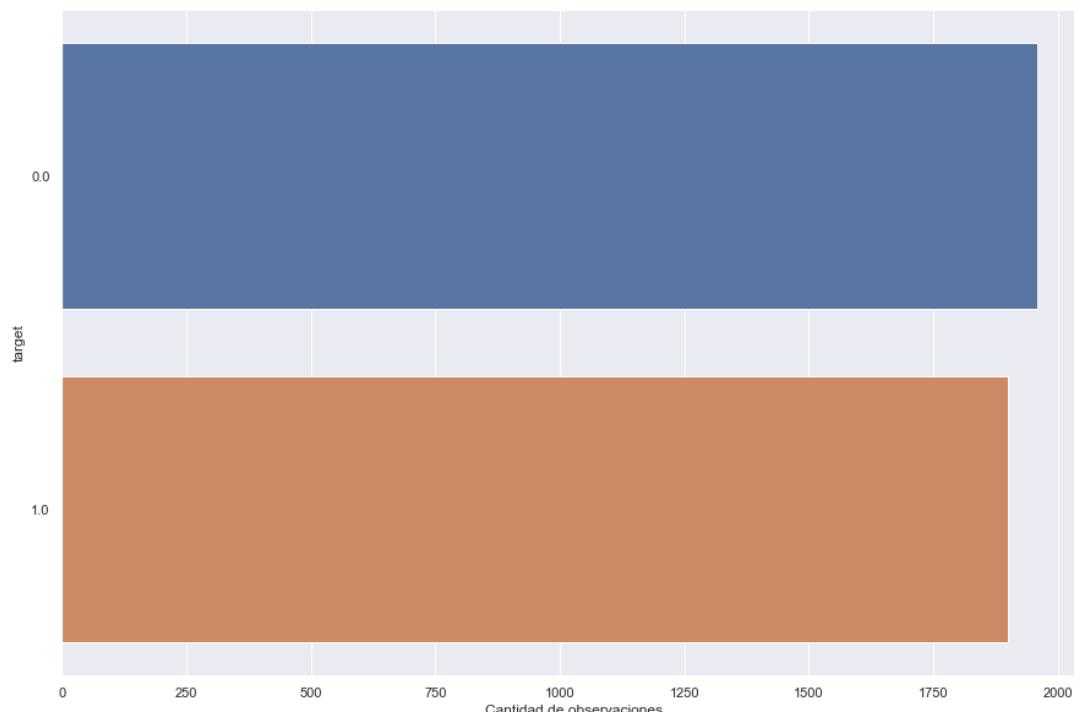


Ilustración 8. Dataset balanceado de correas. Ilustración propia

Fuga de aire

En la Ilustración 9 podemos ver como se ha balanceado correctamente el dataset para Fuga de aire como target igual a 1

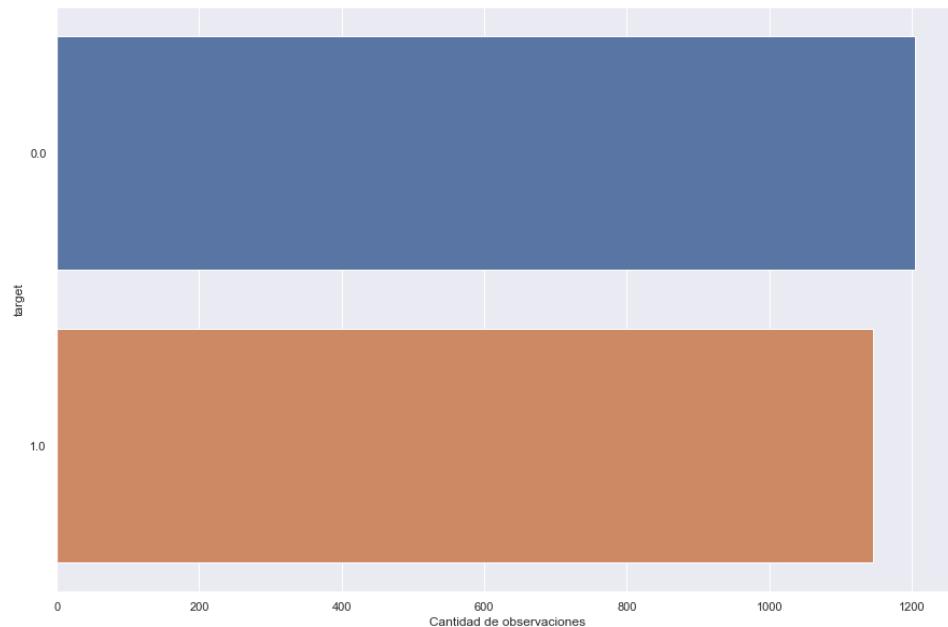


Ilustración 9. Dataset balanceado de Fuga de aire. Ilustración propia

Motor

En la Ilustración 10 podemos ver como se ha balanceado correctamente el dataset para motor como target igual a 1

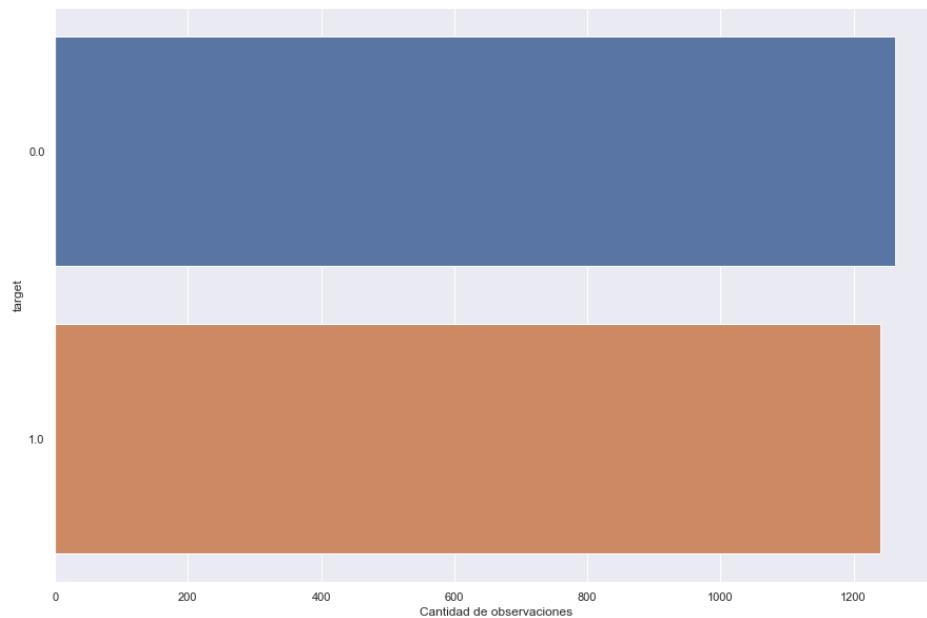


Ilustración 10. Dataset balanceado de Motor. Ilustración propia

Sistema de frenos

En la Ilustración 11 podemos ver como se ha balanceado correctamente el dataset para sistema de frenos como target igual a 1

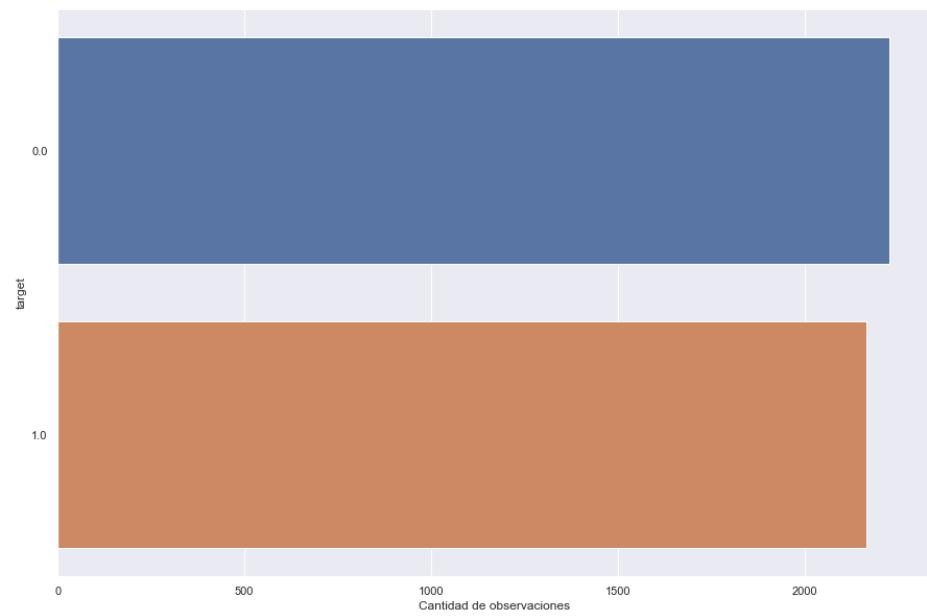


Ilustración 11. Dataset balanceado de Sistema de frenos. Ilustración propia

Sistema de refrigeración

En la Ilustración 11 podemos ver como se ha balanceado correctamente el dataset para sistema de frenos como target igual a 1

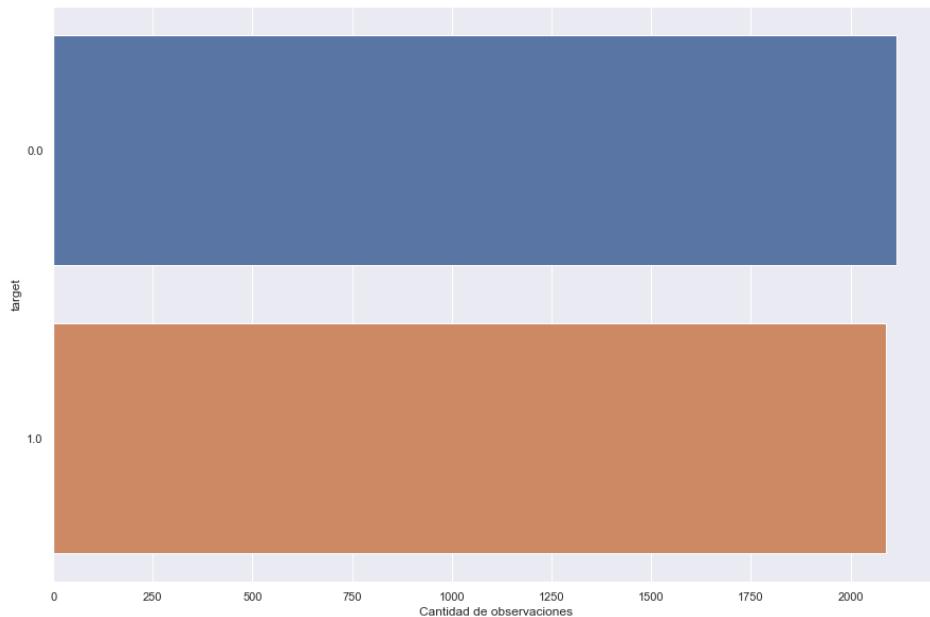


Ilustración 12. Dataset balanceado de Sistema de refrigeración. Ilustración propia

3.4. Análisis de correlaciones

Se determina por criterio experto, reuniones con el equipo y aplicación de la asignatura de Minería de Datos, que teóricamente un 0,8 es una correlación alta, y que si hay una correlación entre dos variables (excluyendo la variable objetivo), de 0.8 o más, deberían ser estudiadas por si hubiera que eliminar una de ellas.

Podemos determinar que no hay correlación con valores altos entre características, por lo que se deduce que no van a generar información similar, y además se observa que no hay una alta correlación con la variable objetivo target, lo que indica que no hay ninguna variable que pueda aportar más información que el resto a la hora de predecir la variable objetivo en función de la correlación.

La correlación más alta que podemos ver es, entre el 60% y el 65% dependiendo del dataset, entre potencia y subtípico, a continuación se muestra el detalle de las correlaciones por cada subáreas.

Correas

A continuación, en la ilustración 13 se muestra en detalle la correlaciones para el dataset de la subárea correas



Ilustración 13. Heatmap correlaciones dataset correas

Fuga de aire

A continuación, en la ilustración 14 se muestra en detalle la correlaciones para el dataset de la subárea fuga de aire.



Ilustración 14. Heatmap correlaciones dataset fuga de aire

Motor

A continuación, en la ilustración 15 se muestra en detalle la correlaciones para el dataset de la subárea Motor.



Ilustración 15. Heatmap correlaciones dataset motor

Sistema de frenos

A continuación, en la ilustración 16 se muestra en detalle la correlaciones para el dataset de la subárea sistema de frenos.



Ilustración 16. Heatmap correlaciones dataset Sistema de frenos

Sistema de refrigeración

A continuación, en la ilustración 17 se muestra en detalle la correlaciones para el dataset de la subárea Sistema de refrigeración.



Ilustración 17. Heatmap correlaciones dataset sistema de refrigeración

3.5. Análisis de componentes principales (PCA)

Como hemos visto en el apartado anterior, las correlaciones no nos dan mucha información ya que las correlaciones con la variable objetivo son muy bajas, para todos los casos, en los 5 datasets, por lo que se va a hacer un análisis de las componentes principales.

El objetivo de la PCA es reducir la dimensionalidad, a costa de la interpretabilidad, utilizando nuevas variables sintéticas no correlacionadas, para describir el conjunto de datos, determinando el orden de las nuevas variables en función de la varianza original que describen.

Al emplearlo sobre la matriz transpuesta y graficando estas en función de las dos primeras variables sintéticas en el caso de que estás expliquen un alto porcentaje de la varianza, podemos ver como se distribuyen las características, y si aportan información similar en función de su varianza.

El primer paso del análisis es comprobar cómo se distribuyen las variables utilizando PCA sobre la traspuesta del conjunto de datos.

Vemos que, con 2 variables sintéticas, podemos representar el 99% de la distribución en varianza de todas las componentes, en todos los casos como se puede ver en las Ilustraciones 18,19,20,21 y 22.

Correas

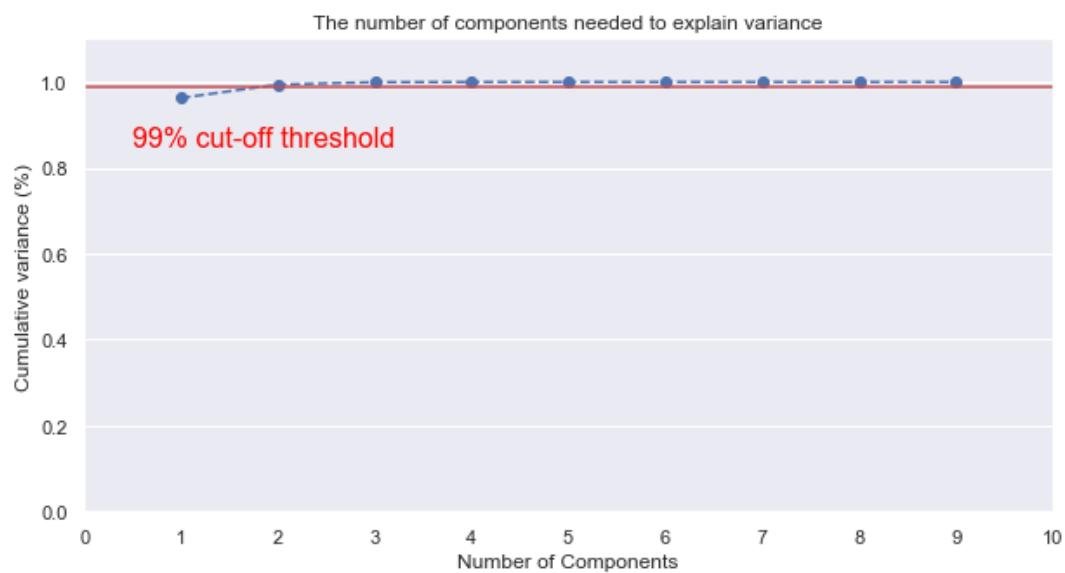


Ilustración 18. Número variables sintéticas que representan el 99% de la varianza para correas. Ilustración propia

Fuga de aire

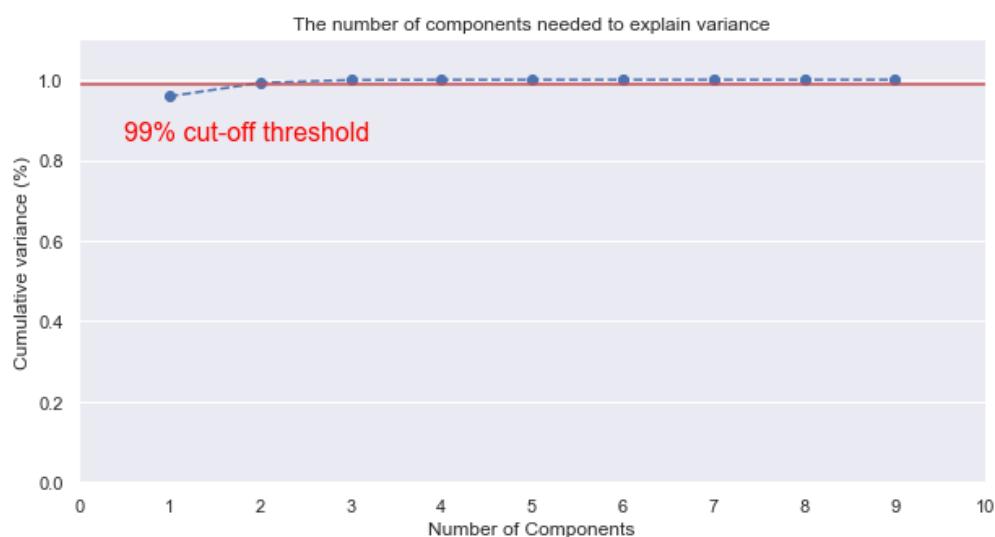


Ilustración 19. Número variables sintéticas que representan el 99% de la varianza para fuga de aire. Ilustración propia

Motor

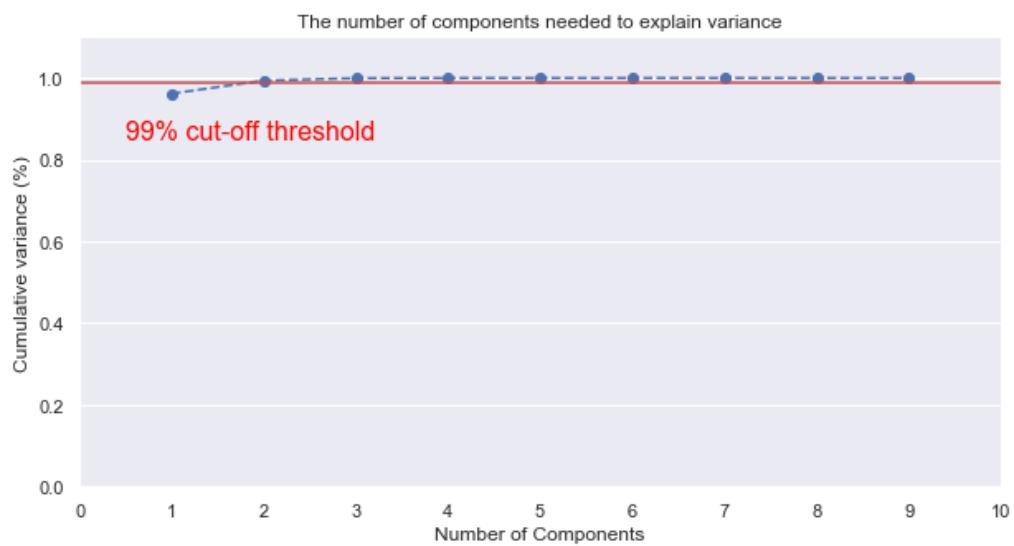


Ilustración 20. Número variables sintéticas que representan el 99% de la varianza para motor. Ilustración propia

Sistema de Frenos

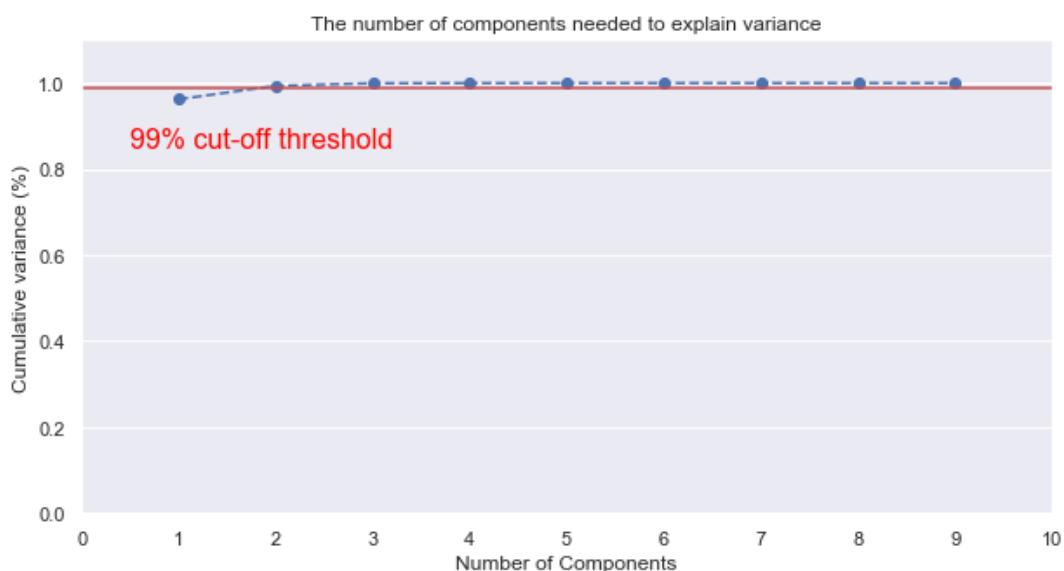


Ilustración 21. Número variables sintéticas que representan el 99% de la varianza para sistema de frenos. Ilustración propia

Sistema de Refrigeración

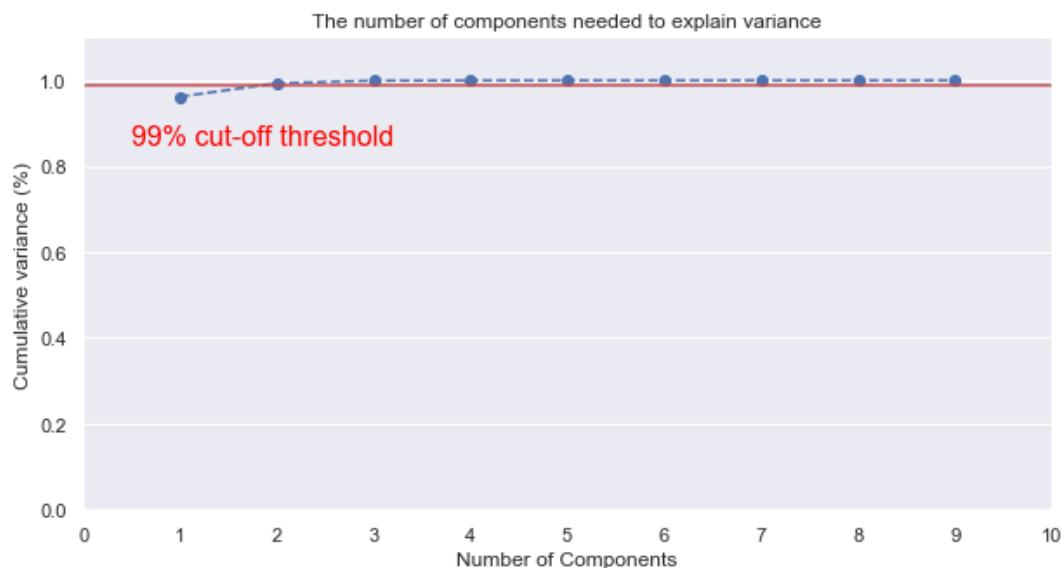


Ilustración 22. Número variables sintéticas que representan el 99% de la varianza para Sistema de refrigeración.
Ilustración propia

Correas

Para el dataset de Correas, obtenemos un ratio de varianza [0.862, 0.101], esto nos indica que en la ilustración 23, tenemos representada con dos dimensiones el 87,2% de la varianza explicada. Explicando la primera dimensión, el eje X el 86.2% de la varianza y el eje Y un 1%.

Viendo la distribución y sabiendo que la posición de las variables nos indica, la varianza explicada, las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar. Por lo que aportarán información similar.

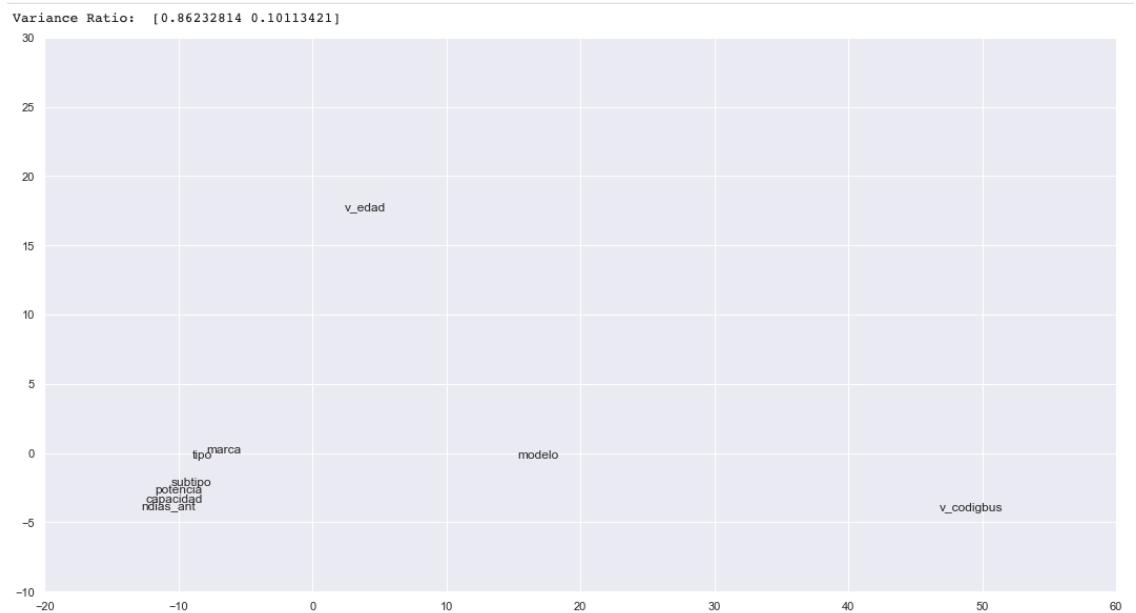


Ilustración 23. Representación gráfica de las dos primeras variables sintéticas para correar. Ilustración propia

Fuga de aire

Para el dataset de fuga de aire obtenemos un ratio de varianza [0.855, 0.111], esto nos indica que en la ilustración 24, tenemos representada con dos dimensiones el 86,6% de la varianza explicada. Explicando la primera dimensión, el eje X el 85,5% de la varianza y el eje Y un 1,1%

Como en el caso anterior, viendo la distribución y sabiendo que la posición de las variables nos indica, la varianza explicada, las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar. Por lo que aportarán información similar.

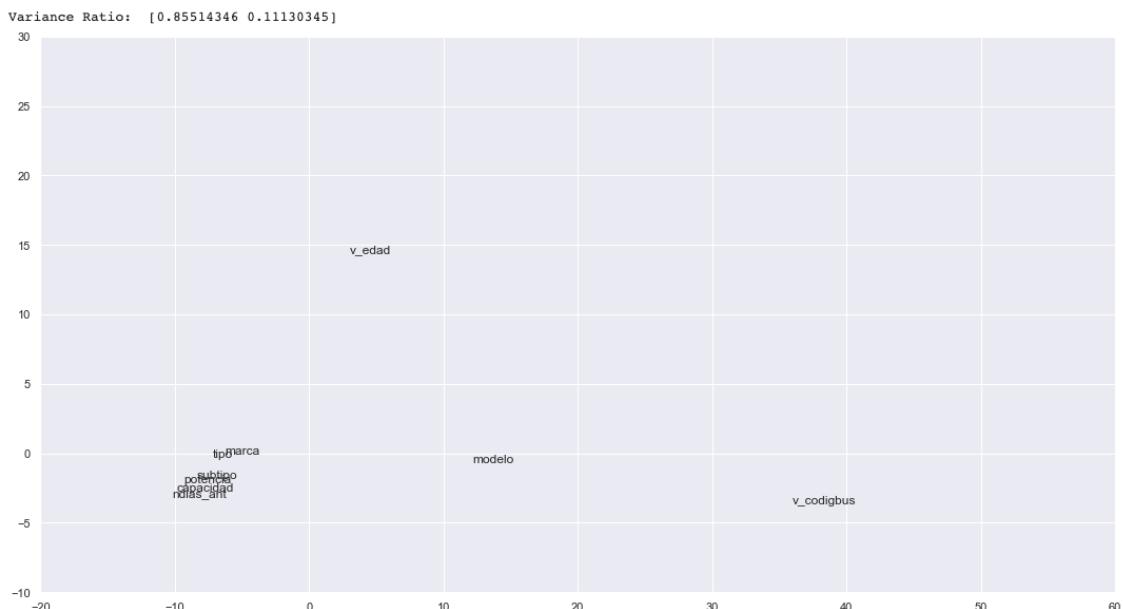


Ilustración 24. Representación gráfica de las dos primeras variables sintéticas para fuga de aire. Ilustración propia

Motor

Para el dataset de motor obtenemos una ratio de varianza [0.850, 0.110], esto nos indica que en la ilustración 25, tenemos representada con dos dimensiones el 86,1% de la varianza explicada. Explicando la primera dimensión, el eje X el 85% de la varianza y el eje Y un 1,1%

Como en el caso anterior, viendo la distribución y sabiendo que la posición de las variables nos indica, la varianza explicada, las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar. Por lo que aportarán información similar.

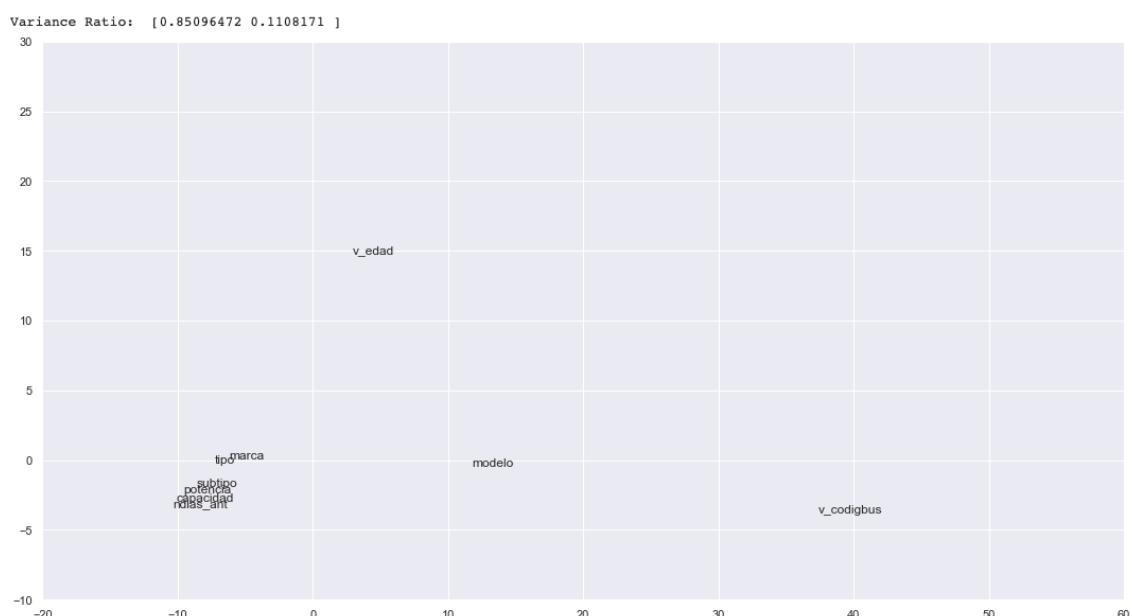


Ilustración 25. Representación gráfica de las dos primeras variables sintéticas para motor. Ilustración propia

Sistema de Frenos

Para el dataset de sistema de frenos obtenemos una ratio de varianza [0.858, 0.104], esto nos indica que en la ilustración 26, tenemos representada con dos dimensiones el 86,8% de la varianza explicada. Explicando la primera dimensión, el eje X el 85,8% de la varianza y el eje Y un 1%

Como en el caso anterior, viendo la distribución y sabiendo que la posición de las variables nos indica, la varianza explicada, las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar. Por lo que aportarán información similar.



Ilustración 26. Representación gráfica de las dos primeras variables sintéticas para sistema de frenos. Ilustración propia

Sistema de Refrigeración

Para el dataset de sistema de refrigeración obtenemos una ratio de varianza [0.848, 0.113], esto nos indica que en la ilustración 27, tenemos representada con dos dimensiones el 85,5% de la varianza explicada. Explicando la primera dimensión, el eje X el 84,8% de la varianza y el eje Y un 1,1%

Como en el caso anterior, viendo la distribución y sabiendo que la posición de las variables nos indica, la varianza explicada, las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar. Por lo que aportarán información similar.



Ilustración 27. Representación gráfica de las dos primeras variables sintéticas para sistema de refrigeración. Ilustración propia

3.5.1. Clustering jerárquico

Distancia entre variables

Obtenemos la distancia entre variables en función de su varianza vemos que las más próximas son **capacidad** y **ndias_ant** en casi todas las subáreas excepto en sistema de refrigeración que se muestra en la ilustración 32, no pudiendo verse una distancia menor a 1, y siendo las más próximas entre sí en el dataset de fuga de aire en torno a 0.6 mostrado en la ilustración 29.

Correas

Para el data set de correas vemos que lo más destacable es la proximidad en función de su varianza, entre **capacidad** y **ndias_ant** en torno a 0.9 como muestra la ilustración 28.

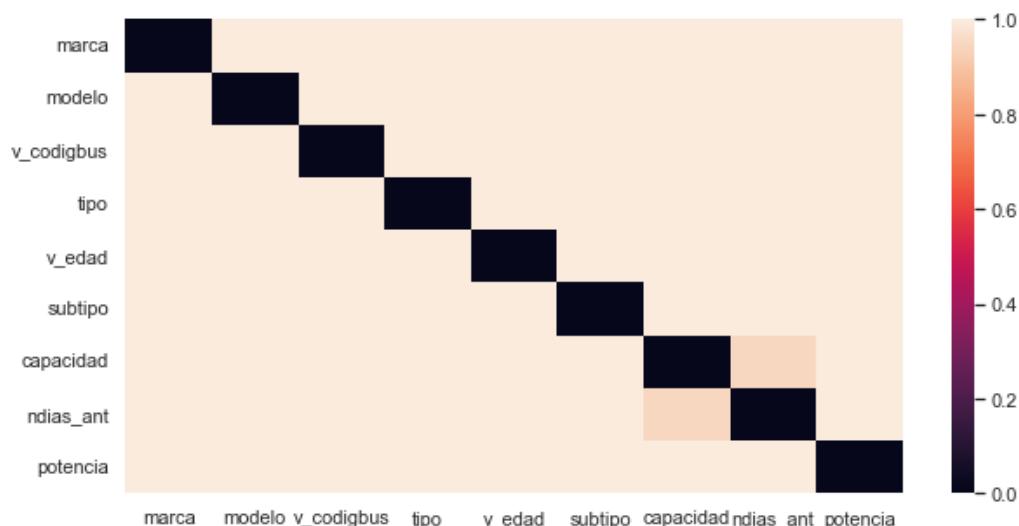


Ilustración 28. Distancia entre variables en función de la varianza explicada en correas. Ilustración propia

Fuga de aire

Para el data set de fuga de aire vemos que lo más destacable es la proximidad en función de su varianza, entre **capacidad** y **ndias_ant** en torno a 0.6 como muestra la ilustración 29.

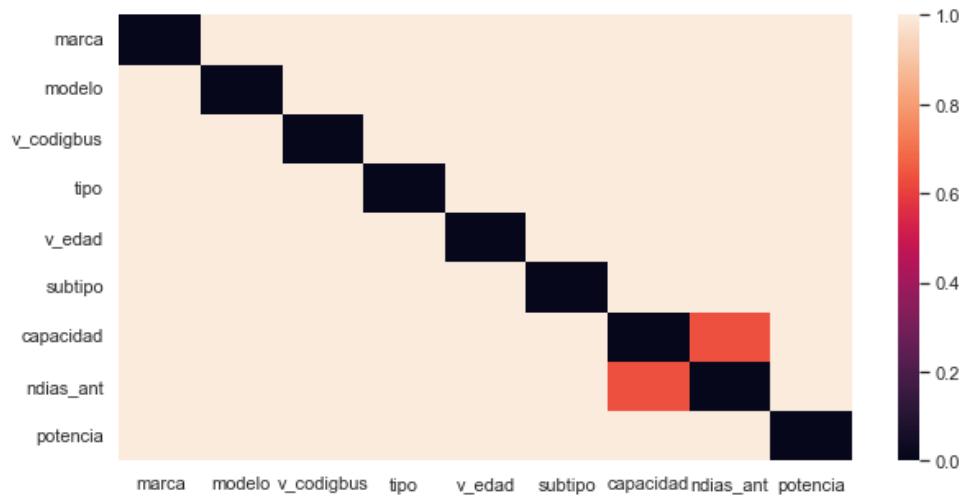


Ilustración 29. Distancia entre variables en función de la varianza explicada en Fuga de aire. Ilustración propia

Motor

Para el data set de motor vemos que lo más destacable es la proximidad en función de su varianza, entre **capacidad** y **ndias_ant** en torno a 0.7 como muestra la ilustración 30.

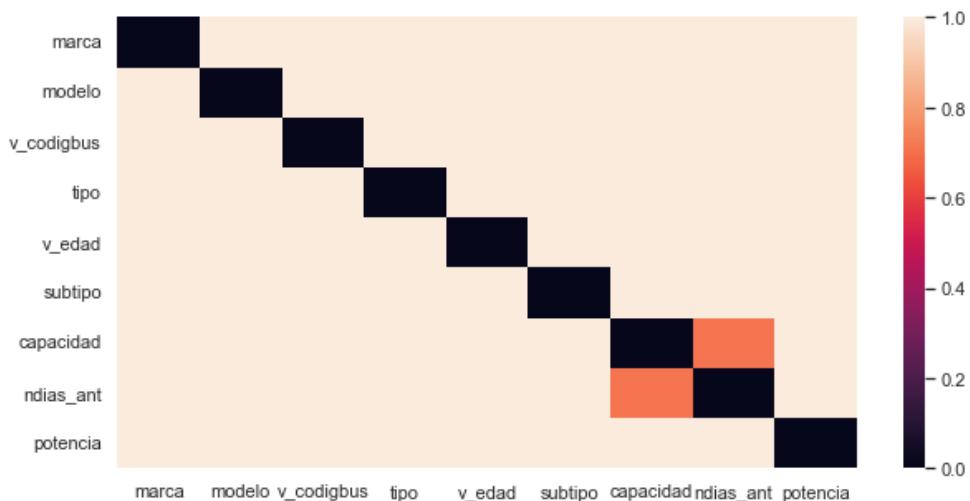


Ilustración 30. Distancia entre variables en función de la varianza explicada en motor. Ilustración propia

Sistema de Frenos

Para el dataset de sistema de frenos vemos que lo más destacable es la proximidad en función de su varianza, entre **capacidad** y **ndias_ant** en torno a 0.9 como muestra la ilustración 31w.

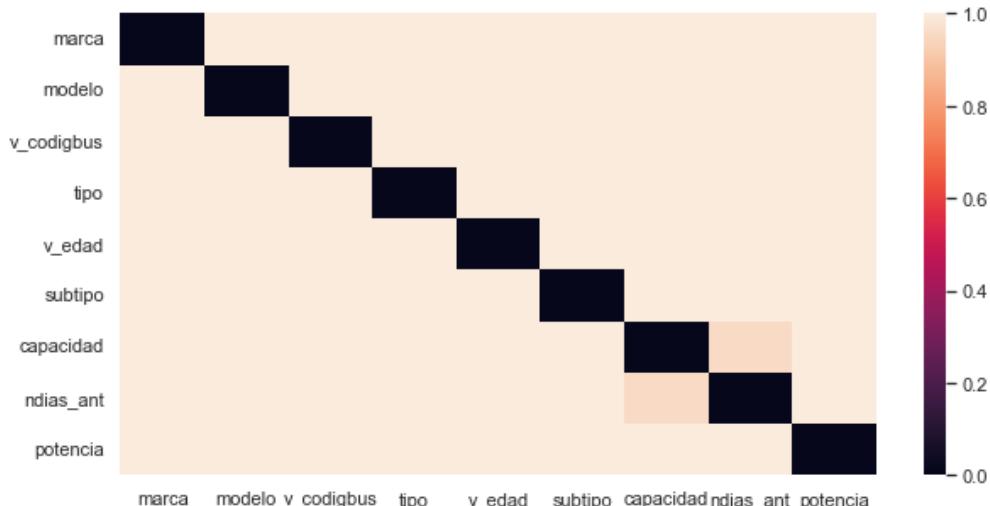


Ilustración 31. Distancia entre variables en función de la varianza explicada en sistema de frenos. Ilustración propia

Sistema de Refrigeración

Para el dataset de sistema de refrigeración no podemos apreciar la proximidad en función de su varianza, entre **capacidad** y **ndias_ant** como muestra la ilustración 32.

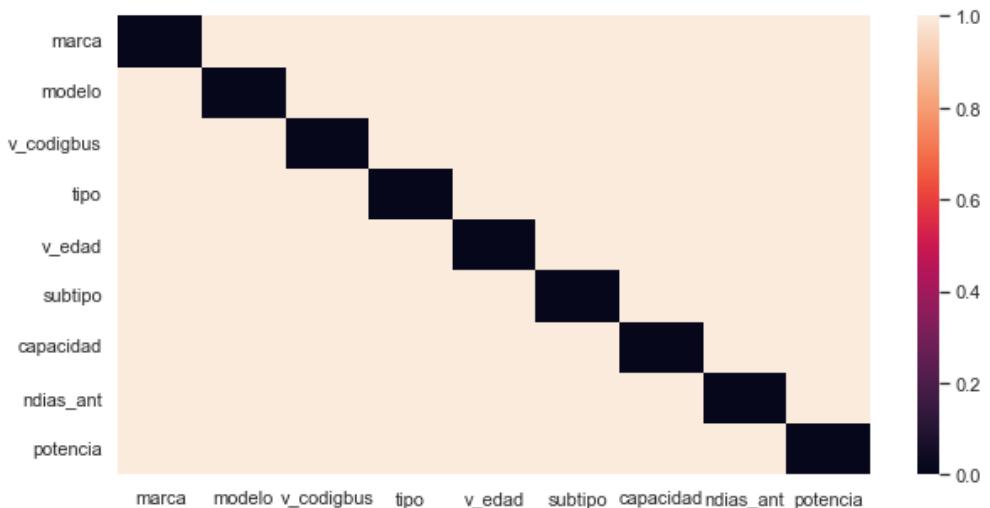


Ilustración 32. Distancia entre variables en función de la varianza explicada en sistemas de refrigeración. Ilustración propia

Dendrograma

El dendrograma es un diagrama en forma de árbol en el cual se muestran la relación jerárquica entre objetos. El uso principal de un dendrograma es encontrar la mejor manera de asignar objetos a grupos. En este caso el dendrograma muestra el agrupamiento jerárquico de las características que se muestran en el diagrama de dispersión de la PCA mostrado en las ilustraciones 23, 24, 25, 26 y 27 correspondiendo respectivamente a correas, fuga de aire, motor, Sistema de frenos y Sistema de refrigeración

Correas

En el dendrograma del dataset de correas, ilustración 33, podemos ver como se agrupan las variables como habíamos visto en el diagrama de dispersión de la ilustración 23 las variables marca y tipo están muy próximas entre sí, y también subtipo, potencia, capacidad, ndias_ant, lo que indica que tienen una varianza muy similar.

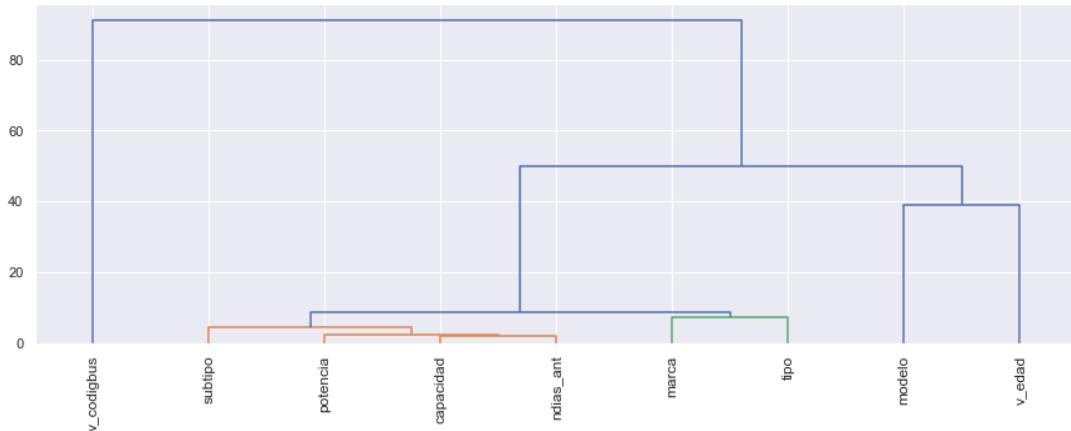


Ilustración 33. Dendrograma características de correas. Ilustración propia

Fuga de aire

En el dendrograma del dataset de fuga de aire, ilustración 34, podemos ver como se agrupan las variables como habíamos visto en el diagrama de dispersión de la ilustración 24 las variables **marca**, **tipo**, **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar

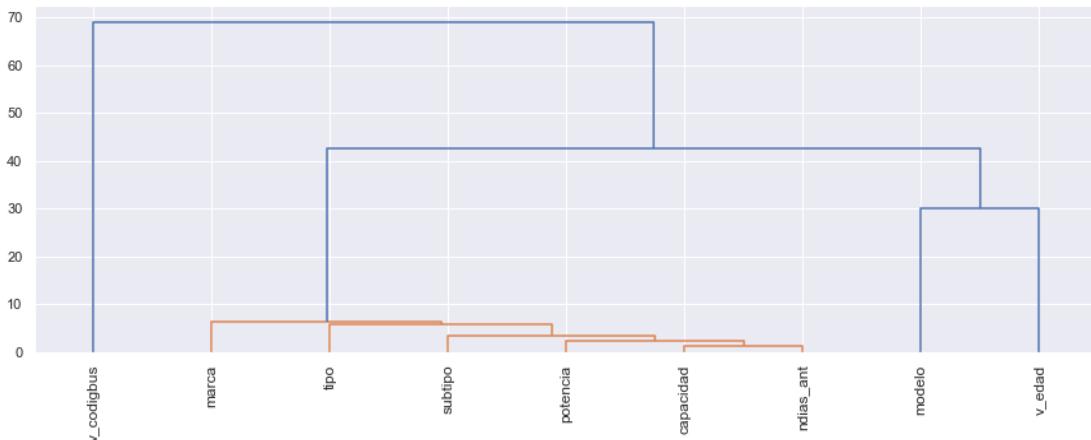


Ilustración 34. Dendrograma características de fuga de aire. Ilustración propia

Motor

Igual que en el dendrograma de correas en el dendrograma del dataset de motor, ilustración 35, podemos ver como se agrupan las variables como habíamos visto en el diagrama de dispersión de la ilustración 25 las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar.

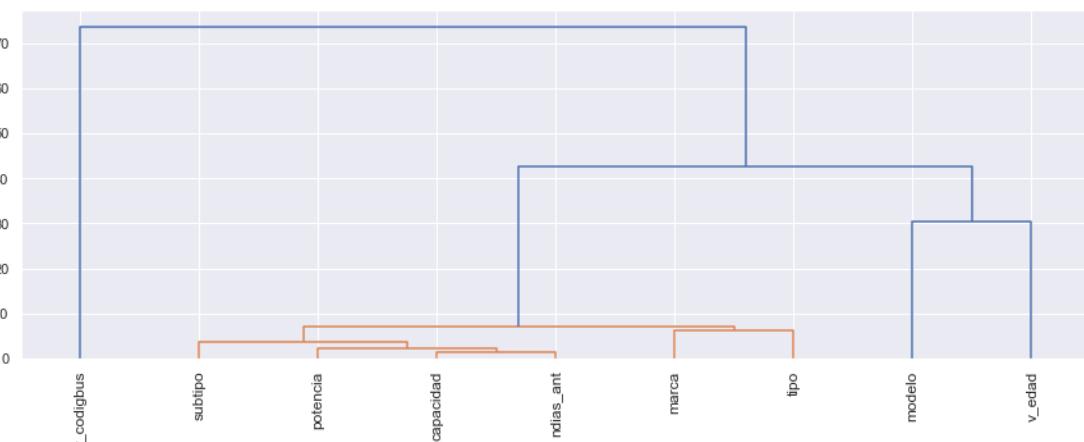


Ilustración 35. Dendrograma características de motor. Ilustración propia

Sistema de Frenos

En el dendrograma del dataset de Sistema de frenos, ilustración 36, podemos ver como se agrupan las variables como habíamos visto en el diagrama de dispersión de la ilustración 26 las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar.

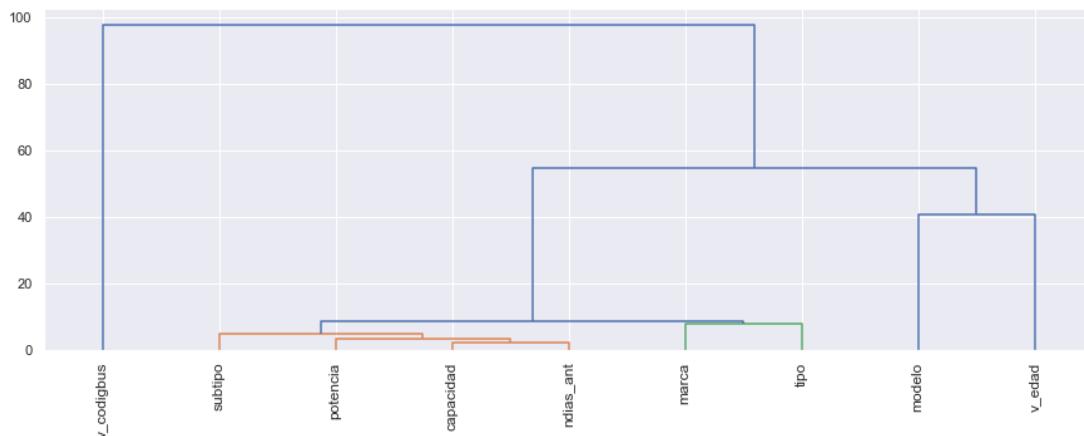


Ilustración 36. Dendrograma características de sistema de frenos. Ilustración propia

Sistema de Refrigeración

En el dendrograma del dataset de Sistema de frenos, ilustración 36, tenemos un resultado muy similar, podemos ver como se agrupan las variables como habíamos visto en el diagrama de dispersión de la ilustración 26 las variables **marca** y **tipo** están muy próximas entre sí, y también **subtipo**, **potencia**, **capacidad**, **ndias_ant**, lo que indica que tienen una varianza muy similar.

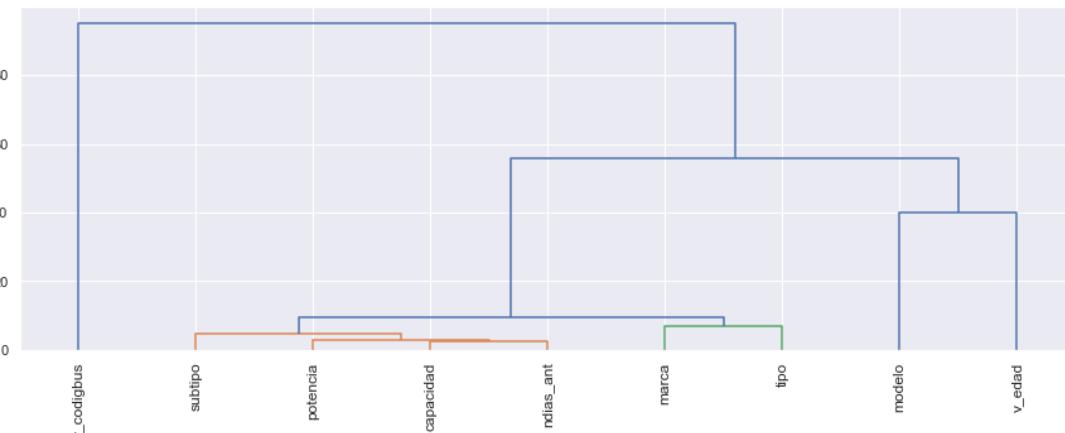


Ilustración 37. Dendrograma características sistema de refrigeración. Ilustración propia

3.6. Reducción de variables PCA

Como se ha explicado previamente el objetivo de la PCA es reducir la dimensionalidad, a costa de la interpretabilidad, utilizando nuevas variables sintéticas no correlacionadas, para describir el conjunto de datos, determinando el orden de las nuevas variables en función de la varianza original que describen.

En este apartado tiene como objeto comprobar si vale la pena reducir la dimensionalidad a razón de sacrificar la interpretabilidad del dataset, aplicando PCA sobre el dataset.

Como podemos ver en las ilustraciones 38, 39, 40, 41 y 42 para explicar el 95% de la varianza necesitamos 8 variables en todos los casos por lo que no compensa perder la interpretabilidad de las variables para ahorrarnos una variable.

Correas

Para el dataset de correas, el cual, se puede ver la ilustración 38, con 8 variables sintéticas se puede explicar algo mas del 95% de la variabilidad, lo que nos indica que no es recomendable perder la interpretabilidad por reducir una variable.

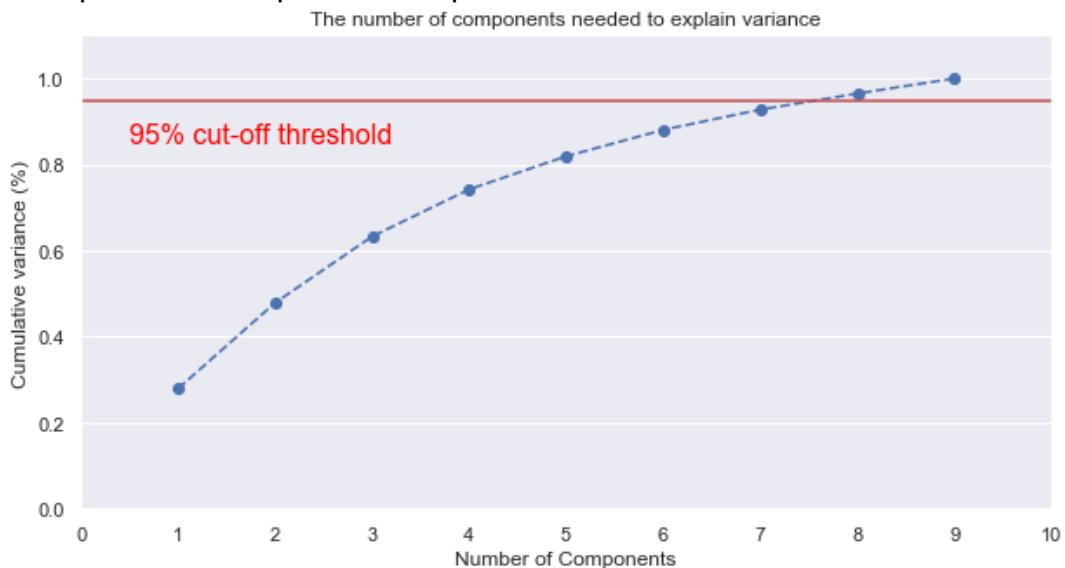


Ilustración 38. Número variables sintéticas que representan el 95% de la varianza para correas. Ilustración propia

Fuga de aire

Para el dataset de fuga de aire, en el cual, se puede ver la ilustración 39, con 8 variables sintéticas se puede explicar algo mas del 95% de la variabilidad, lo que nos indica que no es recomendable perder la interpretabilidad por reducir una variable.

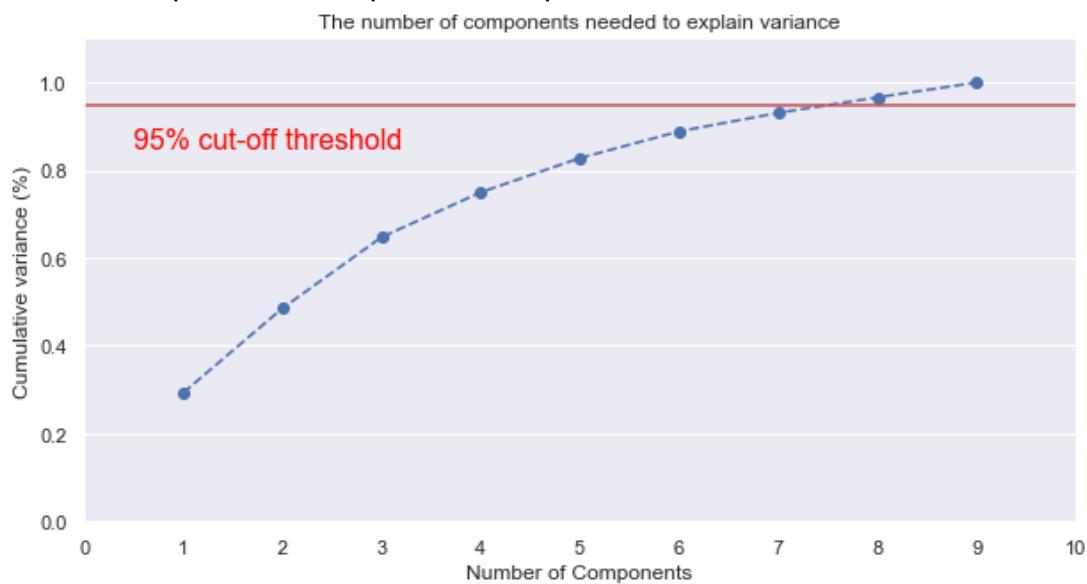


Ilustración 39. Número variables sintéticas que representan el 95% de la varianza para fuga de aire. Ilustración propia

Motor

Para el dataset de motor, en el cual, se puede ver la ilustración 40, con 8 variables sintéticas se puede explicar algo mas del 95% de la variabilidad, lo que nos indica que no es recomendable perder la interpretabilidad por reducir una variable.

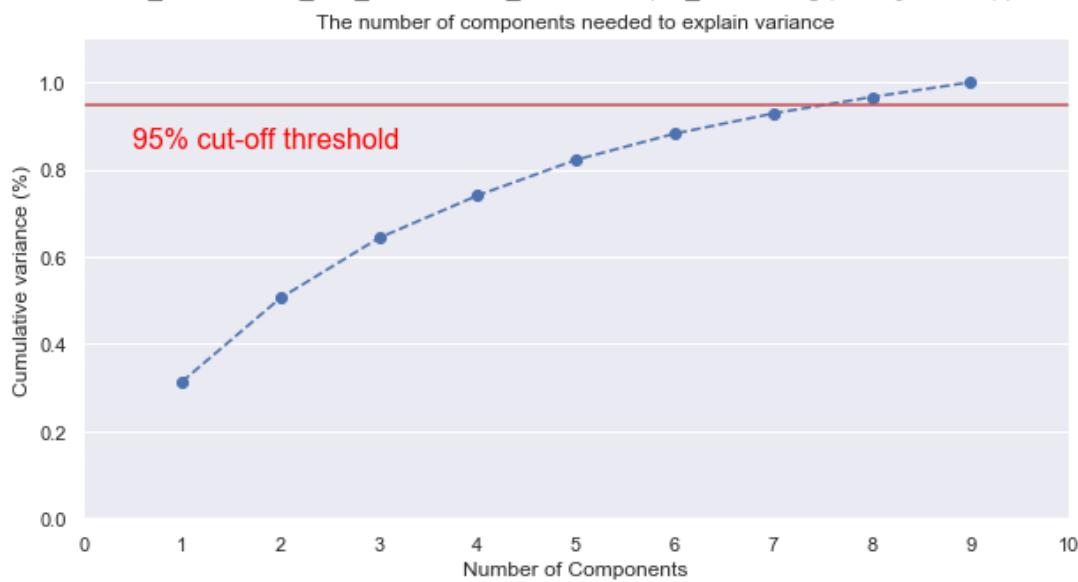


Ilustración 40. Número variables sintéticas que representan el 95% de la varianza para motor. Ilustración propia

Sistema de Frenos

Para el dataset de sistema de frenos, en el cual, se puede ver la ilustración 41, con 8 variables sintéticas se puede explicar algo mas del 95% de la variabilidad, lo que nos indica que no es recomendable perder la interpretabilidad por reducir una variable.

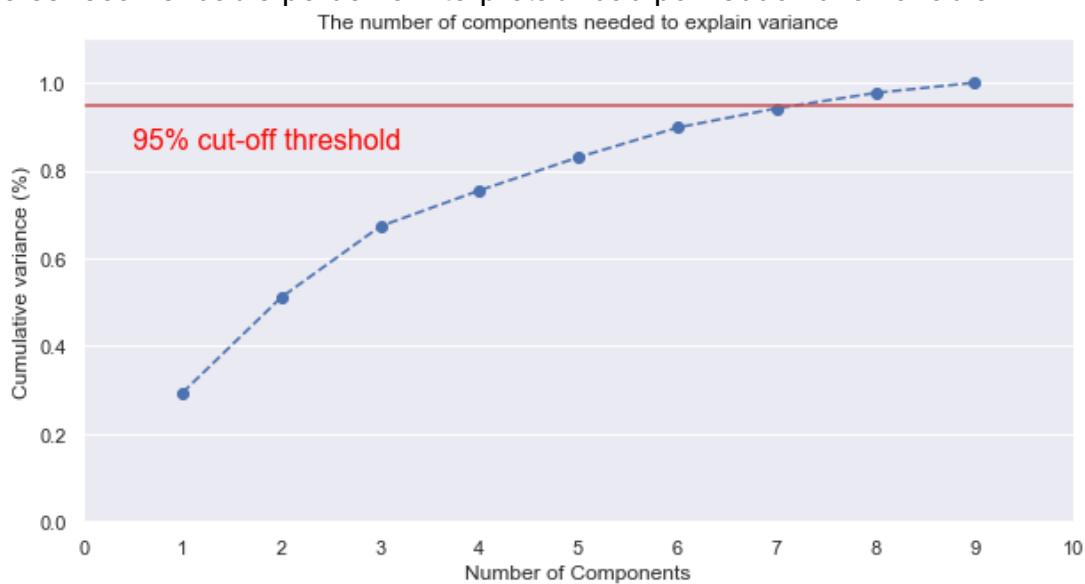


Ilustración 41. Número variables sintéticas que representan el 95% de la varianza para sistema de frenos2. Ilustración propia

Sistema de Refrigeración

Para el dataset de sistema de refrigeración, en el cual, se puede ver la ilustración 42, con 8 variables sintéticas se puede explicar algo más del 95% de la variabilidad, lo que nos indica que no es recomendable perder la interpretabilidad por reducir una variable.

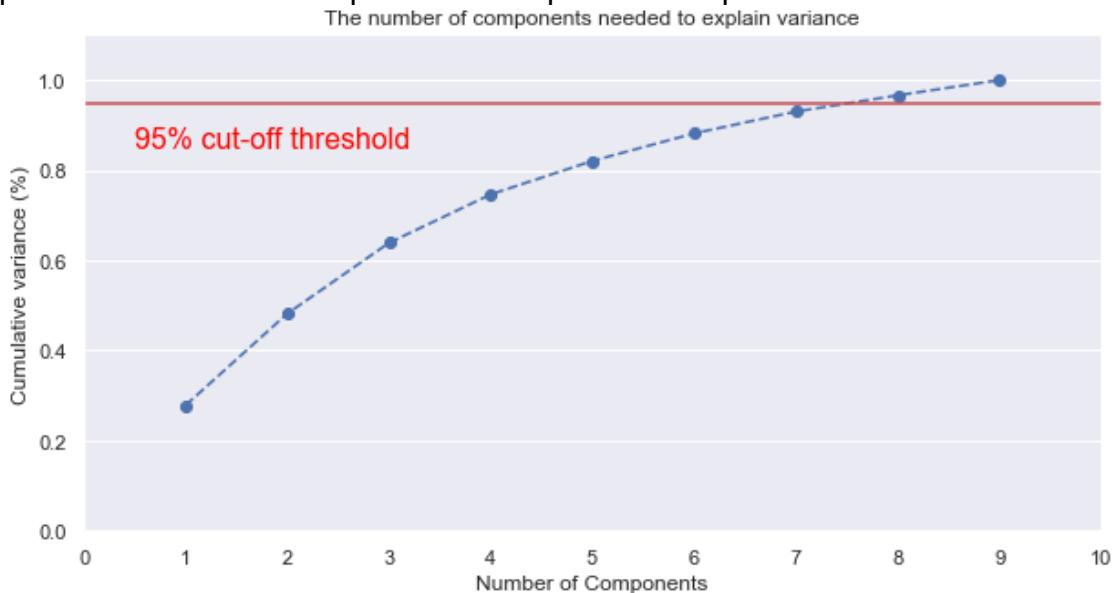


Ilustración 42. Número variables sintéticas que representan el 95% de la varianza para sistema de refrigeración. Ilustración propia

Analizando la distribución del PCA y sabiendo que solo se está mostrando las dos dimensiones con más varianza explicada acumulada, por debajo del 50%, de las 8 actuales,

podemos ver en las ilustraciones 43, 44, 45, 46 y 47 que una parte de los datos esta separada de otra, lo que indica que al menos hay dos grupos diferenciados que forman clusters, se a probado a nombrar cada punto con distintas columnas de características, obteniendo como resultado que los autobuses del subtipo 1 'Normal' forman un clúster y el resto otro.

Correas

Como podemos ver en la ilustración 43. Para el dataset correas vemos que los autobuses de subtipo 1, normal, se distribuyen en función de las dos dimensiones con más varianza explicada, como un conjunto formando clúster y el resto como otro conjunto formando otro clúster, también vemos que el subtipo 5 VTC también se distribuyen formando un conjunto separado, peor con muy pocas observaciones lo que nos indica que con más observaciones se podrían analizar como un tercer clúster.

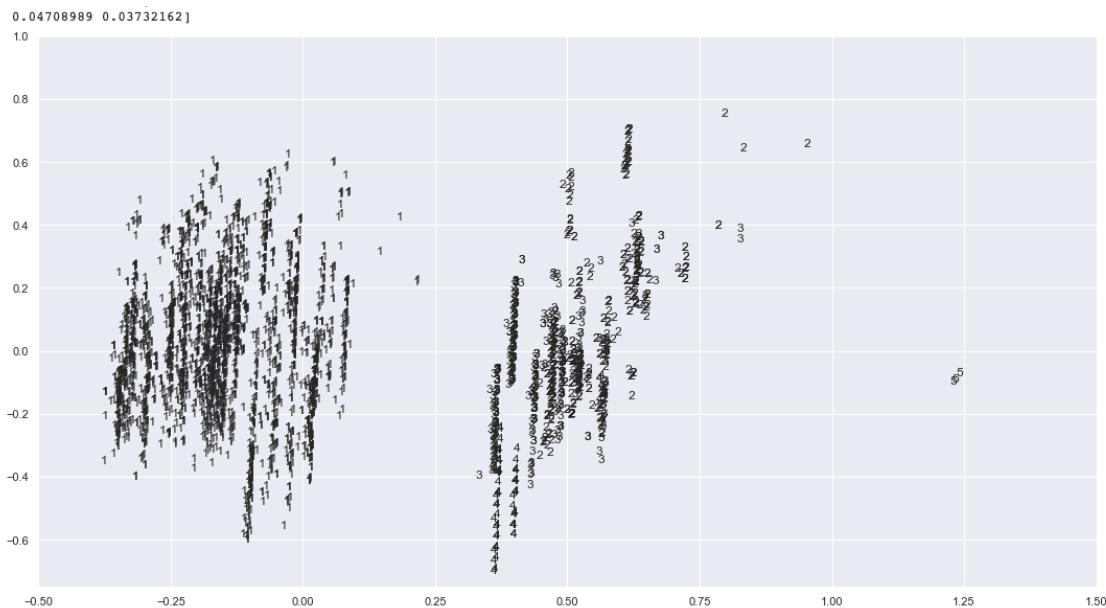


Ilustración 43. Distribución PCA para correas. Ilustración propia

Fuga de aire

Como podemos ver en la ilustración 44. Para el dataset fuga de aire vemos que los autobuses de subtipo 1, normal se distribuyen en función de las dos dimensiones con más varianza explicada, como un conjunto formando clúster y el resto como otro conjunto formando otro clúster, también vemos que el subtipo 5 VTC también se distribuyen formando un conjunto separado, pero con muy pocas. También podemos observar que no se distribuyen exactamente igual, que el dataset de correas, pero si muy parecido.

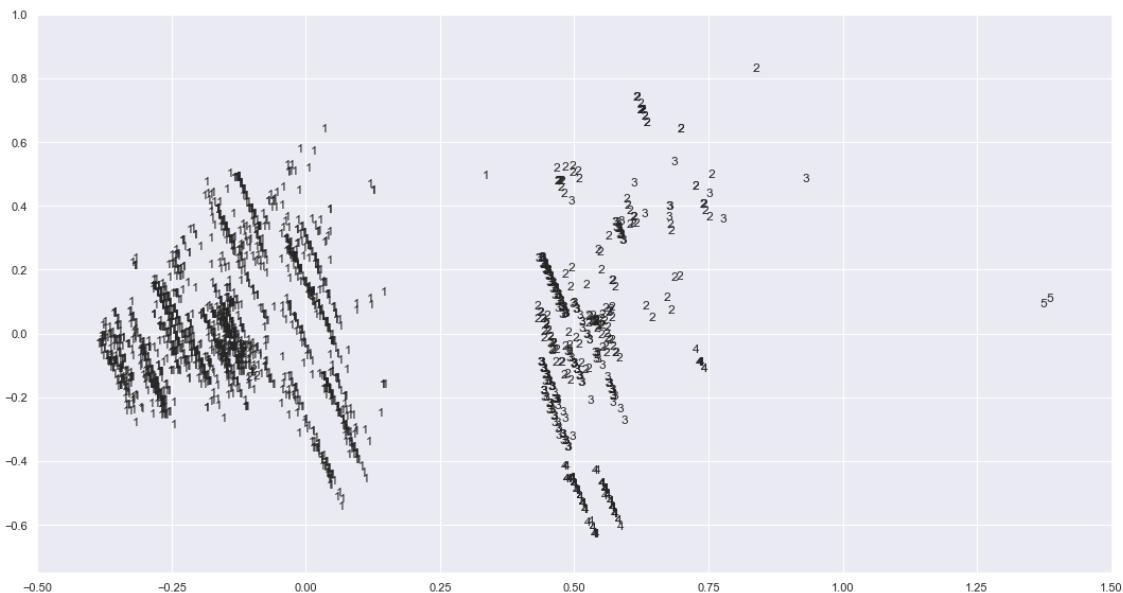


Ilustración 44. Distribución PCA para fuga de aire. Ilustración propia

Motor

Como podemos ver en la ilustración 45. Para el dataset motor vemos que los autobuses de subtipo 1, normal se distribuyen en función de las dos dimensiones con más varianza explicada, como un conjunto formando clúster y el resto como otro conjunto formando otro clúster, volviendo a tener el subtipo 5 VTC distribuido formando un conjunto separado, pero con muy pocas observaciones. También podemos observar que no se distribuyen exactamente igual, que el datasets anteriores, pero si muy parecido. Se observa que hay algunos autobuses del subtipo 1 que están en medio de los dos clústeres.

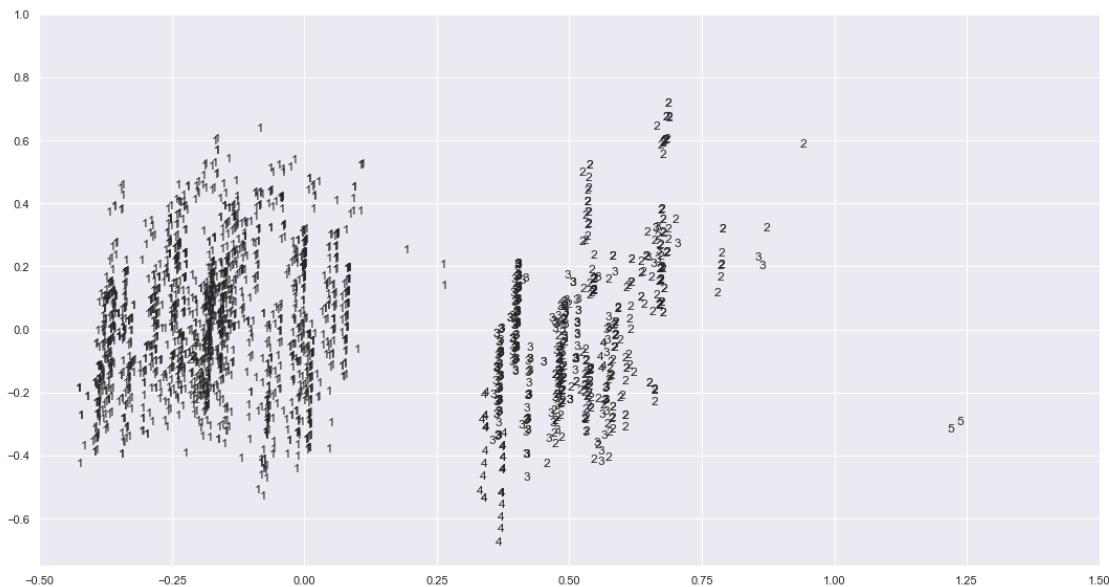


Ilustración 45. Distribución PCA para motor. Ilustración propia

Sistema de Frenos

Como podemos ver en la ilustración 46. Para el dataset sistema de frenos vemos que los autobuses de subtipo 1, normal se distribuyen en función de las dos dimensiones con más varianza explicada, como un conjunto formando clúster aunque en este caso si se mezclan más con el segundo clúster, también vemos nuevamente que el subtipo 5 VTC también se distribuyen formando un conjunto separado, pero con muy pocas observaciones. podemos observar que no se distribuyen exactamente igual, que el datasets anteriores, pero si con la misma tendencia.

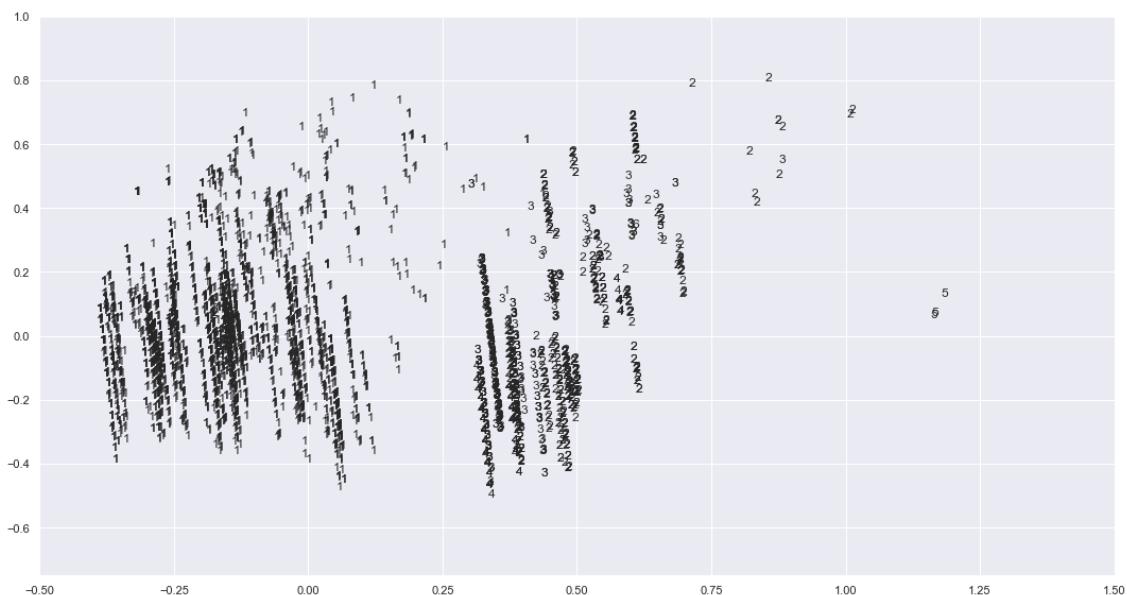


Ilustración 46. Distribución PCA para sistema de frenos. Ilustración propia

Sistema de Refrigeración

Como podemos ver en la ilustración 47. Para el dataset motor vemos que los autobuses de subtipo 1, normal se distribuyen en función de las dos dimensiones con más varianza explicada, como un conjunto formando clúster y el resto como otro conjunto formando otro clúster, volviendo a tener el subtipo 5 VTC distribuido formando un conjunto separado, pero con muy pocas observaciones. También podemos observar que no se distribuyen exactamente igual, que el datasets anteriores, pero si muy parecido. Se observa que hay algunos autobuses del subtipo 1 que están en medio de los dos clusters, pero no se llegan a mezclar quedando los suficientemente separados.

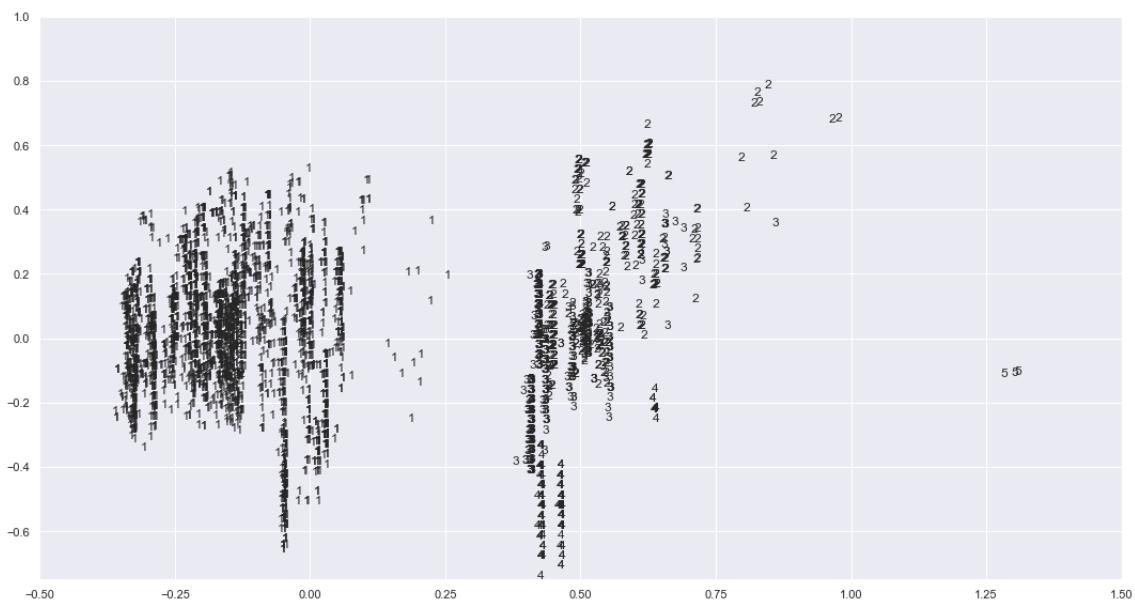


Ilustración 47. Distribución PCA para sistema de refrigeración. Ilustración propia

3.7. Clusterización mediante Kmeans

La agrupación en clusters de Kmeans sirve para dividir un conjunto de datos en K clusters distintos que no se superponen. Para realizar el agrupamiento de Kmeans, se debe especificar el número deseado de agrupamientos K, luego, el algoritmo de Kmeans asignará cada observación exactamente a uno de los K conjuntos de datos.

Ya que, para algunos datasets, no se puede ver claramente la separación y pudiendo haber algún dato de subtipo que no se vea bien al estar muy juntos los datos, vamos a utilizar Kmeans para clusterizar en lugar de usar el subtipo.

Se prueba a implementar Kmeans con distinta cantidad de clusters, verificando que con 2 clusters se obtiene el mejor valor de silueta, que es una medida para cuantificar cuan similar es un objeto a su propio grupo en comparación con otros grupos, por lo tanto, nos indica cual es la mejor opción para clusterizar, además existe una concordancia en los clusters que se han observado en la distribución PCA formados por subtipo. Este hecho coincide en todos los datasets. cómo se mostrará en las ilustraciones posteriores.

Correas

En la ilustración 48 podemos los clusters para el dataset de correas y que el valor de silueta para dos clusters es el mas alto.

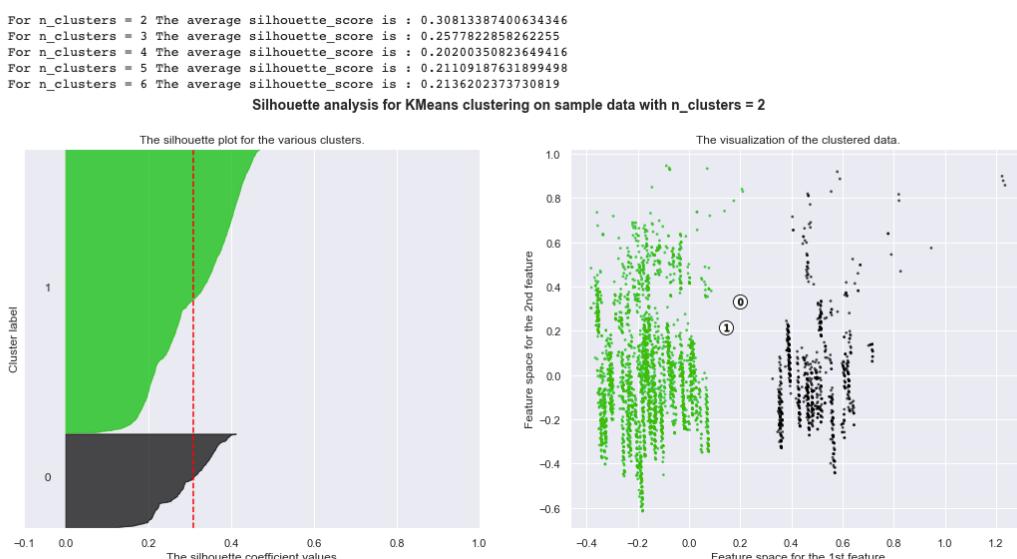


Ilustración 48. Clusters para correas. Ilustración propia

Fuga de aire

En la ilustración 49 podemos los clusters para el dataset de fuga de aire y que el valor de silueta para dos clusters es el mas alto.

```
For n_clusters = 2 The average silhouette_score is : 0.33761035541909395
For n_clusters = 3 The average silhouette_score is : 0.2441292555958088
For n_clusters = 4 The average silhouette_score is : 0.27273726420964195
For n_clusters = 5 The average silhouette_score is : 0.2795532397087194
For n_clusters = 6 The average silhouette_score is : 0.29482311200079453
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
```

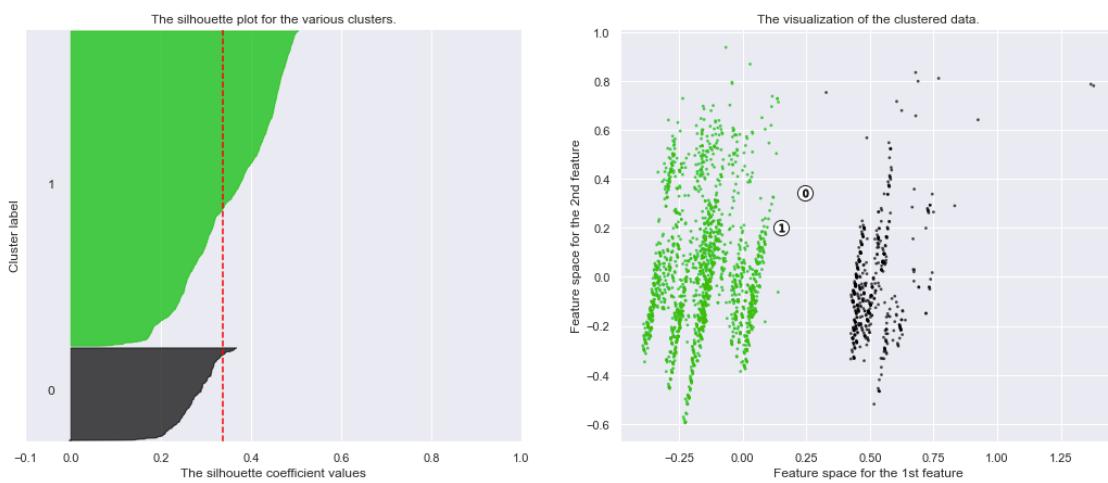


Ilustración 49. Clusters para fuga de aire. Ilustración propia

Motor

En la ilustración 50 podemos los clusters para el dataset de motor y que el valor de silueta para dos clusters es el mas alto.

```
For n_clusters = 2 The average silhouette_score is : 0.32176430522753136
For n_clusters = 3 The average silhouette_score is : 0.2659729319437846
For n_clusters = 4 The average silhouette_score is : 0.27855625303992204
For n_clusters = 5 The average silhouette_score is : 0.2767314348268367
For n_clusters = 6 The average silhouette_score is : 0.1956323993613581
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2
```

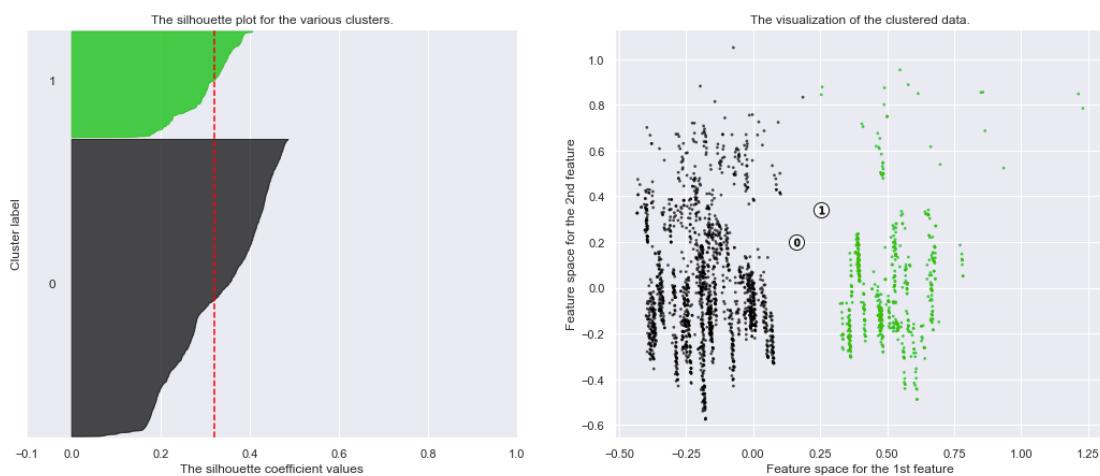


Ilustración 50. Clusters para motor. Ilustración propia

Sistema de Frenos

En la ilustración 51 podemos los clusters para el dataset de sistema de frenos y que el valor de silueta para dos clusters es el más alto.

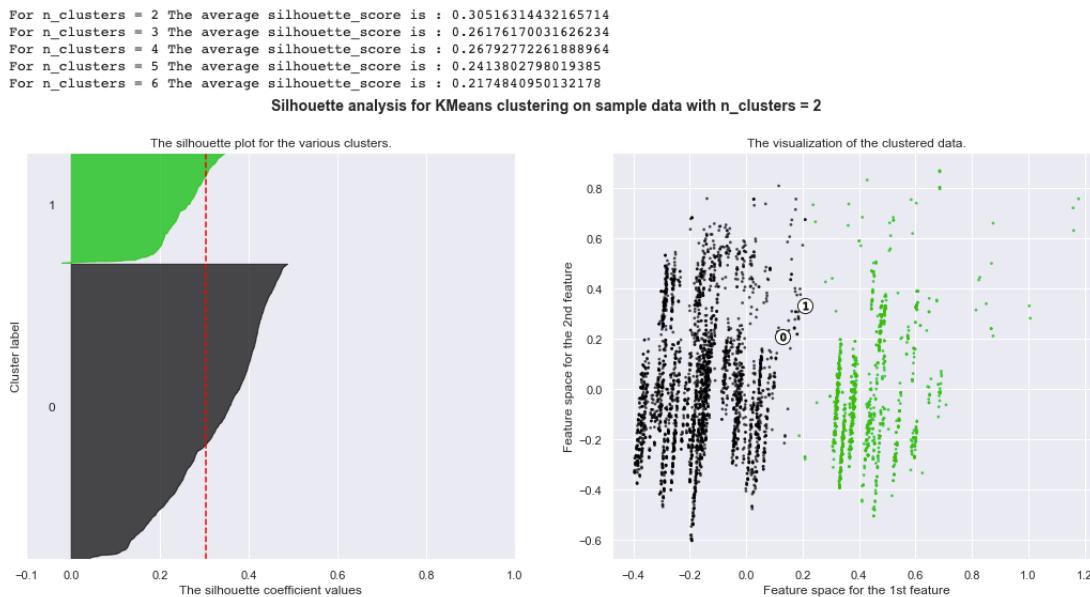


Ilustración 51. Clusters para sistema de frenos. Ilustración propia

Sistema de Refrigeración

En la ilustración 52 podemos los clusters para el dataset de sistema de refrigeración y que el valor de silueta para dos clusters es el mas alto.

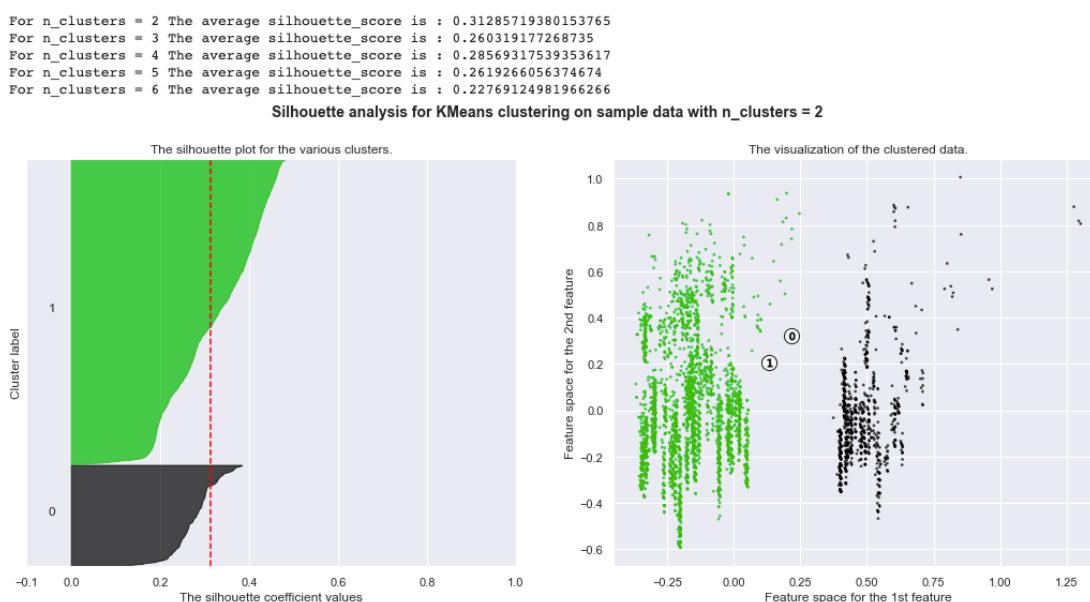


Ilustración 52. Clusters para sistema de refrigeración. Ilustración propia

Añadimos una columna al dataset, **kmeans_cluster**, en la que añadiremos el valor de, a que clúster pertenece, 0 par el primer clúster y 1 para el segundo clúster.

Al analizar nuevamente las correlaciones comprobamos que la clusterización esta fuertemente relacionada con el subtipo, por encima del 90% en todos los casos, lo que confirma los indicios que teníamos de que los valores se agrupan por subtipos en la distribución del PCA. Nuevamente podemos ver estas correlaciones en las ilustraciones posteriores

Correas

Para el dataset de correas vemos que en la ilustración 53 que se correlaciona con subtipo en un 91% también se ve que hay un 79% de correlación con la potencia, lo que nos indica que la potencia también podría generar clúster en la distribución.



Ilustración 53. Correlaciones con kmeans para correas. Ilustración propia

Fuga de aire

Para el dataset de fuga de aire vemos que en la ilustración 54 que se correlaciona con subtipo en un 92% también se ve que hay un 75% de correlación con la potencia, lo que nos indica que la potencia también podría generar clúster en la distribución.



Ilustración 54. Correlaciones con kmeans para fuga de aire. Ilustración propia

Motor

Para el dataset de motor vemos que en la ilustración 55 que se correlaciona con subtipo en un 91% también se ve que hay un 79% de correlación con la potencia, lo que nos indica que la potencia también podría generar clúster en la distribución.



Ilustración 55. Correlaciones con kmeans para motor. Ilustración propia

Sistema de Frenos

Para el dataset de sistema de frenos vemos que en la ilustración 56 que se correlaciona con subtipo en un 90% también se ve que hay un 82% de correlación con la potencia que es la mas fuertes de todos los datasets, lo que nos indica que la potencia también podría generar clúster en la distribución.



Ilustración 56. Correlaciones con kmeans para sistema de frenos. Ilustración propia

Sistema de Refrigeración

Para el dataset de sistema de frenos vemos que en la ilustración 56 que se correlaciona con subtipo en un 91% también se ve que hay un 77% de correlación con la potencia, lo que nos indica que la potencia también podría generar clúster en la distribución.



Ilustración 57. Correlaciones con kmeans para sistema de refrigeración. Ilustración propia

3.8. Modelado Random Forest

Después de varias pruebas se ha visto que el mejor modelo para realizar este tipo de predicción es el Random Forest, obteniendo una precisión similar a SVC (clasificador de vectores de soporte), que tiene como objetivo ajustarse a los datos, devolviendo el hiperplano que "mejor ajuste", que divide o categoriza los datos, el cual tiene un alto coste computacional.

Random Forest es un tipo de modelo que combina diversos árboles de decisión y la salida de cada uno se contará como "un voto" y la opción más votada será la respuesta del Random Forest, al utilizar múltiples árboles se reduce considerablemente el riesgo de overfitting, normalmente da buenos resultados en problemas de clasificación, aunque no funciona tan bien con pocos datos.

El motivo por el que se opta por este tipo de modelo es, que es más simple y no difiere en exceso en resultados, mejorando así el coste computacional.

Además, se ha probado con diferentes modelos que se explicaran brevemente a continuación ya que se han obtenido peores resultados.

Árbol de decisión, DecisionTreeClassifier, destinado a la clasificación, es una estructura de árbol similar a un diagrama de flujo, donde cada nodo interno aplica una prueba en un atributo, cada rama representa un resultado de la prueba y cada nodo hoja (nodo terminal) tiene una etiqueta de clase.

K- vecinos, KNeighborsClassifier sirve esencialmente para clasificar valores buscando los puntos de datos "más similares" (por cercanía) aprendidos en la etapa de entrenamiento.

Se decide realizar un modelo Random Forest para ambos clusters y para cada dataset.

Para encontrar los hiperparámetros se ha usado RandomizedSearchCV y GridSearchCV, que son algoritmos para muestrear de manera efectiva el espacio de búsqueda y encontrar una buena solución. Inicialmente y utilizando finalmente RandomizedSearchCV ya que con un número considerable de iteraciones se consigue valores muy similares a los de

GridSearchCV que, a pesar de ser mucho más exhaustivo, el coste computacional es mucho mayor por lo que se ha utilizado en espacios de búsqueda menores acotados mediante el RandomizedSearchCV. Además de que se ha implementado cross validation que tiene dos pasos principales, dividir los datos en subconjuntos y rotar el entrenamiento y la validación entre ellos, en los algoritmos anteriormente mencionados ya viene integrado, utilizando 5 particiones.

Cuando entrenamos un modelo, dividimos el conjunto de datos en dos conjuntos principales: entrenamiento y prueba. El conjunto de entrenamiento representa todos los ejemplos de los que un modelo está aprendiendo, mientras que el conjunto de prueba simula los ejemplos de prueba.

Para la selección de las características que deseamos utilizar para cada clúster, después de ejecutar el modelo se ha observado que hay características que nos son relevantes, ya que la importancia es 0, por lo que se han seleccionado las características que tuvieran importancia superior a 0, y se ha repetido la ejecución del modelo. A continuación, se detalla cuales se han seleccionado para cada dataset.

Correas

Para el Cluster0 se han seleccionado: marca, modelo, v_codigbus, tipo, v_edad, potencia y subtipo.

Para el Cluster1 se han seleccionado: marca, modelo, v_codigbus, tipo, v_edad y subtipo.

Fuga de aire

Para el Cluster0 se han seleccionado: marca, modelo, v_codigbus, tipo y v_edad.

Para el Cluster1 se han seleccionado: marca, modelo, v_codigbus, tipo, potencia, v_edad y subtipo.

Motor

Para el Cluster0 se han seleccionado: marca, modelo, v_codigbus, tipo, potencia, v_edad y subtipo

Para el Cluster1 se han seleccionado: marca, modelo, v_codigbus, tipo, v_edad y subtipo.

Sistema de frenos

Para el Cluster0 se han seleccionado: marca, modelo, v_codigbus, tipo, v_edad y subtipo

Para el Cluster1 se han seleccionado: marca, modelo, v_codigbus, tipo, potencia, v_edad y subtipo.

Sistema refrigeración

Para el Cluster0 se han seleccionado: marca, modelo, v_codigbus, tipo, potencia, v_edad y subtipo.

Para el Cluster1 se han seleccionado: marca, modelo, v_codigbus, tipo, v_edad y subtipo.

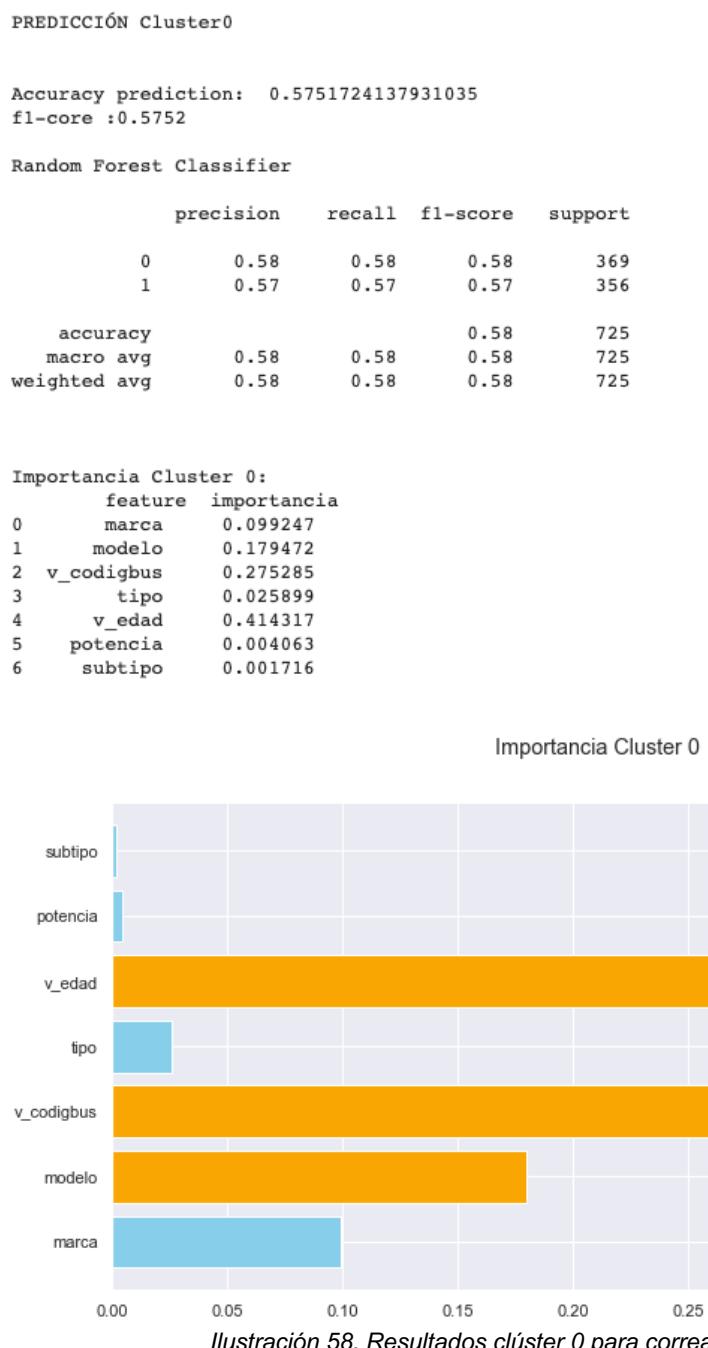
4. Resultados

En el presente apartados se mostrarán los resultados obtenidos para cada una de las diferentes subáreas. los resultados combinados se han obtenido de multiplicar el valor obtenido por un factor de ponderación asignado por el tamaño del clúster. En cada caso será distinto, aunque similares.

4.1. Correas

Como vemos en la ilustración 58 para el clúster 0 (subtipo ‘Normal’) de correas vemos que el modelo es capaz de predecir en un 57 % que el autobús sufre una avería por correas y un 58% de que no es por correas

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo



Como vemos en la Ilustración 59, para el clúster 1 de correas vemos que el modelo es capaz de predecir en un 59 % que el autobús sufre una avería por correas y un 55% de que no es por correas.

Y que las variables más influyentes para determinarlo como en el clúster 0 son, la edad que tiene más importancia que en el caso anterior, el código del autobús y el modelo

PREDICCIÓN Cluster 1

```
Accuracy prediction: 0.5708333333333333
f1-core :0.5708
```

```
Random Forest Classifier
```

	precision	recall	f1-score	support
0	0.56	0.55	0.55	116
1	0.58	0.59	0.59	124
accuracy			0.57	240
macro avg	0.57	0.57	0.57	240
weighted avg	0.57	0.57	0.57	240

Importancia Cluster 1:		
	features	importancia
0	marca	0.055957
1	modelo	0.106844
2	v_codigbus	0.179632
3	tipo	0.010106
4	v_edad	0.622246
5	subtipo	0.025216

Importancia Cluster 1

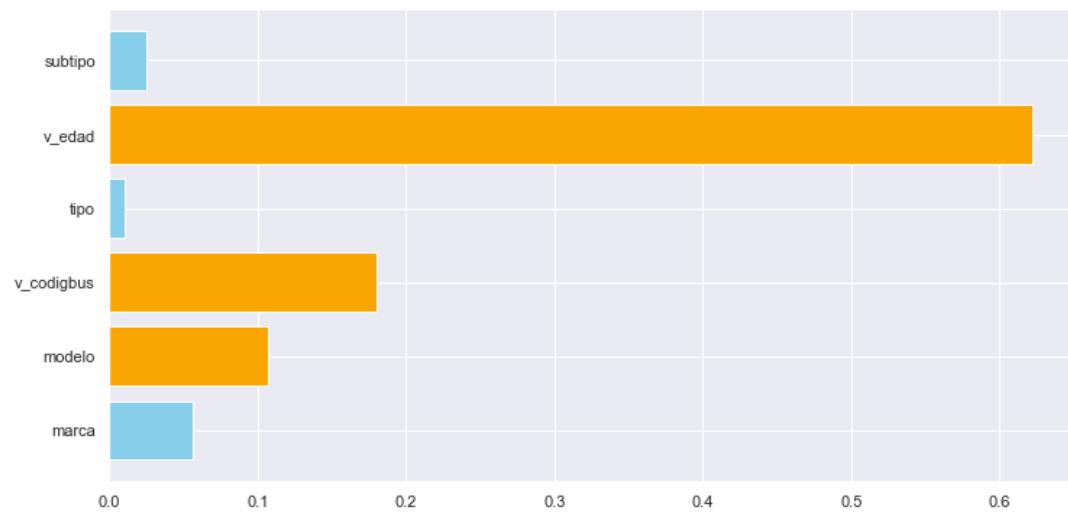


Ilustración 59. Resultados clúster 1 para correas. Ilustración propia

Como se puede ver en la ilustración 60, para el combinando, los resultados, ponderándolos, multiplicando por un factor determinado por el tamaño de cada clúster vemos que el modelo es capaz de predecir en un 57 % que el autobús sufre una avería por correas y un 58% de que no es por correas.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo

PREDICCIÓN Combinada

```
Accuracy prediction: 0.5740932642487047
f1-score :0.5741
```

```
Random Forest Classifier
```

	precision	recall	f1-score	support
0	0.58	0.58	0.58	485
1	0.57	0.57	0.57	480
accuracy			0.57	965
macro avg	0.57	0.57	0.57	965
weighted avg	0.57	0.57	0.57	965

Importancia:

	features	importancia
feature		
marca	marca	0.088425
modelo	modelo	0.161315
potencia	potencia	0.003047
subtipo	subtipo	0.007591
tipo	tipo	0.021951
v_codigbus	v_codigbus	0.251372
v_edad	v_edad	0.466300

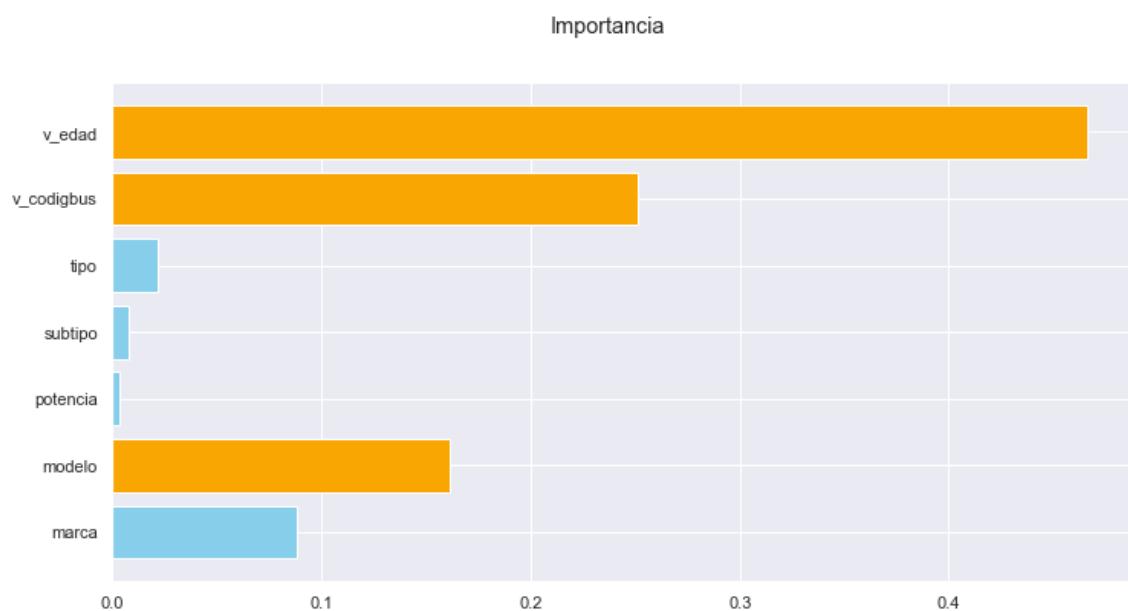


Ilustración 60. Resultados combinados para correas. Ilustración propia

4.2. Fuga de aire

Como se puede ver en la ilustración 61, para el clúster 0 de fuga de aire vemos que el modelo es capaz de predecir en un 72 % que el autobús sufre una avería por correas y un 73% de que no es por fuga de aire.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, la marca y el subtipo

PREDICCIÓN Cluster0

```
Accuracy prediction: 0.7238805970149254
f1-score :0.7239
```

Random Forest Classifier

	precision	recall	f1-score	support
0	0.77	0.69	0.73	72
1	0.68	0.76	0.72	62
accuracy			0.72	134
macro avg	0.73	0.73	0.72	134
weighted avg	0.73	0.72	0.72	134

Importancia Cluster 0:

	feature	importancia
0	marca	0.116865
1	modelo	0.140025
2	v_codigbus	0.166931
3	tipo	0.032668
4	v_edad	0.402967
5	subtipo	0.140546

Importancia Cluster 0

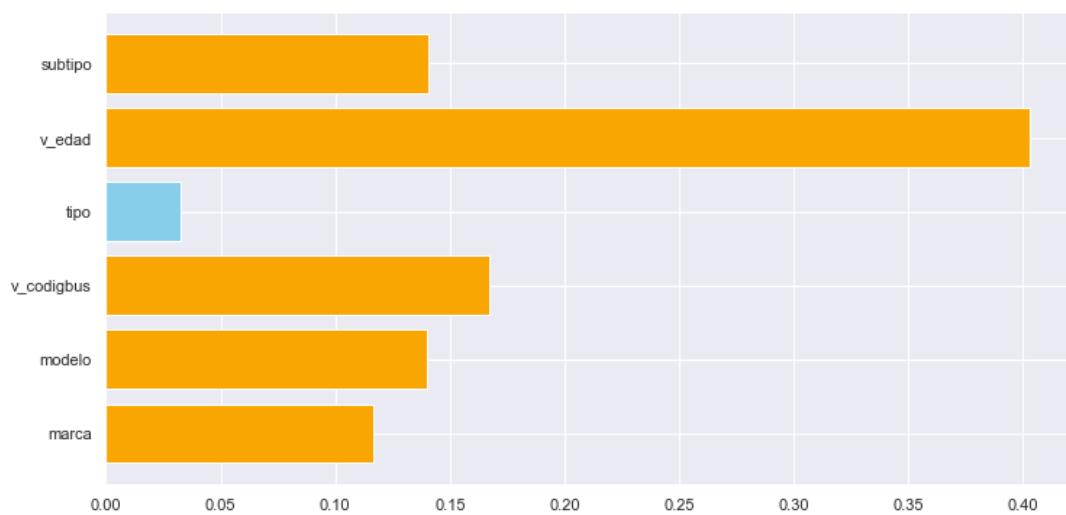


Ilustración 61.Resultados clúster 0 para fuga de aire. Ilustración propia

Como se puede ver en la ilustración 62, para el clúster 1 de fuga de aire vemos que el modelo es capaz de predecir en un 59 % que el autobús sufre una avería por correas y un 59% de que no es por fuga de aire.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, influyendo algo la marca también.

PREDICCIÓN Cluster 1

```
Accuracy prediction: 0.5859030837004405
f1-score :0.5859
```

Random Forest Classifier

	precision	recall	f1-score	support
0	0.60	0.59	0.59	233
1	0.57	0.58	0.58	221
accuracy			0.59	454
macro avg	0.59	0.59	0.59	454
weighted avg	0.59	0.59	0.59	454

Importancia Cluster 1:		
	features	importancia
0	marca	0.076465
1	modelo	0.176243
2	v_codigbus	0.292668
3	tipo	0.020528
4	v_edad	0.432034
5	potencia	0.001231
6	subtipo	0.000831

Importancia Cluster 1

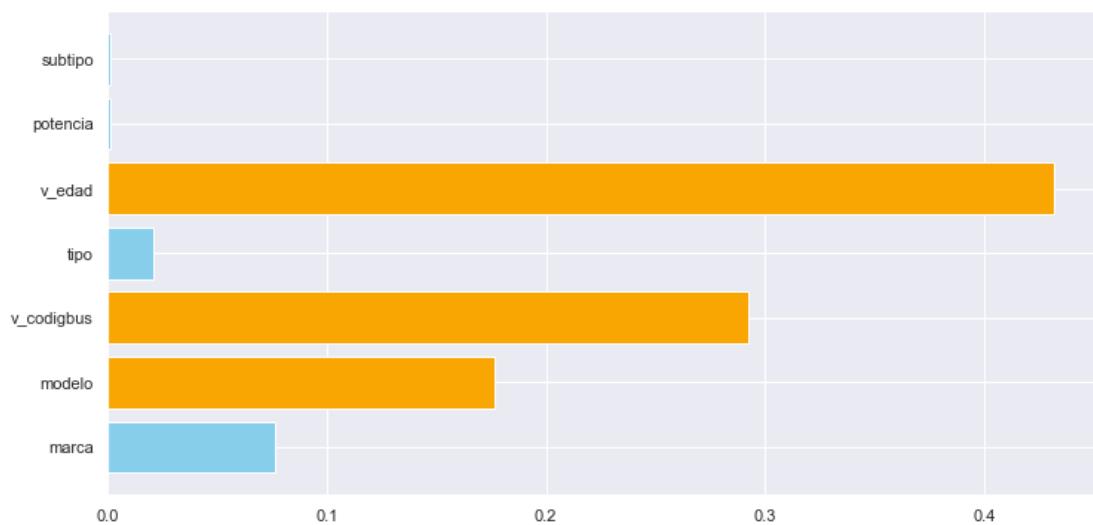


Ilustración 62. Resultados clúster 1 para fuga de aire. Ilustración propia

Como se puede ver en la ilustración 63, para el combinando, los resultados, ponderados multiplicando por un factor determinado por el tamaño de cada clúster, vemos que el modelo es capaz de predecir en un 61 % que el autobús sufre una avería por fuga de aire y un 62% de que no es por fuga de aire.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, influyendo también levemente la marca

PREDICCIÓN Combinada

```
Accuracy prediction: 0.6173469387755102
f1-score : 0.6173
```

Random Forest Classifier

	precision	recall	f1-score	support
0	0.64	0.61	0.62	305
1	0.60	0.62	0.61	283
accuracy			0.62	588
macro avg	0.62	0.62	0.62	588
weighted avg	0.62	0.62	0.62	588

Importancia:

	features	importancia
feature		
marca	marca	0.085757
modelo	modelo	0.167913
potencia	potencia	0.000948
subtipo	subtipo	0.032966
tipo	tipo	0.023320
v_codigbus	v_codigbus	0.263749
v_edad	v_edad	0.425348

Importancia

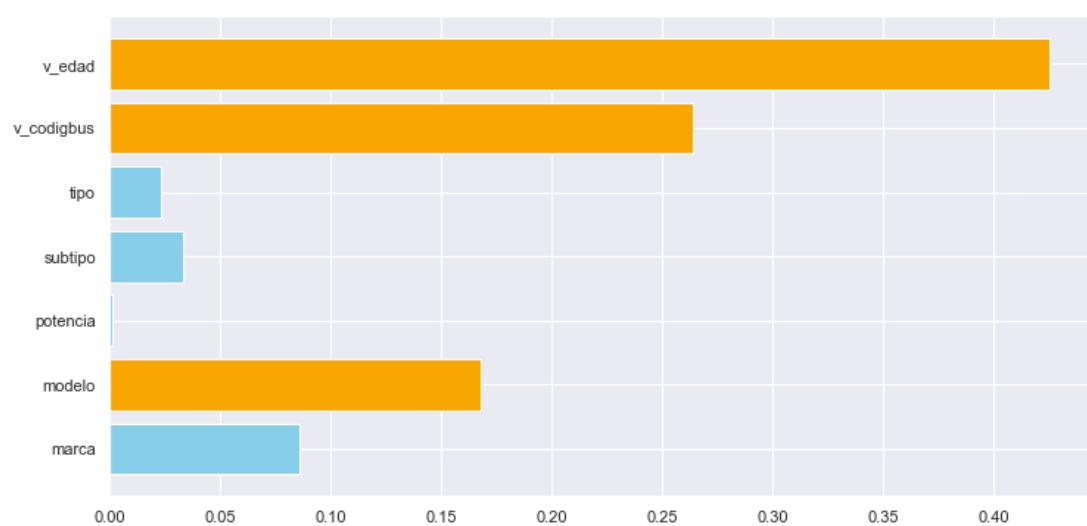


Ilustración 63. Resultados combinados para fuga de aire. Ilustración propia

4.3. Motor

Como se puede ver en la ilustración 64, para el clúster 0 de Motor vemos que el modelo es capaz de predecir en un 54 % que el autobús sufre una avería por correas y un 56% de que no es por Motor.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo.

```
PREDICCIÓN Cluster0

Accuracy prediction: 0.5531453362255966
f1-score :0.5531

Random Forest Classifier

      precision    recall  f1-score   support

          0       0.59      0.54      0.56      248
          1       0.51      0.57      0.54      213

   accuracy                           0.55      461
    macro avg       0.55      0.55      0.55      461
weighted avg       0.56      0.55      0.55      461
```

```
Importancia Cluster 0:
      feature  importancia
0      marca     0.055511
1     modelo     0.173608
2  v_codigbus    0.268832
3      tipo     0.023461
4    v_edad     0.469264
5   subtipo     0.001916
6    potencia    0.007408
```

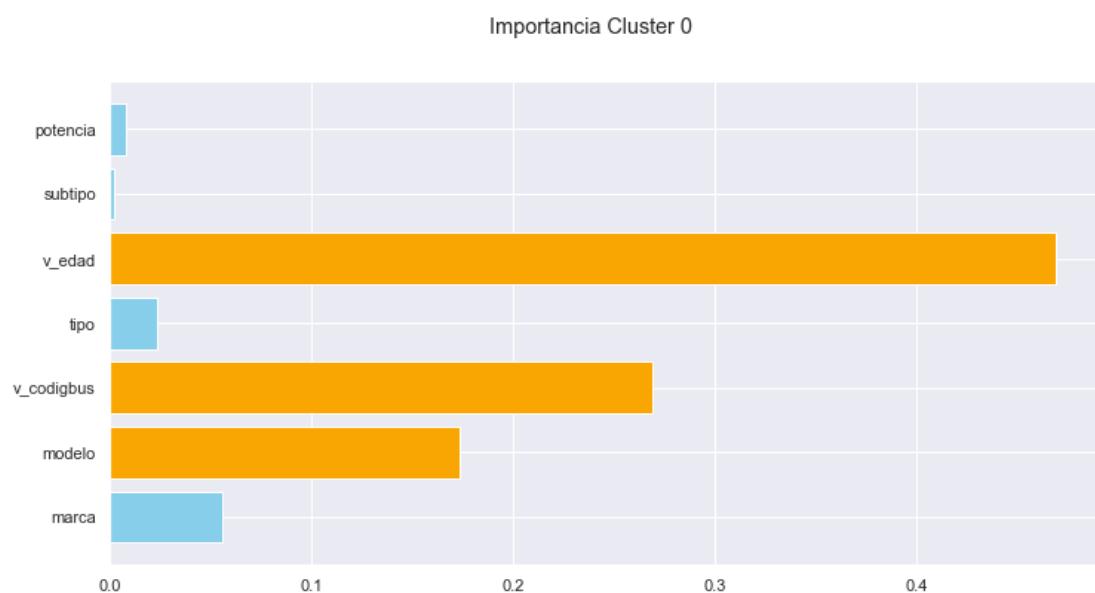


Ilustración 64. Resultados clúster 0 para motor. Ilustración propia

Como se puede ver en la ilustración 65, para el clúster 1 de Motor vemos que el modelo es capaz de predecir en un 59 % que el autobús sufre una avería por correas y un 50% de que no es por Motor.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, se observa también que la marca suele tener influencia para el clúster 1.

PREDICCIÓN Cluster 1

```
Accuracy prediction: 0.5454545454545454
f1-score :0.5455
```

```
Random Forest Classifier
```

	precision	recall	f1-score	support
0	0.49	0.50	0.50	74
1	0.59	0.58	0.59	91
accuracy			0.55	165
macro avg	0.54	0.54	0.54	165
weighted avg	0.55	0.55	0.55	165

Importancia Cluster 1:

	features	importancia
0	marca	0.117725
1	modelo	0.167336
2	v_codigbus	0.195475
3	tipo	0.020060
4	v_edad	0.472466
5	subtipo	0.026939

Importancia Cluster 1

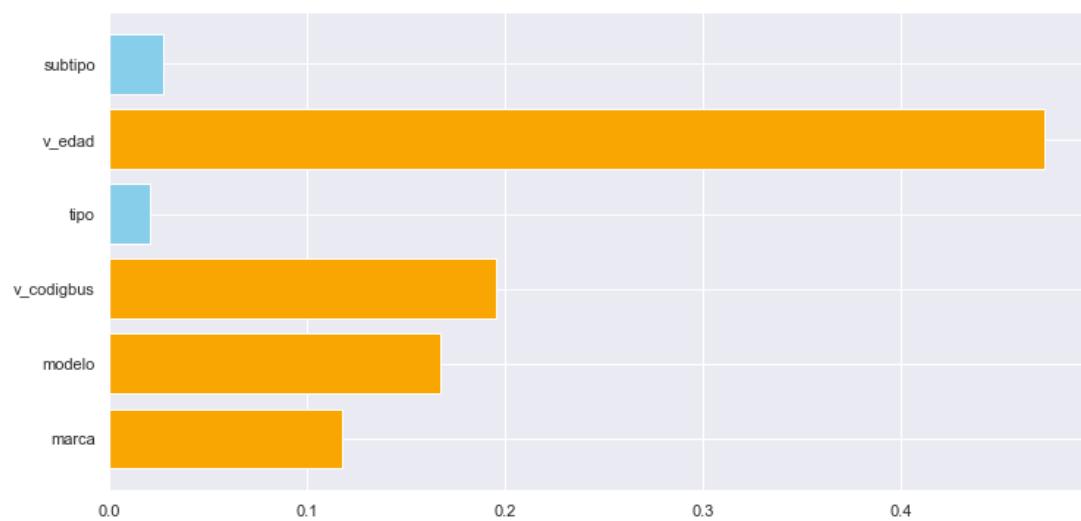


Ilustración 65. Resultados clúster 1 para motor. Ilustración propia

Como se puede ver en la ilustración 66, para el combinando, los resultados, ponderados multiplicando por un factor determinado por el tamaño de cada clúster, vemos que el modelo es capaz de predecir en un 55 % que el autobús sufre una avería Motor y un 55% de que no es por Motor.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, influyendo también levemente la marca

PREDICCIÓN Combinada

```
Accuracy prediction: 0.5511182108626198
f1-score :0.5511
```

Random Forest Classifier

	precision	recall	f1-score	support
0	0.57	0.53	0.55	322
1	0.54	0.58	0.55	304
accuracy			0.55	626
macro avg	0.55	0.55	0.55	626
weighted avg	0.55	0.55	0.55	626

	features	importancia
feature		
marca	marca	0.071686
modelo	modelo	0.171977
potencia	potencia	0.005482
subtipo	subtipo	0.008422
tipo	tipo	0.022577
v_codigbus	v_codigbus	0.249759
v_edad	v_edad	0.470097

Importancia

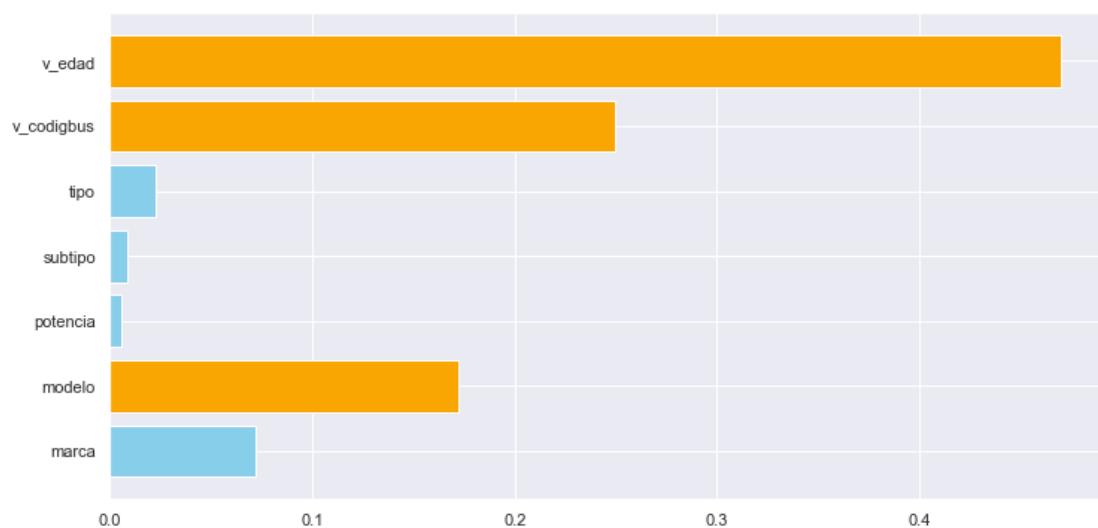


Ilustración 66. Resultados combinados para motor. Ilustración propia

4.4. Sistema de Frenos

Como se puede ver en la ilustración 67, para el clúster 0 de Sistema de Frenos vemos que el modelo es capaz de predecir en un 71% que el autobús sufre una avería por el Sistema Frenos y un 60% de que no es por el Sistema de Frenos.

PREDICCIÓN Cluster0

```
Accuracy prediction: 0.6677852348993288
f1-score :0.6678

Random Forest Classifier

      precision    recall   f1-score   support
          0       0.60     0.60     0.60      125
          1       0.71     0.72     0.71      173

   accuracy                           0.67      298
  macro avg       0.66     0.66     0.66      298
weighted avg       0.67     0.67     0.67      298
```

```
Importancia Cluster 0:
      feature  importancia
0      marca      0.124005
1     modelo      0.189119
2  v_codigbus      0.208087
3      tipo      0.018484
4     v_edad      0.395324
5    subtipo      0.064982
```

Importancia Cluster 0

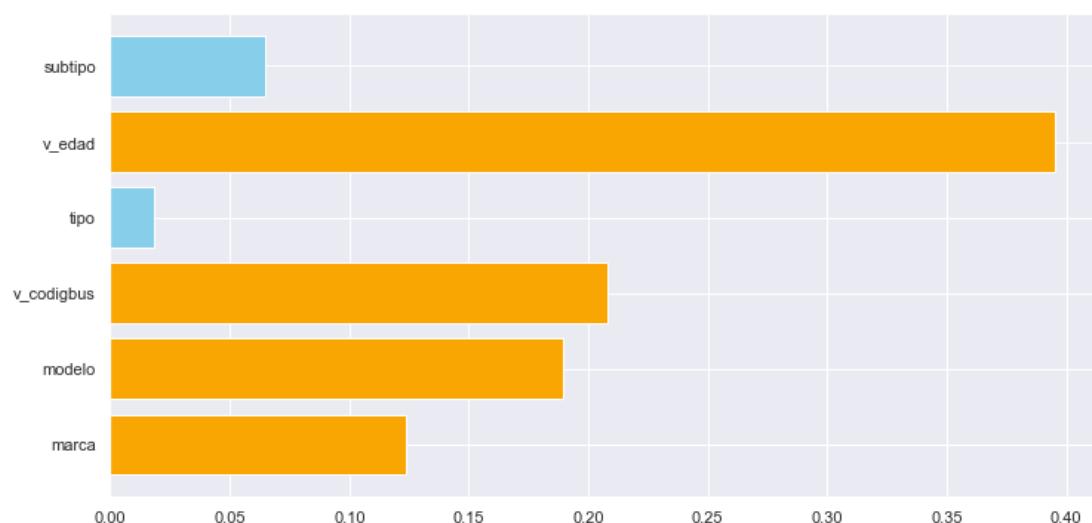


Ilustración 67. Resultados clúster 0 para sistema de frenos. Ilustración propia

Como se puede ver en la ilustración 68, para el clúster 1 de Sistema de Frenos vemos que el modelo es capaz de predecir en un 60% que el autobús sufre una avería por el Sistema Frenos y un 56% de que no es por el Sistema de Frenos.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo.

PREDICCIÓN Cluster 1

```
Accuracy prediction: 0.5792759051186017
f1-score :0.5793
```

Random Forest Classifier

	precision	recall	f1-score	support
0	0.61	0.58	0.60	429
1	0.54	0.58	0.56	372
accuracy			0.58	801
macro avg	0.58	0.58	0.58	801
weighted avg	0.58	0.58	0.58	801

Importancia Cluster 1:

	features	importancia
0	marca	0.060553
1	modelo	0.170889
2	v_codigbus	0.291375
3	tipo	0.028712
4	v_edad	0.440253
5	potencia	0.001023
6	subtipo	0.007196

Importancia Cluster 1

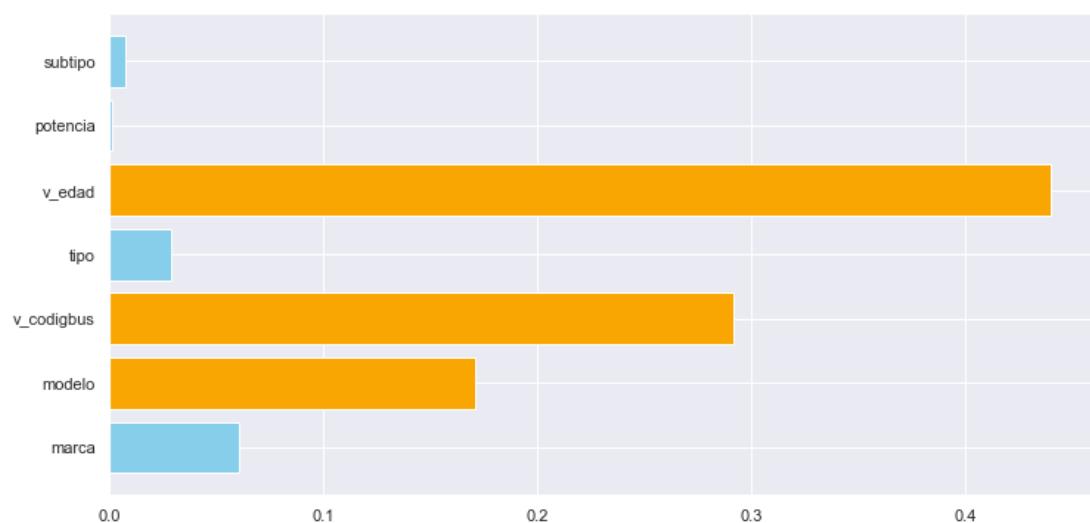


Ilustración 68. Resultados clúster 1 para sistema de frenos. Ilustración propia

Como se puede ver en la ilustración 69, para el combinando, los resultados, ponderados multiplicando por un factor determinado por el tamaño de cada clúster, vemos que el modelo es capaz de predecir en un 61 % que el autobús sufre una avería Motor y un 60% de que no es por Motor.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, influyendo también levemente la marca

PREDICCIÓN Combinada

```
Accuracy prediction: 0.6032757051865332
f1-score :0.6033
```

```
Random Forest Classifier
```

	precision	recall	f1-score	support
0	0.61	0.59	0.60	554
1	0.60	0.62	0.61	545
accuracy			0.60	1099
macro avg	0.60	0.60	0.60	1099
weighted avg	0.60	0.60	0.60	1099

Importancia:

feature	features	importancia
marca	marca	0.077685
modelo	modelo	0.175811
potencia	potencia	0.000747
subtipo	subtipo	0.022798
tipo	tipo	0.025950
v_codigbus	v_codigbus	0.268887
v_edad	v_edad	0.428122

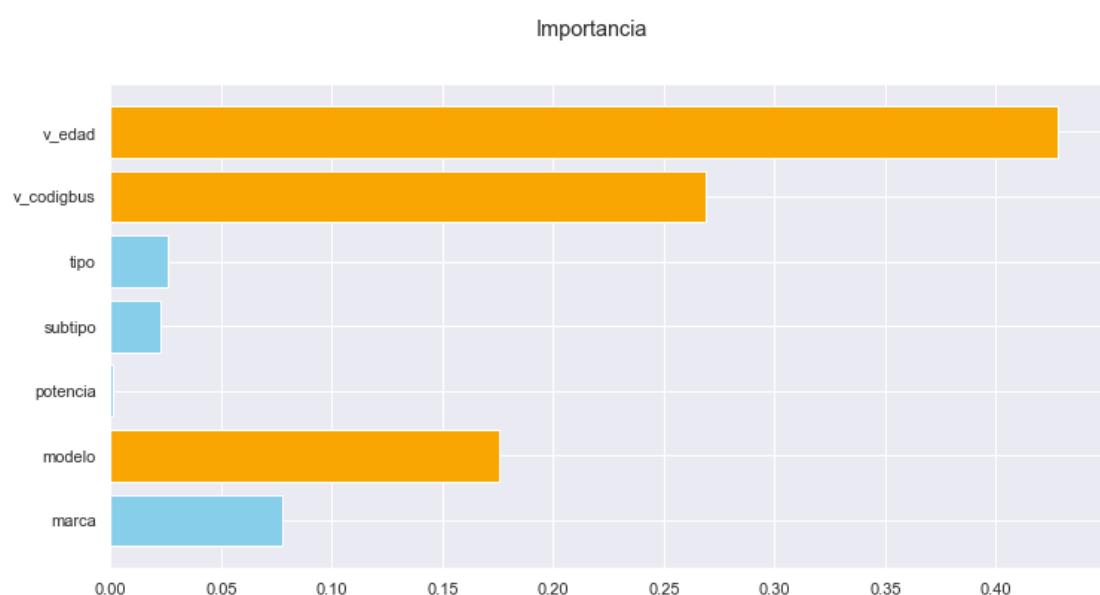


Ilustración 69. Resultados combinados para sistema de frenos. Ilustración propia

4.5. Sistema de Refrigeración

Como se puede ver en la ilustración 70, para el clúster 0 de Sistema de Refrigeración vemos que el modelo es capaz de predecir en un 58% que el autobús sufre una avería por el Sistema Refrigeración y un 57% de que no es por el Sistema de Frenos.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús el modelo y la marca.

PREDICCIÓN Cluster0

```
Accuracy prediction: 0.5759493670886076
f1-score :0.5759
```

Random Forest Classifier

	precision	recall	f1-score	support
0	0.61	0.54	0.57	415
1	0.55	0.61	0.58	375
accuracy			0.58	790
macro avg	0.58	0.58	0.58	790
weighted avg	0.58	0.58	0.58	790

Importancia Cluster 0:

	feature	importancia
0	marca	0.123444
1	modelo	0.220256
2	v_codigbus	0.263065
3	tipo	0.027398
4	v_edad	0.336935
5	potencia	0.006585
6	subtipo	0.022317

Importancia Cluster 0

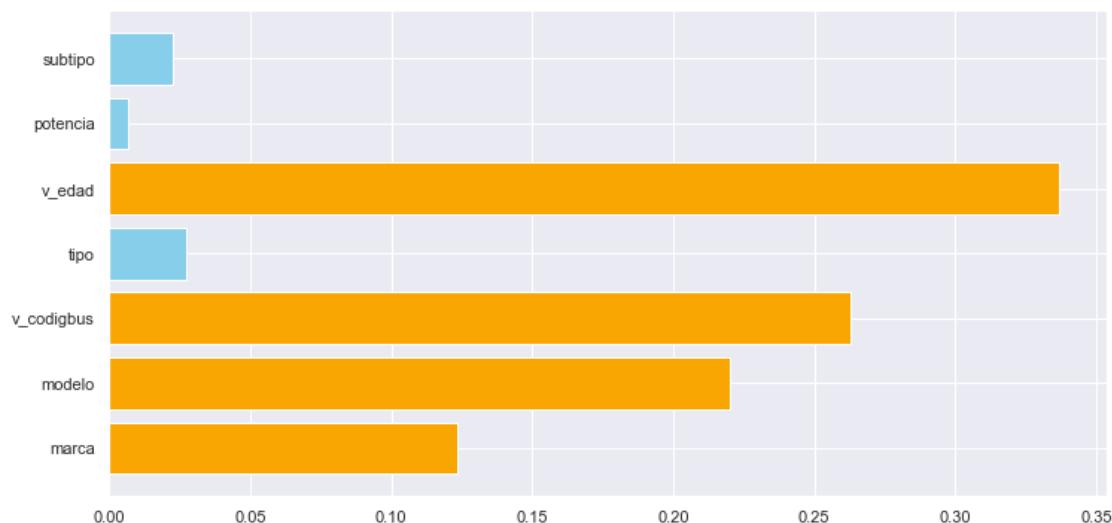


Ilustración 70. Resultados clúster 0 para sistema de refrigeración. Ilustración propia

Como se puede ver en la ilustración 71, para el clúster 1 de Sistema de Refrigeración vemos que el modelo es capaz de predecir en un 59% que el autobús sufre una avería por el Sistema Refrigeración y un 59% de que no es por el Sistema de Frenos. Y que las variables más influyentes para determinarlo son, la edad, el código del autobús el modelo

```
PREDICCIÓN Cluster 1

Accuracy prediction: 0.5877862595419847
f1-score :0.5878

Random Forest Classifier

      precision    recall   f1-score   support
          0       0.60     0.57     0.59      134
          1       0.57     0.60     0.59      128

   accuracy                           0.59      262
  macro avg                           0.59      262
weighted avg                          0.59      262
```

```
Importancia Cluster 1:
      features  importancia
0       marca      0.047133
1     modelo      0.133457
2  v_codigbus     0.172308
3      tipo      0.077931
4   v_edad      0.485774
5    subtipo     0.083397
```



Ilustración 71. Resultados clúster 1 para sistema de refrigeración. Ilustración propia

Como se puede ver en la ilustración 72, para el combinando, los resultados, ponderados multiplicando por un factor determinado por el tamaño de cada clúster, vemos que el modelo es capaz de predecir en un 58 % que el autobús sufre una avería Motor y un 58% de que no es por Motor.

Y que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, influyendo también la marca

PREDICCIÓN Combinada

```
Accuracy prediction: 0.5788973384030418
f1-score :0.5789
```

Random Forest Classifier

	precision	recall	f1-score	support
0	0.61	0.55	0.58	549
1	0.55	0.61	0.58	503
accuracy			0.58	1052
macro avg	0.58	0.58	0.58	1052
weighted avg	0.58	0.58	0.58	1052

Importancia:	features	importancia
feature		
marca	marca	0.104366
modelo	modelo	0.198556
potencia	potencia	0.004938
subtipo	subtipo	0.037587
tipo	tipo	0.040032
v_codigbus	v_codigbus	0.240376
v_edad	v_edad	0.374145

Importancia

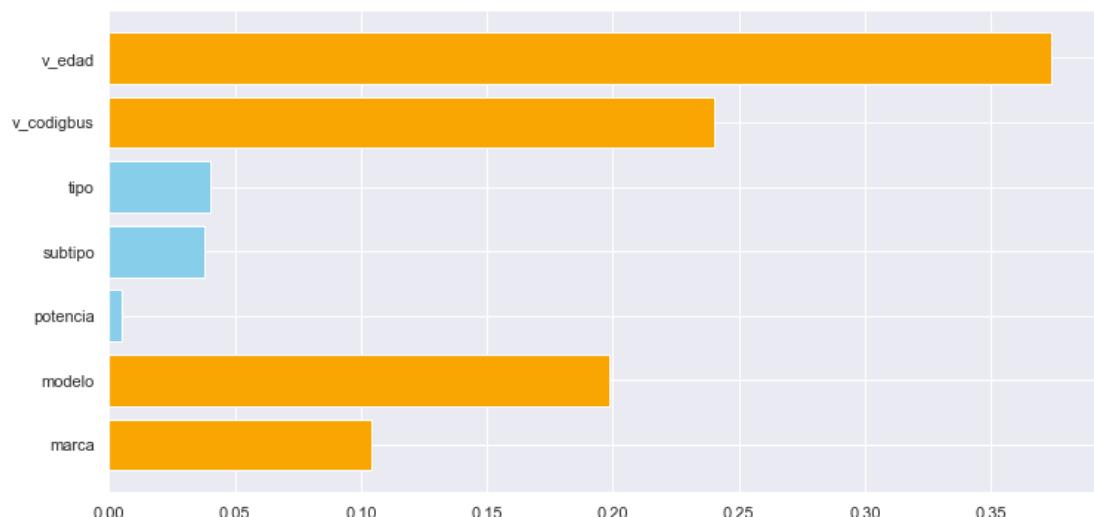


Ilustración 72. Resultados combinados para sistema de refrigeración. Ilustración propia

5. Conclusiones

Como conclusión podemos decir que las predicciones a partir de las características del autobús no son muy relevantes ya que no se dispone de datos que se puedan relacionar con el desgaste para llevar a una reparación por dichas piezas, o bien porque el número de observaciones no era muy alto.

Pudiendo mejorarse los resultados aumentando el número de observaciones o añadiendo características que puedan darnos información sobre el desgaste de dichos grupos de operaciones o con combinación de ambas.

Lo que sí se ha observado es que los datos por subtipo ‘Normal’ se distribuyen claramente de manera distinta que el resto de los subtipos. Y que cuanto más influye la marca en la predicción, los resultados mejoran.

Los casos más relevantes son:

La precisión para saber si hay fuga de aire en los autobuses de subtipo ‘Normal’ es del 72%, dándonos la información de que las variables más influyentes para determinarlo son, la edad, el código del autobús y el modelo, la marca y el subtipo.

La precisión para saber si hay un fallo en el sistema de Freno en los autobuses de subtipo normal es del 67%, dándonos la información de que las variables más influyentes para determinarlo son, la edad, el código del autobús, el modelo y la marca.

5.1. Futuras líneas de mejora

Buscar de forma descriptiva con un software de Business Intelligence (Power BI), las relaciones que existen entre Fuga de aire y Sistema de frenado, con la edad, el código del autobús, el modelo, la marca y el subtípo.

Ampliar el número de observaciones para cada uno de los subgrupos y ampliar el número de variables medibles para cada observación.

Cambio de objetivo, averiguar en cuanto tiempo volverá al taller por una avería concreta.

6. Bibliografía

- [1] KAMBATLA, Karthik, et al. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 2014, vol. 74, no 7, p. 2561-2573.
- [2] O'Reilly Media, Inc. Big Data Now: 2012. "O'Reilly Media, Inc.", 2012.
- [3] ABC. Big data: ¿vidas privadas al alcance de todos?. URL: <http://www.abc.es/tecnologia/informatica-software/20131028/abci-entrevista-data201310221252.html> [13 de Febrero del 2016]
- [4] BBVA Innovation Center. Proyecto Big Data. URL: <http://www.centrodeinnovacionbbva.com/proyectos/big-data> [14 de Febrero del 2016]
- [5] BASANTA-VAL, P., et al. Improving the predictability of distributed stream processors. *Future Generation Computer Systems*, 2015, vol. 52, p. 22-36.
- [6] SALMERON, J. (15 de Enero de 2016). "¿Qué herramientas necesitas para iniciarte en Big Data?". Recuperado el 21 de Septiembre de 2018, de inLabFIB: <https://inlab.fib.upc.edu/es/blog/que-herramientas-necesitas-para-iniciarte-en-big-data>
- [7] Richard Szeliski. *Computer Vision: Algorithms and Applications*. SpringerVerlag New York, Inc., New York, NY, USA, 1st edition, 2010.
- [8] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.
- [9] Sylvain Arlot and Alain Celisse. 2009. A survey of cross-validation procedures for model selection. 4, (2009), 40–79. DOI:<https://doi.org/10.1214/09-SS054>
- [10] W Hu, W Hu, and S Maybank. 2008. AdaBoost-Based Algorithm for Network Intrusion Detection. *IEEE Trans. Syst. Man, Cybern. Part B* 38, 2 (2008), 577–583. DOI:<https://doi.org/10.1109/TSMCB.2007.914695>
- [11] Sanjay Kumar Palei and Samir Kumar Das. 2009. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. *Saf. Sci.* 47, 1 (January 2009), 88–96. DOI:<https://doi.org/10.1016/J.SSCI.2008.01.002>