TRABAJO FIN DE MÁSTER

MÁSTER EN CIENCIA DE DATOS

Minería de datos y relación entre tipos de inteligencias

La inteligencia energética, más allá de la inteligencia emocional

Autor

José Pérez-Parras Toledano

Directores

María Del Carmen Pegalajar Jimenez

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

—

Granada, Septiembre de 2019

Minería de datos y relación entre tipos de inteligencias

José Pérez-Parras Toledano

Palabras clave: Minería de datos, Psicología, Clustering, Análisis de datos, Aprendizaje supervisado, Aprendizaje no supervisado, Regresión, Clasificación, Redes neuronales.

Resumen

Este trabajo de Fin de Máster ha podido llevarse a cabo gracias a los datos recogidos en la investigación desarrollada por Elena Pérez Moreiras (doctoranda) en su tesis doctoral que lleva por título "La inteligencia energética, más allá de la inteligencia emocional", dirigida por Joan Boada (catedrático de la Universidad de Rovira y Virgili). Para la demostración de la inteligencia energética en el ámbito de la psicología. El objetivo de este trabajo es realizar un análisis y minería de datos de un gran número de instancias asociadas a las respuestas de un conjunto de test psicológicos que intentan medir diferentes características de los individuos. A partir de estas muestras, intentaremos obtener algunas conclusiones sobre los datos recopilados por el estudio psicológico.

Data mining and relations between different types of intelligences

José Pérez-Parras Toledano

Keywords: Data Mining, Psicology, Clustering, Data Analysis, Supervised Learning, Unsupervised learning, Regression, Classification, Neural networks.

Abstract

This Final year project has been undergone with the help of the data collected for the research made by Elena Pérez Moreiras in her PhD which is named after this title .Energetic Intelligence, far after the the emotional intelligence". Directed by Joan (professor of Rovida & Virgili's University). For the demonstration of the energetic intelligence in the field of psychology. The main task of this project is to make an analysis and data mining of the great ammount of psicological tests which measures different aspects of each of the individuals tested. From theses tests, we would try to reach out some conclusions about the data collected in this study.

Agradecimientos

Muchas gracias a todos mis amigos por apoyarme durante la duración de este trabajo. En especial a mis padres y a María (aunque haya terminado este trabajo a pesar de los viajes). Sobretodo dar las gracias a mis tutores por contestar tantos correos (a todas horas) y ayudarme tanto durante este trabajo. También a mis compañeros del máster por haber pasado un año inolvidable y haber aprendido tanto, sobretodo por esas milnos de descanso en el Peruano.

Muchas gracias a todos. De corazón.

Índice general

1. Introducción **1**

2. Análisis de los datos proporcionados **3**

2.1. Análisis de las variables en conjunto 3
2.2. Análisis de las variables de carácter personal 4
2.2.1. Género 6
2.2.2. Edad 7
2.2.3. Estado civil 9
2.2.4. Situación laboral 9
2.2.5. Formación 10
2.2.6. Antigüedad en la empresa 11
2.2.7. Antigüedad en el cargo 12
2.2.8. Antigüedad en la profesión 14
2.2.9. Grupo profesional (CIUO) 15
2.2.10. Sector 16
2.2.11. Banca Internacional 17
2.2.12. Número de empleados en la empresa 17
2.2.13. Ámbito de actuación 18
2.2.14. Situación de la empresa en los últimos 12 meses 19
2.3. Análisis de las variables psicológicas y organizacionales de los encuestados 19
2.3.1. Flow-Fluir 20
2.3.2. Autoeficacia 22
2.3.3. Autoestima 23
2.3.4. Personalidad-OPEARS 25
2.3.5. Flourishing-Florecimiento 27
2.3.6. Engagement 28
2.3.7. Satisfacción con la vida 30
2.3.8. Inteligencia emocional 32
2.3.9. Energía (Inteligencia energética) 34
2.3.10. Calidad personal y Organizacional 36

3. Modelos de Machine Learning usados **41**
3.1. Regresión Lineal 41
3.2. Nearest Neighbour 42
3.3. SVM 42
3.4. Decision Tree 43
3.5. Random Forest 43
3.6. Multi-Layer Perceptron 43
3.7. Principal Component Analysis (PCA) 44
3.8. K-Means 44
3.9. Hierarchical Clustering 45
3.10. DBSCAN 45

4. Relaciones entre variables **47**
4.1. Correlación entre variables y *heatmap* 47
4.2. Correlación usando Random Forest 51
4.2.1. V1 (Flow-Fluir) 53
4.2.2. V9 (Flourishing-Florecimiento) 54
4.2.3. V19 (Self Drive Direction) 55
4.2.4. V20 (Consciencia energética) 56
4.2.5. V21 (Energía total) 57
4.2.6. Conclusiones 58

5. Procedimiento de los experimentos a realizar **59**

6. Uso de modelos de regresión para los distintos problemas **65**
6.1. Métricas usadas 65
6.1.1. Mean Squared Error (MSE) 65
6.1.2. Explained Variance Score 66
6.1.3. R2 Score 66
6.2. Modelos y parámetros escogidos 66
6.2.1. Linear Regression 67
6.2.2. Nearest Neighbor Regressor 67
6.2.3. SVM Regressor 67
6.2.4. Decision Tree Regressor 67
6.2.5. Random Forest Regressor 68
6.2.6. Multi-Layer Perceptron Regressor 68
6.3. Resultados obtenidos 70
6.3.1. Conclusiones 70
6.4. Repetición de los experimentos sin las variables personales . . 71
6.4.1. V1 (Flow-Fluir) 72
6.4.2. V9 (Flourishing-Florecimiento) 73
6.4.3. V19 (Self Drive Direction) 74
6.4.4. V20 (Consciencia energética) 74
6.4.5. V21 (Energía total) 75

6.4.6. Conclusiones 75

7. Aplicación de técnicas de aprendizaje no supervisado 77
7.1. Visualización de los datos utilizando Principal Component Analysis (PCA) 77
7.2. Métricas utilizadas para medir nuestros clusters 80
7.2.1. Silhouette Coefficient 80
7.2.2. Davies-Bouldin Index 81
7.3. Modelos escogidos y resultados 81
7.3.1. K-Means 82
7.3.2. Hierarchical Clustering 83
7.3.3. DBSCAN 87
7.4. Conclusiones 88

8. Uso de modelos de clasificación multiclase 89
8.1. Creación de intervalos en nuestros datos 89
8.1.1. Visualización de la distribución de clases por variable 90
8.1.2. V9 (Flurishing-Florecimiento) 92
8.1.3. V19 (Self Drive Direction) 94
8.1.4. V20 (Consciencia energética) 96
8.1.5. V21 (Energía total) 98
8.2. Aplicación de técnicas de aprendizaje supervisado (multiclase) 99
8.2.1. Modelos escogidos 100
8.2.2. Métricas usadas 100
8.2.3. Resultados con todas las variables del problema 101
8.2.4. Resultados sin usar las variables sociales 101
8.2.5. Representación de los árboles creados por Random Forest 105

9. Conclusiones y Trabajos Futuros 111

Bibliografía 116

Apéndices 116

A. Resultados obtenidos con todas las variables para regresión 117
A.1. V1 (Flow-Fluir) 117
A.2. V9 (Flourishing-Florecimiento) 118
A.3. V19 (Self Drive Direction) 119
A.4. V20 (Consciencia energética) 119
A.5. V21 (Energía total) 120

B. Resultados obtenidos con todas las variables en clasificación 123
B.1. V1 (Flow-Fluir) 123
B.2. V9 (Flourishing-Florecimiento) 124
B.3. V19 (Self Drive Direction) 124
B.4. V20 (Consciencia energética) 125
B.5. V21 (Energía total) 125

Índice de figuras

2.1. Género de los encuestados. 6
2.2. Boxplot de Edad. 7
2.3. Edad de los encuestados. 8
2.4. Estado civil de los encuestados. 9
2.5. Situación laboral de los encuestados. 10
2.6. Formación de los encuestados. 10
2.7. Boxplot de antigüedad en la empresa de los encuestados. . . . 11
2.8. Antigüedad en la empresa de los encuestados. 12
2.9. Boxplot de antigüedad en el cargo de los encuestados. 13
2.10. Antigüedad en el cargo de los encuestados. 13
2.11. Boxplot de antigüedad en profesión de los encuestados. 14
2.12. Antigüedad en profesión de los encuestados. 15
2.13. Grupo profesional de los encuestados. 15
2.14. Sector de los encuestados. 16
2.15. Banca Internacional. 17
2.16. Tamaño de la empresa. 17
2.17. Ámbito de actuación de la empresa. 18
2.18. Situación en la empresa en el último año. 19
2.19. Boxplot de Flow-Fluir. 21
2.20. Boxplot de Autoeficacia. 23
2.21. Boxplot de Autoestima. 24
2.22. Boxplot de OPERAS. 26
2.23. Boxplot de Flourishing-Florecimiento. 28
2.24. Boxplot de Engagement. 30
2.25. Boxplot de Satisfacción con la vida. 31
2.26. Boxplot de Inteligencia emocional. 33
2.27. Boxplot de Inteligencia energética. 36
2.28. Boxplot de Calidad personal y Organizacional. 39
2.29. Boxplot de Calidad personal y Organizacional total. 40

3.1. Ejemplo de regresión lineal. 42
3.2. Ejemplo de red neuronal. 44

4.1. Heatmap con todas las variables. 48
4.2. Heatmap con las variables sociales. 49
4.3. Heatmap con las variables psicológicas. 50
4.4. Importancia de las variables para V1 según Random Forest. . 53
4.5. Importancia de las variables para V9 según Random Forest. . 54
4.6. Importancia de las variables para V19 según Random Forest. 55
4.7. Importancia de las variables para V20 según Random Forest. 56
4.8. Importancia de las variables para V21 según Random Forest. 57

5.1. Diagrama del proceso del experimento. 60

6.1. Ejemplo de red neuronal con nomenclatura. 69

7.1. Visualización de los datos con PCA. 78
7.2. Visualización de los datos con PCA sin la variable Género. . . 79
7.3. Visualización de los datos con PCA sin las variables sociales. 80
7.4. Visualización de los clusters formados por KMeans. 82
7.5. Visualización de los clusters formados por Hierarchical Clustering (Ward). 83
7.6. Visualización de los clusters formados por Hierarchical Clustering (Complete linkage). 84
7.7. Visualización de los clusters formados por Hierarchical Clustering (Average linkage). 85
7.8. Visualización de los clusters formados por Hierarchical Clustering (Single linkage). 86
7.9. Visualización de los clusters formados por DBSCAN. 87

8.1. Histograma de las clases asignadas a Flow-Fluir. 90
8.2. Visualización PCA sobre Flow-Fluir multiclase. 91
8.3. Histograma de las clases asignadas a Flourishing-Florecimiento. 92
8.4. Visualización PCA sobre Flourishing-Florecimiento multiclase. 93
8.5. Histograma de las clases asignadas a Self Drive Direction. . . 94
8.6. Visualización PCA sobre Self Drive Direction multiclase. . . . 95
8.7. Histograma de las clases asignadas a Consciencia energética. . 96
8.8. Visualización PCA sobre Consciencia energética multiclase. . 97
8.9. Histograma de las clases asignadas a Energía total. 98
8.10. Visualización PCA sobre Energía total multiclase. 99
8.11. Ejemplo de Decision Tree para la variable V1 (1). 105
8.12. Ejemplo de Decision Tree para la variable V1 (2). 106
8.13. Ejemplo de Decision Tree para la variable V1 (3). 106
8.14. Ejemplo de Decision Tree para la variable V1 (4). 106
8.15. Ejemplo de Decision Tree para la variable V9 (1). 107
8.16. Ejemplo de Decision Tree para la variable V9 (2). 107
8.17. Ejemplo de Decision Tree para la variable V9 (3). 107
8.18. Ejemplo de Decision Tree para la variable V9 (4). 108

8.19. Ejemplo de Decision Tree para la variable V21 (1). 109
8.20. Ejemplo de Decision Tree para la variable V21 (2). 109
8.21. Ejemplo de Decision Tree para la variable V21 (3). 109
8.22. Ejemplo de Decision Tree para la variable V21 (4). 109

Índice de cuadros

2.1. Variables sociales de nuestro problema. 6
2.2. Análisis numérico de la variable edad. 7
2.3. Análisis numérico de la antigüedad en la empresa. 11
2.4. Análisis numérico de la antigüedad en el cargo. 12
2.5. Análisis numérico de la antigüedad en la profesión. 14
2.6. Variables psicológicas en nuestro problema. 20
2.7. Análisis numérico de Flow-Fluir. 21
2.8. Análisis numérico de Autoeficacia. 22
2.9. Análisis numérico de Autoestima. 24
2.10. Análisis numérico de OPERAS. 25
2.11. Análisis numérico de Flourishing-Florecimiento. 28
2.12. Analisis numérico de Engagement. 29
2.13. Análisis numérico de Satisfacción con la vida. 31
2.14. Análisis numérico de Inteligencia emocional. 33
2.15. Análisis numérico de Inteligencia energética. 35
2.16. Análisis numérico de Calidad personal y Organizacional. . . . 38

5.1. Variables de los experimentos. 63

6.1. Parámetros a mejorar para kNN Regressor. 67
6.2. Parámetros a mejorar para SVM Regressor. 67
6.3. Parámetros a mejorar para Decision Tree Regressor. 67
6.4. Parámetros a mejorar para Random Forest. 68
6.5. Parámetros a mejorar para la Red Neuronal. 69
6.6. Experimentos con únicamente variables psicológicas. 72
6.7. Resultados de regresión para V1 sin variables personales. . . 73
6.8. Resultados de regresión para V9 sin variables personales. . . 73
6.9. Resultados de regresión para V19 sin variables personales. . . 74
6.10. Resultados de regresión para V20 sin variables personales. . . 74
6.11. Resultados de regresión para V21 sin variables personales. . . 75

7.1. Resultados obtenidos por los métodos no supervisados. 88

8.1. Intervalos y valores para las clases creadas. 89

8.2. Resultados multiclase para V1 sin las variables sociales. . . . 102
8.3. Resultados multiclase para V9 sin las variables sociales. . . . 102
8.4. Resultados multiclase para V19 sin las variables sociales. . . . 103
8.5. Resultados multiclase para V20 sin las variables sociales. . . . 103
8.6. Resultados multiclase para V21 sin las variables sociales. . . . 104

A.1. Resultados de regresión para V1. 117
A.2. Resultados de regresión para V9. 118
A.3. Resultados de regresión para V19. 119
A.4. Resultados de regresión para V20. 120
A.5. Resultados de regresión para V21. 121

B.1. Resultados multiclase para la variable V1. 123
B.2. Resultados multiclase para la variable V9. 124
B.3. Resultados multiclase para la variable V19. 124
B.4. Resultados multiclase para la variable V20. 125
B.5. Resultados multiclase para la variable V21. 125

Capítulo 1

Introducción

Este trabajo de Fin de Máster ha podido llevarse a cabo gracias a los datos recogidos en la investigación desarrollada por Elena Pérez Moreiras (doctoranda) en su tesis doctoral que lleva por título "La inteligencia energética, más allá de la inteligencia emocional", dirigida por Joan Boada (catedrático de la Universidad de Rovira y Virgili). Para la demostración de la inteligencia energética en el ámbito de la psicológica. El objetivo de este trabajo es realizar un análisis y minería de datos de un gran número de instancias asociadas a las respuestas de un conjunto de test psicológicos que intentan medir diferentes características de los individuos. A partir de estas muestras, intentaremos obtener algunas conclusiones sobre los datos recopilados por el estudio psicológico.

Dentro de las tareas que cubriremos con este proyecto, estarán:

- Realizar un análisis de las variables del problema a estudiar.
- Realizar un estudio correlacional predictivo para descubrir qué variables predictoras afectan a la Inteligencia Energética.
- Realizar un estudio sobre las variables a predecir con métodos de regresión.
- Realizar un estudio con técnicas de aprendizaje no supervisado en nuestros datos.
- Realizar un estudio sobre las variables a predecir con métodos de clasificación multiclase.

Para ello primero realizaremos un primer análisis de todos los datos e intentaremos encontrar relaciones estadísticas entre las variables de nuestro problema, intentando relacionar la inteligencia energética con otro tipo de variables psicológicas o personales. Viendo la relación existente entre las

mismas, es decir, validando la relación existente entre los distintos tipos de inteligencias.

Para la predicción de la inteligencia energética de los sujetos, usaremos tres vertientes para la misma. La primera será abordar el problema con modelos de regresión, puesto que las variables psicológicas están tratadas como variables continuas, cada una dentro de un rango de aplicación. Posteriormente tras el estudio con técnicas de regresión, se usarán técnicas de modelado no supervisado para nuestros datos, intentando detectar agrupaciones de nuestros encuestados en función de sus puntuaciones obtenidas en la encuesta. De esta manera intentaremos detectar características de los individuos en función del cluster al que pertenezcan. Gracias a todo el conocimiento extraído en el estudio desarrollado , se podrán etiquetar los encuestados con valores multiclase para sus valores de inteligencia energética, pudiendo entonces aplicar algoritmos de clasificación multiclase. Finalmente, se mostrarán los resultados obtenidos por cada uno de los enfoques y las conclusiones derivadas de todo el estudio del problema.

La forma en que los datos van a ser tratados para el estudio del problema, así como el análisis de los datos y los modelos usados serán descritos en las siguientes secciones.

Capítulo 2

Análisis de los datos proporcionados

2.1. Análisis de las variables en conjunto

En este apartado realizaremos un primer estudio estadístico y una visualización de nuestros datos, el cual nos ayudará a comprender mejor el problema con el que estamos tratando. Las encuestas proporcionadas nos dan un total de **856** personas encuestadas. Cada una de estas personas cuenta con **47** variables a analizar, de las cuales existirán 5 variables que estudiaremos más a fondo, puesto que serán las variables a predecir en nuestro problema: **Flow-Fluir**, **Flourishing-Florecimiento**, **Self Drive Direction**, **Consciencia energética** y **Energía total**.

Cada uno de los encuestados cuenta con **19** variables de tipo personal y empresarial, sobre la situación laboral, familiar, etc. . Estas variables serán analizadas en la siguiente sección. Además, se dispone de **28** variables de carácter psicológico, creadas a partir de las respuestas dadas a la encuesta aplicada. De estas variables, nos centraremos en predecir 5 de ellas asociadas a la inteligencia energética.

Todas las variables utilizadas en la encuesta pueden tener relación con las variables a predecir y este será el primer punto a analizar en nuestro trabajo, la relaciones que existan entre las distintas variables de nuestro problema.

Estos datos son bastante variados, al haber sido rellenados en una encuesta online puede que algunas de estas personas no haya respondido correctamente o que existan valores nulos en nuestros datos. Debemos de tener especial cuidado con dichos valores puesto que influyen en el modelo a entrenar. Por ello, haciendo una primera observación obtenemos que solamente

existen dos variables que no han sido completadas por todos nuestros encuestados.

Se puede observar que son datos que algunas personas pudieran no saber o no querer proporcionar. Debemos de tener cuidado con este tipo de variables, pues al haber sido introducidos por los usuarios sin ningún tipo de restricción, de una manera abierta y sin dar a escoger en un conjunto, no podemos acotarlas en unos valores fijos.

Todas las variables proporcionadas por el problema serán mostradas y explicadas en los siguientes apartados.

2.2. Análisis de las variables de carácter personal

Se ha realizado un pequeño análisis de las variables personales de los encuestados. Para poder posteriormente crear grupos sociales con los mismos o analizar si los resultados psicológicos obtenidos tienen relación con los atributos sociales de cada persona encuestada. Las variables asociadas a los datos personales de los encuestados y sus posibles valores se pueden ver en la siguiente tabla 2.1:

Género	1. Hombre. 2. Mujer.
Edad	(Años)
Estado civil	1. Casado/a o pareja de hecho. 2. Soltero/a. 3. Divorciado/a o separado/a. 4. Viudo/a.
Situación laboral-ocupacional	1. Empleado/a. 2. Desempleado/a. 3. Prejubilado/a o jubilado/a. 4. Voluntario/a. 5. Estudiante. 6. De baja.
Nivel de formación académica	1. Sin estudios. 2. Estudios primarios finalizados. 3. Estudios secundarios finalizados. 4. Estudios de máster/doctorado finalizados.
Unidad de Negocio-División	(A introducir por el usuario)
Antigüedad en la empresa-institución	(Años)
Cargo	(A introducir por el usuario)
Número total de profesionales/empleados a su cargo	(A introducir por el usuario)
Antigüedad en el cargo	(Años)

Antigüedad en esta profesión	(Años)
Grupo profesional (CIUO)	1. Personal directivo. 2. Profesionales, científicos e intelectuales. 3. Técnicos profesionales de nivel medio. 4. Administrativos. 5. Trabajadores de servicios, vendedores de comercio y mercados. 6. Agricultores. 7. Oficiales y operarios. 8. Trabajadores no cualificados. 9. Ocupaciones militares y policía. 10. Otro.
Sector	1. Banca-entidades financieras o similares. 2. Seguros o similares. 3. Comercio, Transporte y Distribución. 4. Industria. 5. Ingeniería, Construcción, Obra Pública, Energía o similares. 6. Hostelería, Turismo o similares. 7. Sanidad, Farmaceúticas o similares. 8. Organismos oficiales, Ejército o similares. 9. Cultura, Arte, Educación o similares. 10. Consultoría, Asesoría, Servicios Profesionales o similares. 11. Informática, IT, Software, Tecnología o similares. 12. Otros.
¿Considera que su empresa, división y cargo entran dentro del grupo "Banca internacional¿	Sí. No.
Número de empleados en empresa	(Número de personas)
Factual anual aproximada	(A introducir por el usuario en Euros o Dólares)
Ámbito de actuación de la empresa	1. Nacional o USA. 2. Internacional.

Situación de la empresa en los últimos 12 meses	1. Nada estable. 2. Un poco estable. 3. Bastante estable. 4. Muy estable. 5. Totalmente estable.

Cuadro 2.1: Variables sociales de nuestro problema.

A continuación, se mostrarán los diversos resultados de las encuestas para cada una de las variables. Sobre las variables a introducir por el usuario (aquellas que no se le ofrecen opciones al usuario), tienen un estudio más complicado puesto que los resultados obtenidos por las mismas varían mucho de un encuestado a otro, no habiendo un mismo patrón a analizar. Por esta razón se han dejado fuera del estudio. Para las variables numéricas también se ha realizado un pequeño análisis estadístico que podrá observarse cuando lleguemos a cada una de estas variables.

2.2.1. Género

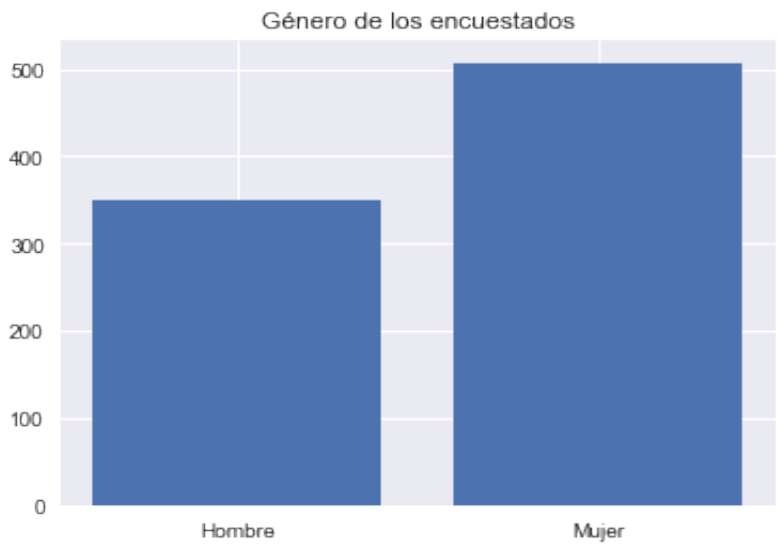

Figura 2.1: Género de los encuestados.

Atendiendo a la variable relacionada con el Género, como puede observarse en la figura 2.1 hay una mayor participación femenina en la encuesta. Aunque esta participación está un poco desbalanceada, dicho desbalance no se considera muy grave ya que dicho porcentaje de participación es de: 40 %

de hombres y 60 % de mujeres. Lo cual no es desproporcionado conforme a nuestra sociedad.

2.2.2. Edad

Primero analizaremos esta variable de forma numérica, para posteriormente mostrar una gráfica por grupos de edad . Se mostrarán los valores más importantes para esta variable, así como un diagrama de caja y bigotes para ver su distribución y rango efectivo.

Mínimo	16
Máximo	76
Media	44.34
Desv. Típica	11.29

Cuadro 2.2: Análisis numérico de la variable edad.

Siendo el boxplot como sigue:

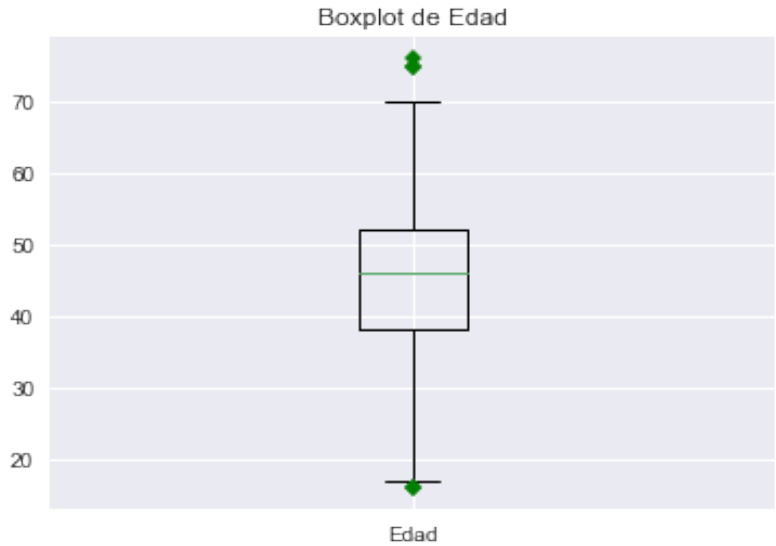

Figura 2.2: Boxplot de Edad.

Como se puede observar en esta tabla 2.2 y boxplot 2.2 mostrado, la mayoría de los encuestados entran entre el rango de 35-55. Siendo la desviación típica de 10 años. Además existen algunos outliers tanto por encima como por debajo, siendo el mínimo 16 años (la edad mínima para completar el test).

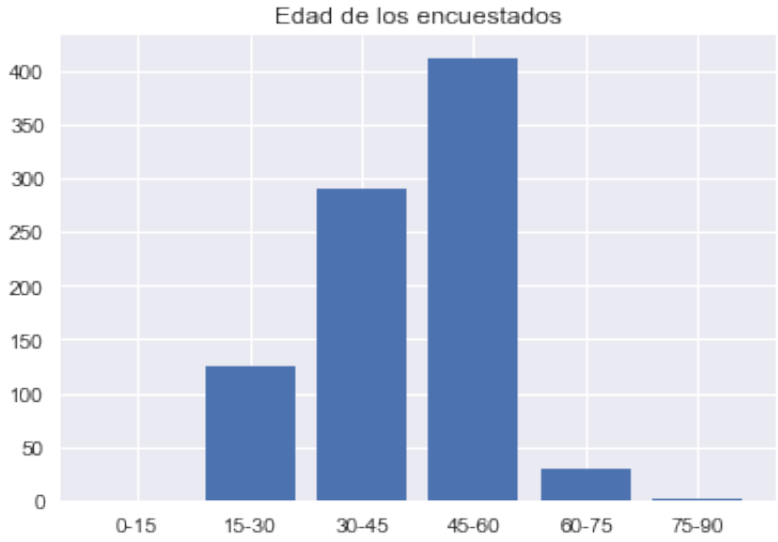

Figura 2.3: Edad de los encuestados.

La edad de los encuestados está mayormente concentrada en una edad media entre 30 y 60 años. Este es un gran rango de edad lo cual nos proporciona encuestados de casi todas las edades. Sin embargo, apenas existe representación para mayores de 60 años. Para menores de 15 no es posible esta representación puesto que se deben de tener 16 años para cumplimentar la encuesta.

2.2.3. Estado civil

La mayoría de los encuestados como puede verse en este gráfico 2.4, pertenecen al grupo de casados, esto mayoritariamente puede deberse a la edad predominante de los encuestados, la cual como recordamos está entre 30 y 60 años. Existe una representación de todos los estados civiles.

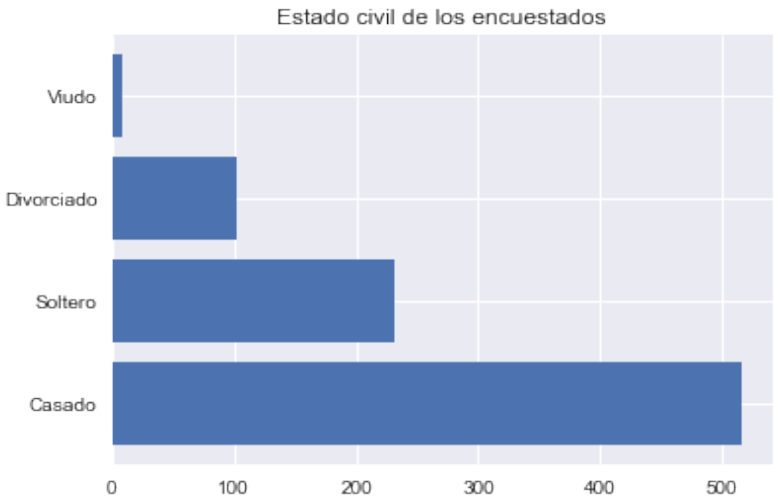

Figura 2.4: Estado civil de los encuestados.

2.2.4. Situación laboral

En la figura 2.5 sí se puede ver una clara diferencia y desbalance entre nuestros encuestados. Puesto que una mayoría aplastante pertenece al grupo de "Empleados", esto puede deberse a la edad de los mismos, viendo que apenas hay jubilados o prejubilados, los cuales serán los representados por una mayor edad. También hay que hacer notar la curiosidad de que no exista ningún encuestado de baja entre casi las 1000 encuestas recogidas. Aunque este desbalance tiene una parte positiva y es la existencia de una parte relacionada sobre la empresa para la que el encuestado trabaja. Puesto que existen bastantes variables a estudiar sobre la situación de la empresa en la que trabajan los encuestados. Dichas preguntas serían imposibles de contestar para las personas que están en situación de desempleo.

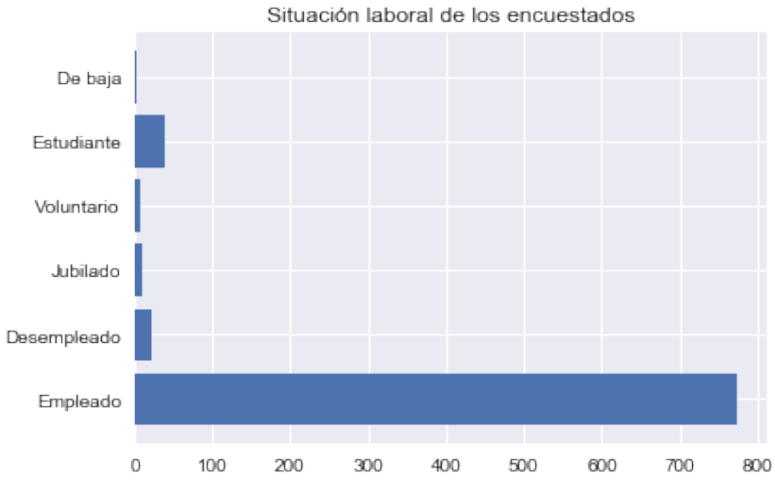

Figura 2.5: Situación laboral de los encuestados.

2.2.5. Formación

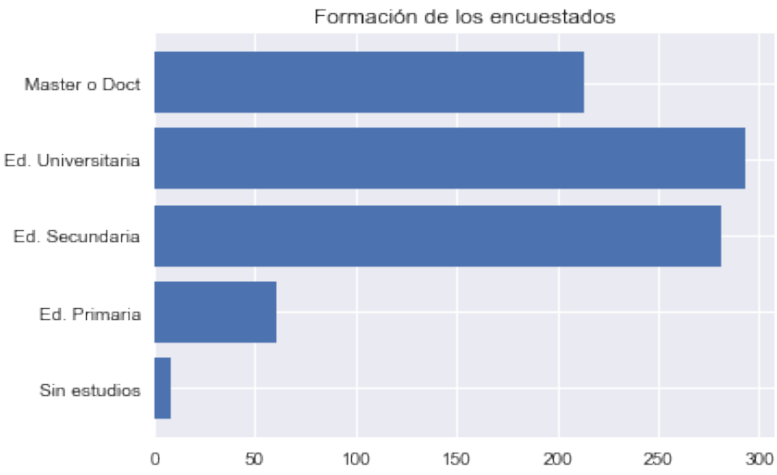

Figura 2.6: Formación de los encuestados.

En esta gráfica 2.6 existe una mayor distribución entre los grupos, aunque puede verse que una mayoría de los encuestados tienen estudios superiores (Al menos graduado universitario). Y que la mayor parte ha completado el recorrido hasta educación secundaria. Lo que vendrá relacionado con el trabajo predominante entre los encuestados.

2.2.6. Antigüedad en la empresa

Para esta variable también mostraremos su análisis numérico, así como un boxplot de la misma.

Mínimo	0
Máximo	58
Media	12.9
Desv. Típica	11.13

Cuadro 2.3: Análisis numérico de la antigüedad en la empresa.

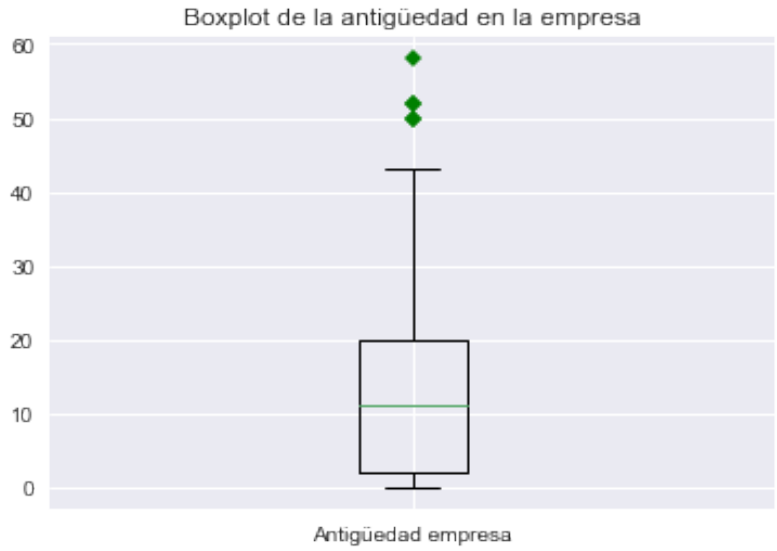

Figura 2.7: Boxplot de antigüedad en la empresa de los encuestados.

Como puede verse en la tabla 2.3 y boxplot 2.7, lo normal para los encuestados en este campo según los datos mostrados será haber estado poco tiempo en la empresa, como mucho 20 años. Sin embargo el número de años en la empresa puede alcanzar hasta los 58 años.

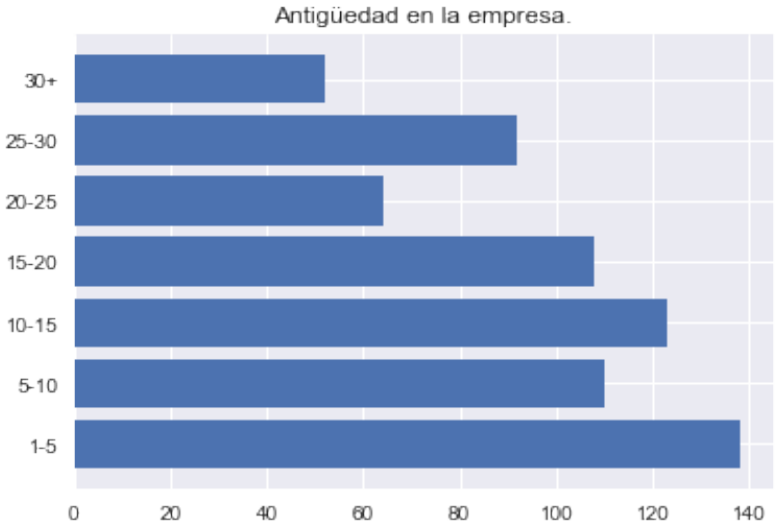

Figura 2.8: Antigüedad en la empresa de los encuestados.

Esta gráfica 2.8 donde se representa la antigüedad en la empresa de los encuestados, la cual está muy relacionada con la edad de los encuestados, puesto que la mayoría de los encuestados estaban en situación laboral. Podemos ver que una gran parte lleva poco tiempo en la empresa, por lo que aún puede estar habituándose a la misma. Esto puede observarse, debido a que el grupo mayoritario en esta gráfica es el perteneciente a personas que llevan de 1 a 5 años en la empresa.

2.2.7. Antigüedad en el cargo

Mínimo	0
Máximo	45
Media	8.98
Desv. Típica	9.03

Cuadro 2.4: Análisis numérico de la antigüedad en el cargo.

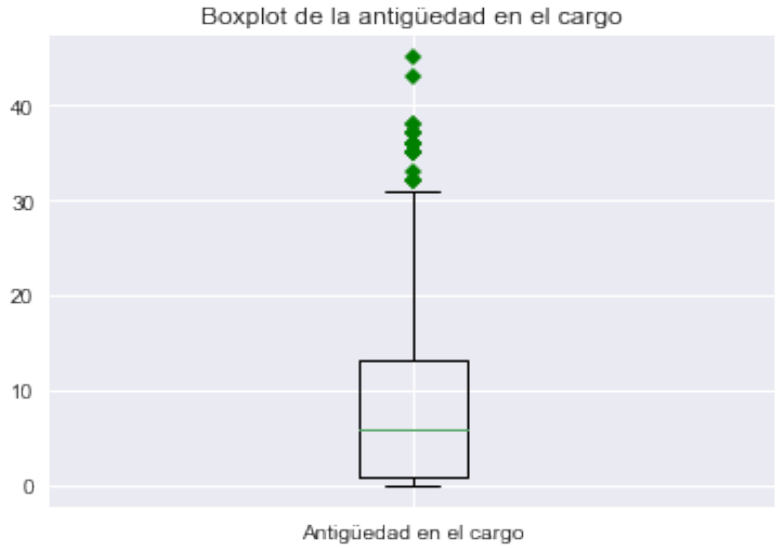

Figura 2.9: Boxplot de antigüedad en el cargo de los encuestados.

Podemos observar en la tabla 2.4 y boxplot 2.9, como ocurría para los valores de los años en la empresa, se ven repetidos para este campo de años en el cargo. Lo normal será que los valores sean bajos, entre 1 y 13 años. Mientras que existirán bastantes encuestados que superen los 30 años en el mismo cargo. Aunque estos sean considerados outliers.

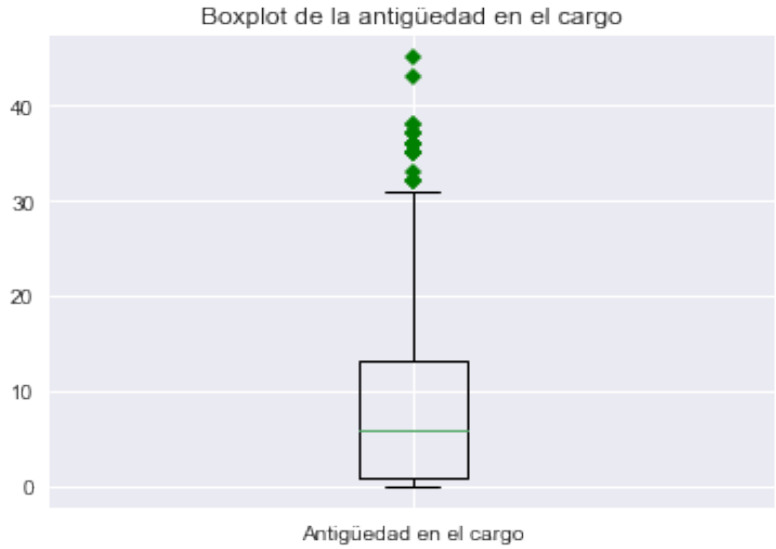

Figura 2.10: Antigüedad en el cargo de los encuestados.

Como vemos en la figura 2.10 la antigüedad en el cargo sí es más variante, pues aún manteniendo la empresa los encuestados han podido cambiar de cargo dentro de la misma empresa, por lo que resulta más difícil acumular años en esta gráfica.

2.2.8. Antigüedad en la profesión

Mínimo	0
Máximo	58
Media	16.4
Desv. Típica	11.25

Cuadro 2.5: Análisis numérico de la antigüedad en la profesión.

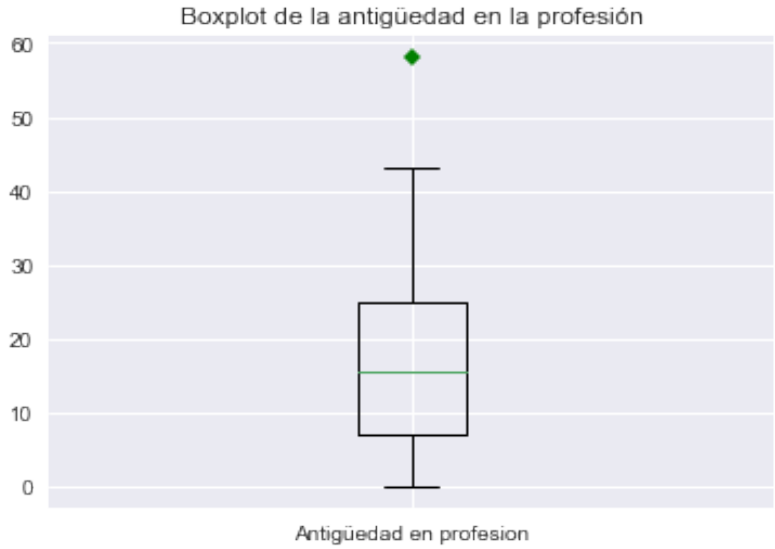

Figura 2.11: Boxplot de antigüedad en profesión de los encuestados.

Como se observa en la tabla 2.4 y boxplot 2.11. Este campo sigue la misma proporción que los anteriores, concentrando su rango y valores entre los 5 y 25 años. Teniendo un outlier que ha estado casi 60 años en la misma profesión. Este campo tiene mayor rango real, puesto que el tiempo que pasa uno en una profesión es mayor que el que se pasa en un cargo, que normalmente va cambiando con los años.

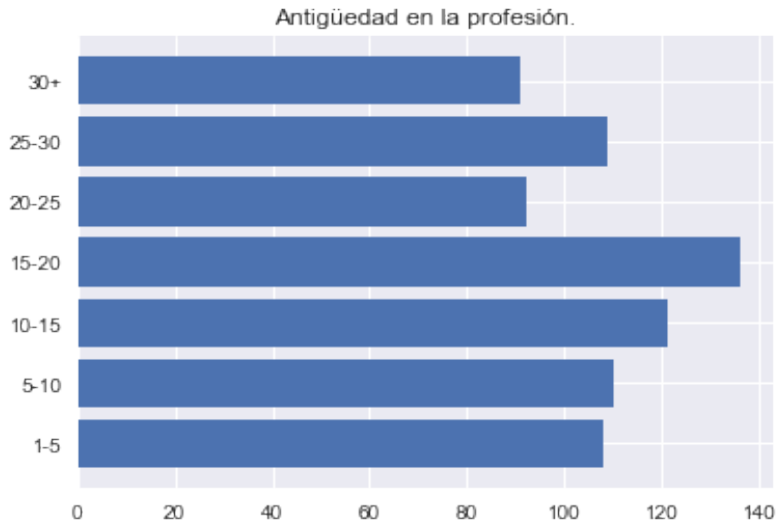

Figura 2.12: Antigüedad en profesión de los encuestados.

La figura 2.12 muestra la antigüedad en la profesión, la cual está ampliamente relacionada con la edad y el término de estudios, puesto que nos muestra el número de años que pasaron entre el término de estudios y la edad de los sujetos (mayormente, puesto que no suele existir cambio de profesión aunque sí de cargo).

2.2.9. Grupo profesional (CIUO)

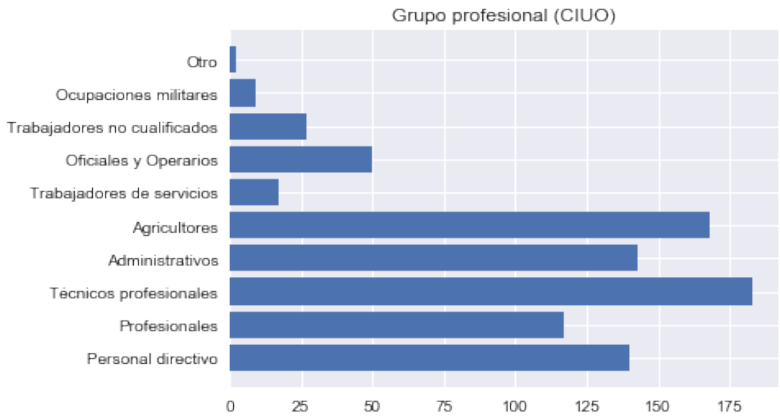

Figura 2.13: Grupo profesional de los encuestados.

En este gráfico 2.13 también existe más igualdad de distribución entre los grupos, siendo los grupos mayoritarios los técnicos profesionales y el sector

de la agricultura. Sorprendentemente, también exista una amplia parte de los encuestados que trabajan en el sector de Personal directivo, esto puede estar relacionado al alto nivel de estudios entre los encuestados. Siendo el grupo minoritario el militar.

2.2.10. Sector

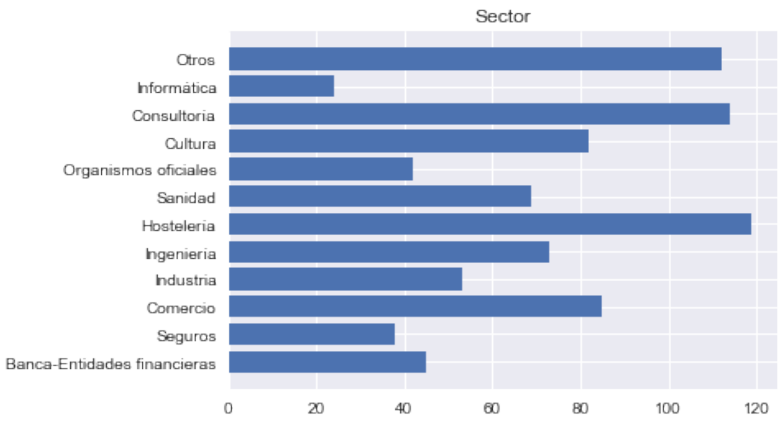

Figura 2.14: Sector de los encuestados.

Algo a destacar sobre esta gráfica 2.14, relacionándolos con el Grupo profesional, es que existe una mayoría de trabajadores que trabajan en el sector de la Hostelería, sin embargo, como se pudo ver anteriormente, el grupo profesional relacionado con trabajadores de servicios era apenas representativo, por lo que existirán muchos encuestados con otro grupo profesional dentro de dicho sector. Siendo el de la hostelería el sector predominante, seguido muy de cerca por el de la consultoría.

2.2.11. Banca Internacional

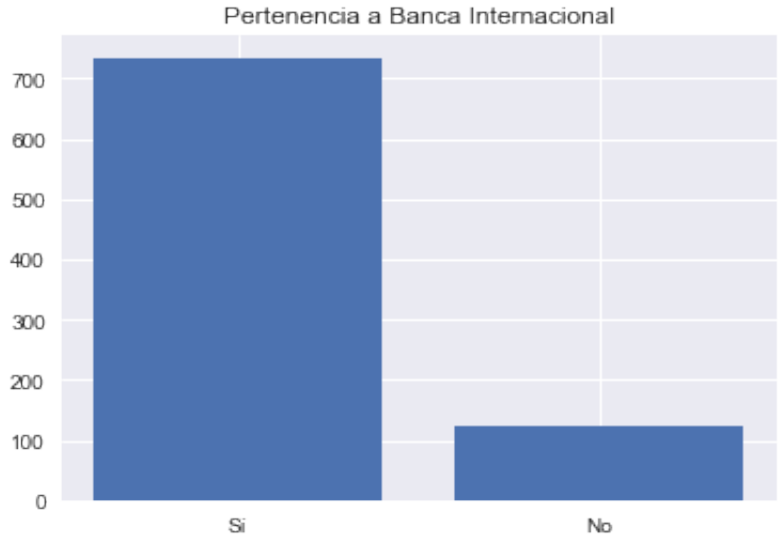

Figura 2.15: Banca Internacional.

Como muestra la figura 2.15 la mayoría de los encuestados pertenecen a empresas que forman parte del programa "Banca Internacional".

2.2.12. Número de empleados en la empresa

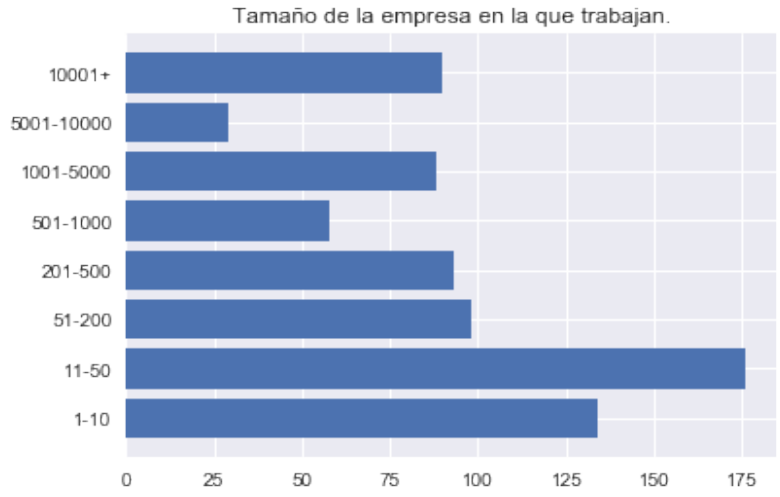

Figura 2.16: Tamaño de la empresa.

Esta gráfica 2.16 nos muestra una distribución bastante uniforme de los encuestados entre empresas de todo tipo de tamaños, predominando aún así las pequeñas y medianas empresas. Sin embargo, todas y cada una de ellas tienen representación en las encuestas.

2.2.13. Ámbito de actuación

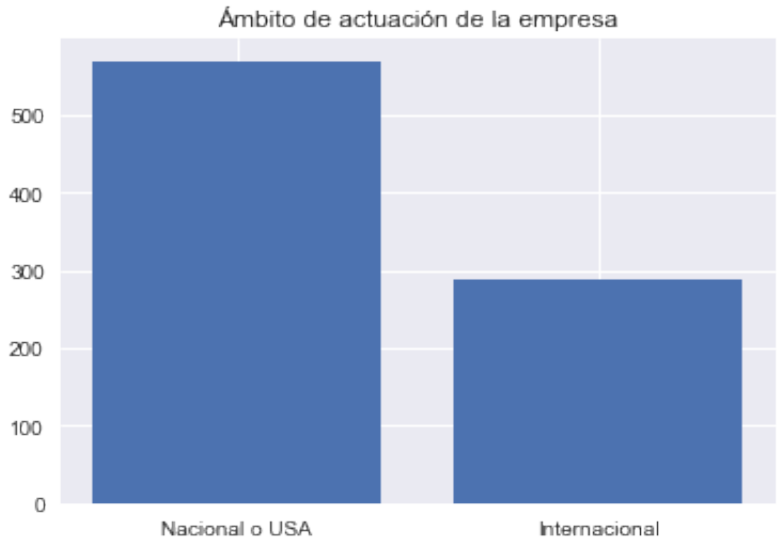

Figura 2.17: Ámbito de actuación de la empresa.

El ámbito de actuación de la mayoría de las empresas cuyos empleados han sido encuestados pertenecen a un ámbito nacional, como muestra la figura 2.17.

2.2.14. Situación de la empresa en los últimos 12 meses

Figura 2.18: Situación en la empresa en el último año.

La situación predominante mostrada es "Bastante estable", como podemos ver en la figura 2.18 y a partir de ahí los grupos van de mayor a menor en buenas condiciones de la empresa. Predominando como vemos los grupos de estabilidad mayor sobre los de una estabilidad menor. En las siguientes secciones comprobaremos si esta situación afecta o no a las cuestiones psicológicas a estudiar.

2.3. Análisis de las variables psicológicas y organizacionales de los encuestados

En esta sección analizaremos las variables psicológicas estudiadas en nuestro problema, para ello primero enumeraremos y describiremos las variables recogidas en la siguiente tabla 2.6:

Atributo	Variables relacionadas
Flow-Fluir	V1
Autoeficacia	V2
Autoestima	V3
Personalidad (OPERAS)	V4 - Extraversión V5 - Estabilidad emocional V6 - Responsabilidad V7 - Amabilidad. V8 - Apertura a la experiencia
Flourishing-Florecimiento	V9
Engagement	V10 - Vigor V11 - Dedicación V12 - Absorción V13 - Engagement total
Satisfacción con la vida	V14
Inteligencia emocional	V15 - Percepción V16 - Comprensión V17 - Regulación V18 - Inteligencia emocional total
Energía (Inteligencia energética)	V19 - Self Drive Direction V20 - Consciencia energética V21 - Energía total
Calidad Personal y Organizacional	V22 - Vitalidad emocional V23 - Estrés organizacional V24 - Estrés emocional V25 - Estrés físico V26 - Abandono V27 - Riesgo salud V28 - Calidad personal y organizacional total

Cuadro 2.6: Variables psicológicas en nuestro problema.

A continuación se describirán en qué consisten estas características psicológicas y los rangos efectivos entre los que se mueven cada una de ellas.

2.3.1. Flow-Fluir

El flow o fluir es un estado de consciencia que se genera en las personas cuando viven experiencias óptimas, normalmente asociadas a la realización de determinadas actividades en las que es necesario involucrar todas las capacidades de atención y habilidad. Dichas experiencias óptimas han sido definidas a través de la investigación científica, como experiencias donde la

idea de fluir es la más ilustrativa del significado descrito por los investigadores, de tal manera que se describe como un estado donde la persona se encuentra completamente absorta en una actividad para su propio placer y disfrute, durante el cual, el tiempo vuela y las acciones, pensamientos y movimientos se suceden a otras sin pausa. Todo el ser está envuelto en esta actividad, y la persona utiliza sus destrezas y habilidades llevándolas hasta el extremo. Tiene que ver con la actividad motivada intrínsecamente, es decir, parte del interior de la persona. Fue formulada por Csikszentmihalyi en 1975[5].

Para medir esta característica se ha usado la escala EBF 9, es una escala abreviada de 9 ítems, uno para cada una de las dimensiones propuestas en la teoría del flow. Las respuestas se presentan en formato Likert con cinco alternativas de respuesta (1 = "Nunca experimento estas sensaciones", 5 = "Siempre experimento estas sensaciones"). Es un estudio realizado con deportistas de distintos niveles de dedicación (Godoy-Izquierdo et al. 2009)[**godoy2009nivel**], los autores obtuvieron una fiabilidad (α de Cronbach) de .76 para la medida de rasgo de la EBF y de .80 para la de estado.

Mínimo	9
Máximo	45
Media	32.61
Desv. Típica	4.91

Cuadro 2.7: Análisis numérico de Flow-Fluir.

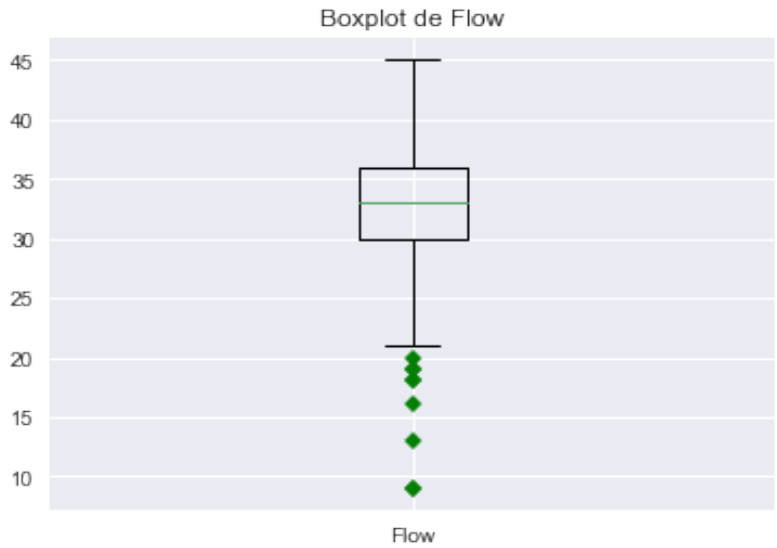

Figura 2.19: Boxplot de Flow-Fluir.

Esta variable tiene un rango formal establecido entre 9 y 45. Esto se ha realizado gracias a 9 preguntas cuya respuesta posible varía del 1 al 5. Sumándose las puntuaciones obtenidas en cada una de ellas.

Como podemos ver en la tabla 2.7 y boxplot 2.19, nuestros datos cumplen con dicho rango. Estando concentrados la mayoría de los encuestados en valores altos, entre 30 y 36. Lo cual es una buena puntuación para este aspecto psicológico, debido a lo dicho anteriormente vemos una concentración en los valores altos para esta variables, sabiendo que a cuanto mayor valor en esta variable más capacidad para Fluir tendremos. Siendo, como se ve en la figura del Boxplot, que lo raro sea tener valores bajos para este carácter psicológico.

2.3.2. Autoeficacia

La **autoeficacia** se define como el sentimiento de confianza en las propias capacidades para poder manejar de forma adecuada los distintos estresores de la vida cotidiana. Fue propuesto por Bandura (1992, 1997)[2]. Tiene que ver con la capacidad de las personas para afrontar las diversas circunstancias de la vida, tomar decisiones y/o realizar actuaciones destinadas al logro de los propósitos que se han planteado, enfrentar obstáculos y/o superar las circunstancias adversas. Involucra la creencia acerca de las propias capacidades para organizar y ejecutar acciones encaminadas a alcanzar resultados. Está relacionada con las capacidades personales para identificar las oportunidades que se tienen alrededor e interpretar los obstáculos que se perciben. A mayor autoeficacia, mayor interés intrínseco por esforzarse y perseverar pese a las dificultades; mayor capacidad para visualizar los problemas como desafíos, oportunidades de aprendizaje y, por ende, de progresar en las propias habilidades.

La escala de Autoeficacia Generalizada (EAG) de Baessler y Schwarzer (1996)[30], se compone de 10 ítems tipo Likert de 4 puntos. Mide el nivel de autoeficacia general, entendida ésta como un constructo global que hace referencia a la creencia estable de las personas sobre su capacidad para manejar adecuadamente una amplia gama de estresores de la vida cotidiana (Bandura, 1992)[38]. La adaptación para la población española realizada por Sanjúan, Pérez y Bermúdez (2000)[27] mostró una consistencia interna de 0.87 y una correlación dos mitades de 0.88.

Mínimo	10
Máximo	40
Media	31.91
Desv. Típica	5.22

Cuadro 2.8: Análisis numérico de Autoeficacia.

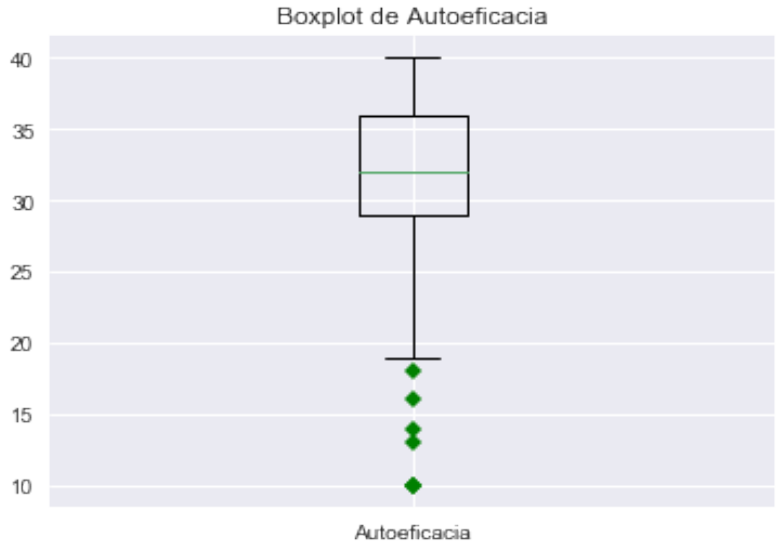

Figura 2.20: Boxplot de Autoeficacia.

El rango formal y establecido para esta variables es de 10 a 40. Creándose dicho rango gracias a 10 preguntas cuyos valores pueden oscilar entre 1 y 4 puntos. Siendo el total la suma de todas estas preguntas.

En nuestros datos 2.8 y 2.20, dicha variable también oscila entre 10 y 40. Estando la mayoría de los encuestados en un rango de 31+-5. Siguiendo la proporción de la variable anterior es lo mayoritario es tener un rango alto (una buena autoeficacia) en esta variable y lo raro tener valores muy bajos, considerando incluso algunos como outliers.

2.3.3. Autoestima

La autoestima supone una evaluación global de la consideración positiva o negativa de uno mismo. Es sólo un componente del más amplio concepto de sí mismo, que Rosenberg define como "la totalidad de los pensamientos y sentimientos de la persona con referencia a sí mismo como objeto". Además de la autoestima, la autoeficacia y la auto-intensidad son una parte importante del concepto de sí mismo. En las últimas décadas, numerosos estudios han constatado la influencia de la autoestima en el bienestar psicosocial, señalando su importancia como recurso personal capaz de reducir los efectos potencialmente negativos de los eventos vitales estresantes.

Uno de los instrumentos más ampliamente utilizados en investigaciones psicológicas y psiquiátricas es la escala de autoestima de Rosenberg (1965)[25], habiendo sido traducida a 28 idiomas, y validada interculturalmente en más de 50 países. La muestra original en la que se desarrolló la

escala en la década de 1960 consistió en 524 estudiantes de tercer y cuarto año de secundaria de 10 escuelas seleccionadas al azar en el estado de Nueva York. La escala mostró una alta fiabilidad: correlaciones test-retest en el rango de 0.82 hasta 0.88.

Mínimo	10
Máximo	40
Media	33.25
Desv. Típica	4.43

Cuadro 2.9: Análisis numérico de Autoestima.

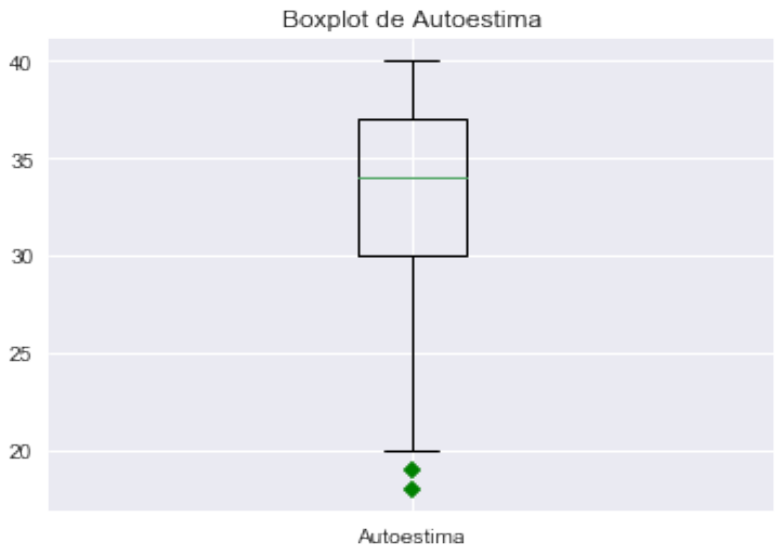

Figura 2.21: Boxplot de Autoestima.

La escala de Roesenberg para la autoestima cuenta con 10 preguntas. Pudiendo obtener cada una de ellas entre 1 y 4 puntos. Siendo el total la suma de ellas. El rango formal de esta escala va desde 10 a 40 puntos. Sin embargo, una puntuación por debajo de los 25 puntos es ya preocupante, puesto que muestra que el sujeto tiene una autoestima baja.

Para esta variable de autoestima, como podemos observar en la tabla 2.9 y boxplot 2.21, el rango en nuestros datos estará entre 10 y 40. Estando la mayoría de los encuestados en un rango cercano a 33, valor relativamente alto siendo lo raro tener valores bajos en esta variable. La mayoría de nuestros encuestados consideran que tienen una autoestima elevada, estando esta entre los 30 y los 40 puntos.

2.3.4. Personalidad-OPEARS

El test *Overall Personality Assessment Scale* (OPERAS)[36] es un cuestionario basado en el modelo de los cinco grandes factores de personalidad. Según este modelo, la conducta humana depende principalmente de cinco rasgos de personalidad: extraversión, responsabilidad, estabilidad emocional, amabilidad y apertura a la experiencia:

- La escala **Extraversión** evalúa el nivel de sociabilidad de una persona, incluyendo características como la tendencia a ser hablador, la preferencia por las situaciones sociales animadas, etc.

- La escala **Estabilidad emocional** evalúa, tal como su nombre indica, la estabilidad emocional de una persona, que se caracteriza por la tendencia a sentir calma, con poca propensión a sentir emociones negativas tales como la ansiedad, la inseguridad, la tristeza o el miedo.

- La escala **Responsabilidad** evalúa el grado de responsabilidad de las personas, e incluye características como la planificación, la organización, la meticulosidad y la eficacia.

- La escala **Amabilidad** evalúa la tendencia a ser amable y a considerar los sentimientos y los derechos ajenos, incluyendo características como la empatía, la cooperación, la franqueza, la confianza en los demás, etc.

- Finalmente, la escala **Apertura a la Experiencia** evalúa la disposición de una persona a conocer otras formas de pensar y el interés por vivir nuevas experiencias, incluyendo características tales como la imaginación, la curiosidad, el interés por la cultura y el arte, etc.

El test contiene un total de 40 ítems. En relación a las propiedades psicométricas, los resultados muestran un buen ajuste del test a la estructura de 5 factores. Las fiabilidades de los factores son: 0.89 para Extraversión, 0.84 para Estabilidad Emocional, 0.81 para Responsabilidad, 0.73 para Amabilidad y 0.84 para Apertura a la Experiencia.

	V4 (Extraversión)	**V5 (Estabilidad emocional)**	**V6 (Responsabilidad)**	**V7 (Amabilidad)**	**V8 (Apertura a la experiencia)**
Mínimo	13	-4	13	13	5
Máximo	77	77	74	89	72
Media	49.83	50.1612	49.87	52	51.80
Desv. Estándar	10.11	10.53	9.84	10.73	10.11

Cuadro 2.10: Análisis numérico de OPERAS.

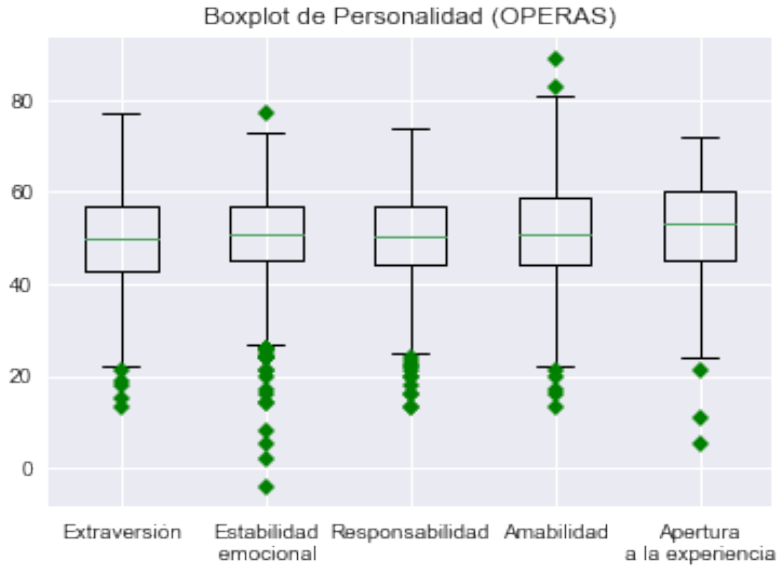

Figura 2.22: Boxplot de OPERAS.

Estas variables son las que más error o problema hemos tenido al analizar, puesto que el rango real para ellas que nos dieron junto con los datos prestados es el siguiente.

- Extraversión: 7 a 35
- Estabilidad emocional: 7 a 35
- Responsabilidad: 7 a 35
- Amabilidad: 7 a 35
- Apertura a la experiencia: 7 a 35

Observando la tabla 2.10 y el boxplot 2.22. Podemos ver como el rango obtenido por nuestro análisis no concuerda para nada con el rango formal que debería de tener estas variables. No sabemos si esto se tratan de outliers o el rango formal estaba mal definido. Puesto que observando el boxplot, la mayoría de las variables están concentradas en el intervalo de 40 a 60, el cual ni siquiere pertenece al rango formal. A continuación pasaremos a analizar las variables una a una. La variable **V4** relacionada con la **extraversión**, puede verse que oscila entre los valores 13 y 77, estando la mayoría de los valores entre 39 y 59. Lo cual son valores más pequeños y más cercanos a un valor central distinguiéndose de las variables anteriores. La variable **V5** que representa la **estabilidad emocional** es la única variable que tiene como mínimo un número negativo, no sabiendo si se trata de un error o realmente

se ha alcanzado dicha puntuación para este valor. Viendo que existen bastantes encuestados con valores muy bajos y cercanos al 0. Aún siendo su rango de -4 a 77, la mayoría de los encuestados estarán concentrados en torno a la media de 50, siendo este un valor bastante alto con una oscilación de 10 puntos. La variable **V6** que representa la **responsabilidad** oscila entre los valores de 13 a 74, estando concentrados la mayoría de los encuestados en torno a la media de 50 con una oscilación menor a 10 puntos, siendo esta la menor oscilación entre las variables vistas en este apartado, aunque se debe a que su rango también es menor. También existen un pequeño número de encuestados concentrados en torno al mínimo. La variable **V7** que representa la **amabilidad**, oscila su rango entre los valores 13 y 89. Teniendo a la mayoría de encuestados en torno a 52 con 10 puntos de oscilación. En esta variable los outliers mínimos no se acercan tanto al mínimo presentado. La variable **V8** que presenta la **apertura a la experiencia**, es la que menos outliers presenta, teniendo su rango entre 5 y 72, concentrando así como todas estas variables a su mayoría entre la centralidad alta de 52 con un rango de 10 puntos. Es la variable con un menor número de outliers mínimos y ninguno máximo.

2.3.5. Flourishing-Florecimiento

El "Flourishing" o Crecimiento Personal se refiere a las evaluaciones positivas que la persona realiza sobre determinados aspectos de su vida. El florecimiento mide la valoración que hacen las personas acerca de sus relaciones sociales, propósito y sentido de vida, auto-eficacia y auto-respeto. Tiene que ver con el sentimiento de felicidad y bienestar que la persona posee. La tesis de Diener y Biswas-Diener (2008)[15] se concreta en la siguiente fórmula de la felicidad:

Bienestar subjetivo = estados afectivos positivos - estados afectivos negativos + satisfacción vital + crecimiento personal ("flourishing")

Por tanto para estos autores los componentes esenciales para llevar una buena vida son tener una vida llena de emociones positivas (de todo tipo), emociones negativas ocasionales, y que gire en torno a actividades en las que disfrutemos y que estén guiadas hacia el logro de ciertos valores que nos planteamos como metas y que le dan sentido al esfuerzo que nos exigen.

La escala de florecimiento fue creada para evaluar el éxito auto percibido en las áreas tales como las relaciones sociales, sentimientos de competencia, propósito y optimismo. Está compuesta por ocho items en un formato de respuesta tipo Likert, con un rango de 1 (completamente en desacuerdo) a 7 (completamente de acuerdo). La escala provee una calificación única de bienestar psicológico y tiene comprobadas propiedades psicométricas: buen nivel de consistencia interna (α de Cronbach 0.87), y la confiabilidad tempo-

ral de la escala es moderadamente alta (0.71). Igualmente, está fuertemente correlacionada con los resultados en otras escalas de bienestar psicológico y sentimientos (Diener et al., 2009)[6].

Mínimo	5
Máximo	35
Media	29.94
Desv. Típica	4.12

Cuadro 2.11: Análisis numérico de Flourishing-Florecimiento.

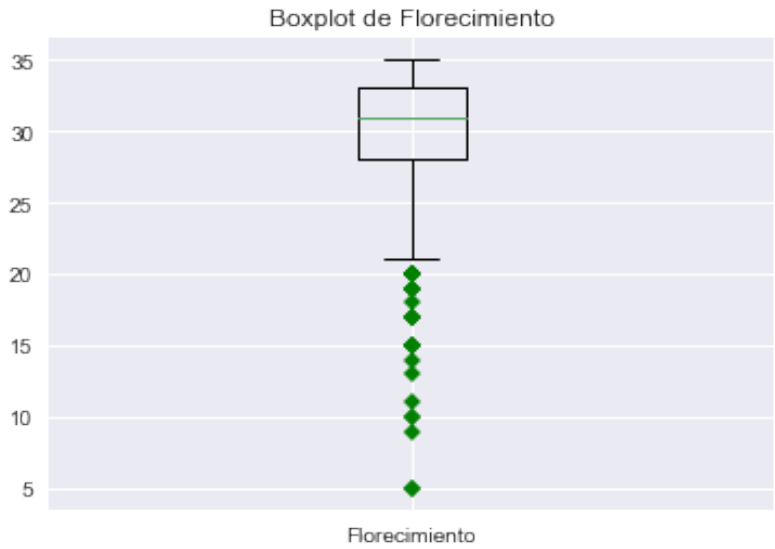

Figura 2.23: Boxplot de Flourishing-Florecimiento.

Según los datos prestados, esta variable debería de tener un rango real de 8 a 56. Sin embargo, dista bastante del rango obtenido en nuestros análisis. Observando la tabla 2.11 y el boxplot 2.23 Esta variable toma valores entre 5 y 35, estando la mayoría de los encuestados entorno a 30, con una oscilación de 4 puntos. Siendo el valor de 30 bastante alto, por ello existen un pequeño número de encuestados que aunque su valor sea alto, en torno a 20, son considerados outliers mínimos. Esto se debe a la gran cantidad de instancias situadas en torno a un valor de 30.

2.3.6. Engagement

El engagement en el trabajo es un constructo motivacional positivo relacionado con el trabajo, se caracteriza por el vigor, la dedicación y la ab-

sorción (Schaufeli, Salanova, González-Romá & Bakker, 2002)[29]:

- El **vigor** implica altos niveles de energía, resistencia y activación mental mientras se trabaja, deseo y predisposición a esforzarse, invertir esfuerzo en el trabajo que se está realizando, y persistencia incluso cuando se presentan dificultades en el camino.

- La **dedicación** se refiere a una alta implicación laboral junto con la manifestación de un sentimiento de significación, entusiasmo, inspiración, orgullo y reto por el trabajo. Es decir, involucrarse, entusiasmarse, estar orgulloso e inspirado en el trabajo, así como poseer un sentimiento de importancia y desafío.

- La **absorción** se produce cuando la persona está totalmente concentrada en su trabajo, mientras se experimenta que el tiempo le pasa volando y se tienen dificultades a la hora de desconectar de lo que se está haciendo debido a las fuertes dosis de disfrute y concentración experimentadas.

El **engagement**, más que un estado específico y momentáneo es un estado afectivo-cognitivo (de emoción y pensamiento) más persistente y no focalizado en un objeto, evento o situación particular. Las personas con engagement se sienten enérgica y eficazmente unidas a sus actividades laborales y capaces de responder a las demandas de su puesto con absoluta eficacia. Schaufeli, Bakker y Salanova (2006)[28] utilizaron los datos recopilados en 10 países ($n = 14521$) para desarrollar esta versión reducida de la escala, que consta de nueve ítems agrupados en tres factores: tres ítems de vigor, tres ítems de dedicación y tres ítems de absorción (UWES-9, por sus siglas en inglés), con una escala tipo Likert de siete anclajes los cuales comienzan desde nunca o ninguna vez, hasta siempre o todos los días. Los índice de fiabilidad α de Cronbach son los siguientes: vigor ($\alpha = 0.82$), dedicación ($\alpha = 0.86$) y absorción ($\alpha = 0.8$).

	V10 (Vigor)	**V11 (Dedicación)**	**V12 (Absorción)**	**V13 (Engagement total)**
Mínimo	3	3	3	3
Máximo	21	21	21	63
Media	15.84	15.71	15.77	47.33
Desv. Estándar	3.81	4.45	4.18	11.52

Cuadro 2.12: Analisis numérico de Engagement.

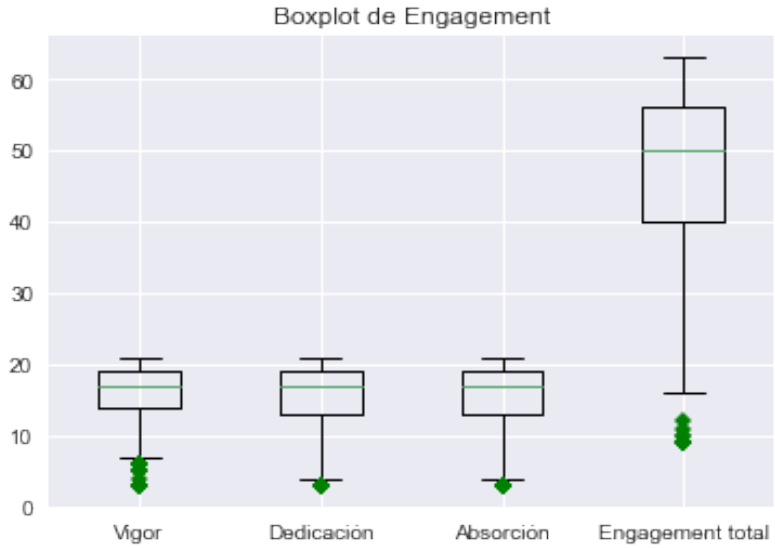

Figura 2.24: Boxplot de Engagement.

Los rangos reales de todas estas variables serán de 3 a 21. Siendo el Engagement total la suma de todas ellas, obteniendo por lo tanto un rango real de 9 a 63. Compuesta cada una de las variables por 3 afirmaciones las cuales pueden obtener una puntuación del 1 al 7. Siendo el total de cada variable la suma de las mismas.

Para estas variables psicológicas, como puede observarse en la tabla 2.12 y boxplot 2.24, **V13** será el total y la suma de las anteriores. Por lo tanto, analizaremos ésta directamente. Su valor oscila entre 9 y 63, siendo el rango efectivo de la misma bastante alto, teniendo una media con un valor cercano a 48 y con una oscilación de 11 puntos. Teniendo esta variable únicamente outliers mínimos, significando esto que la mayoría de la población obtuvo un alto nivel de puntuación para la misma. Las otras variables **V10 (Vigor)**, **V11 (Dedicación)** y **V12 (Absorción)**, siguen la misma proporción que la variable total analizada.

2.3.7. Satisfacción con la vida

La satisfacción con la vida es un juicio cognitivo, basado en la comparación que la persona realiza entre su propia vida y unos estándares autoimpuestos, que conllevan una evaluación global de la vida. Se trata de la percepción personal sobre la propia situación en la vida partiendo de los propios objetivos, expectativas, valores e intereses influenciados por el contexto cultural de referencia. Se relaciona con un sentimiento personal de bienestar o de felicidad. Está muy relacionada con los logros personales y

con circunstancias sociales como el nivel educativo personal, la riqueza de la nación o la riqueza personal.

La escala de satisfacción con la vida (SWLS; Diener, Emmons, Larsen y Griffin, 1985)[7] está formada por un conjunto de 5 afirmaciones frente a las cuales las personas deben indicar su nivel de acuerdo o desacuerdo en una escala tipo Likert de 7 puntos. Las buenas propiedades psicométricas de la SWLS han sido confirmadas durante los últimos veinte años. Se ha demostrado una alta consistencia interna de la escala, con coeficientes α de Cronbach oscilando de 0.79 a 0.89 (Blais, Vallerand, Pelletier, y Briere, 1989; Diener, et al. 1985; Pavot, Diener, Colvin y Sandvik, 1991)[34, 17, 23].

Mínimo	5
Máximo	35
Media	26.55
Desv. Típica	5.7

Cuadro 2.13: Análisis numérico de Satisfacción con la vida.

Figura 2.25: Boxplot de Satisfacción con la vida.

El rango real de esta variable es de 5 a 35. Puesto que está compuesta por 5 afirmaciones de 7 puntos cada una de ellas. Esta variable **V14 (Satisfacción con la vida)** como se puede observar en la tabla 2.13 y boxplot 2.25, oscila en nuestros datos entre los valores 5 y 35, teniendo la mayoría de los valores en torno al valor 26, con una oscilación de 5 puntos. Como se ha venido viendo en las variables anteriores, solo existen outliers marcados

como mínimos, por debajo en este caso del valor 15.

2.3.8. Inteligencia emocional

La inteligencia emocional (IE) es una parte de una inteligencia social que incluye la capacidad para controlar nuestras emociones y las de los demás, de discriminar entre ellas y de usar esta información para guiar nuestros comportamientos y pensamientos. Tiene que ver con las destrezas con las que podemos ser conscientes de nuestras propias emociones así como, de nuestra capacidad para regularlas. Elementos básicos de una persona emocionalmente inteligente son la capacidad para percibir, asimilar, comprender y regular las emociones propias y la de los demás (Mayer y Salovey, 1997)[19]. Está ampliamente relacionada con la capacidad para influir positivamente en otras personas, esencia de liderazgo. Muchos autores han estudiado esta relación, entre ellos Richard Boyatzis, que habla del "Líder Resonante" como la persona que es capaz de sintonizar con los sentimientos de las personas y los encauza en una dirección emocionalmente positiva, se conecta con sus interlocutores llegando incluso a transformar positivamente su estado de ánimo en un momento complicado. Su equipo exhibe optimismo y entusiasmo. Gestiona y canaliza las emociones para que el grupo alcance sus objetivos y alienta en sus subordinados cierto grado de bienestar que les lleva a compartir las ideas y aprender los unos de los otros. Ejerce empatía y ayuda a aprovechar al máximo los recursos disponibles para una mayor eficiencia en la organización.

La "Spanish Modified Trait Meta-Mood Scale-24" (TMMS-24) es una versión reducida del TMMS-48, realizada por el grupo de investigación de Málaga (Fernández-Berrocal, Extremera y Ramos, 2004)[8]. Está basada en Train Meta-Mood Scale-48 (TMMS) del grupo de investigación de Salovey y Mayer. Contiene tres dimensiones claves con 8 items cada una de ellas: Atención Emocional, Claridad de Sentimientos y Reparación Emocional. La Atención emocional expresa la capacidad de sentir y expresar los sentimientos de forma adecuada. La Claridad de Sentimientos el grado de comprensión de los estados emocionales y la Reparación Emocional la capacidad de regular los estados emocionales correctamente. Su fiabilidad para cada componente es: Atención (0.9), Claridad (0.9) y Reparación (0.86). Asimismo, presenta una fiabilidad test-retest adecuada.

	V15 (Percepción)	**V16 (Comprensión)**	**V17 (Regulación)**	**V18 (Inteligencia emocional total)**
Mínimo	8	8	8	24
Máximo	40	40	40	120
Media	26.22	28.90	28.99	84.13
Desv. Estándar	6.76	6.33	6.34	14.79

Cuadro 2.14: Análisis numérico de Inteligencia emocional.

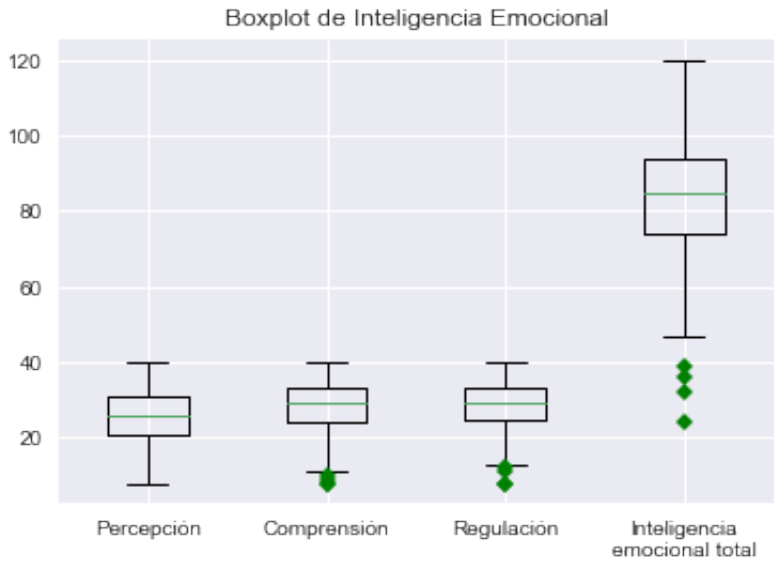

Figura 2.26: Boxplot de Inteligencia emocional.

Los rangos reales de cada una de las variables de la Inteligencia emocional, oscilan entre 8 y 40. Siendo la variable Inteligencia Emocional total la suma de todas ellas, obteniendo por lo tanto un rango real de 24 a 120. Para calcular cada una de las variables se han realizado 8 preguntas en las cuales se puede obtener una puntuación de 1 a 5.

Mirando detenidamente la tabla 2.14 y boxplot 2.26. En este conjunto de variables establecidas para la Inteligencia Emocional, también hablaremos únicamente de la variable **V18** (Inteligencia emocional total), puesto que representa el total y suma de las tres anteriores. En esta variable el rango oscila entre 24 y 120 puntos. Siendo la media del mismo de 84 puntos y teniendo una oscilación bastante alta, debido a su rango, de 15 puntos. Los valores que toman esta variable si están más centrados que las variables vistas anteriormente. Teniendo los outliers nuevamente en forma de outliers mínimos.

2.3.9. Energía (Inteligencia energética)

La inteligencia Energética es un nuevo constructo que se refiere a la capacidad de la persona para identificar las energías que habitan dentro y fuera de sí mismo, distinguirlas unas de otras y utilizar esa información para la consecución de objetivos propios y colectivos.

La inteligencia Energética es el resultado continuo, holístico e indivisible de la actuación de 5 factores: 4 inteligencias y un factor adicional, la consciencia energética. Podemos definirlos de la siguiente manera:

- **Inteligencia Corporal y de Movimiento:** Alude a la capacidad de la persona de reconocer, atender, dar respuesta y utilizar de forma consciente y constructiva la información que recibe de su propio cuerpo y del movimiento que este experimenta en cada momento (autores que hablan de la importancia del cuerpo y el movimiento en la capacidad del ser humano para crecer, evolucionar, desarrollar su talento y potencial, entre otros son: Wallon, Ajuriaguerra, Alcaide, Berruezo, Camps, Cuddy, Eckman, Bloch, Trigo, Duch y Mèlich, y Maturana).

- **Inteligencia Emocional:** Alude a la capacidad de las personas para reconocer sus propias emociones y las ajenas, distinguirlas unas de otras y poder utilizar esa información para la consecución de objetivos (autores que hablan de la importancia de la inteligencia emocional en la capacidad del ser humano para crecer, evolucionar, desarrollar su talento y potencial, entre otros son: Bar On, Mayer, Salovey, Goleman, Boyatzis, McKeen y Schreiber).

- **Inteligencia Lingüístico-Conceptual:** Alude a la capacidad de la persona para crear realidad presente y futura a través de su lenguaje y pensamiento. Es una competencia relacionada con el uso eficaz del lenguaje para crear contextos enriquecedores, positivos (internos y externos), en los que la comunicación se canalice hacia la consecución de logros, resolución de problemas, creación del bienestar y progreso (autores que hablan de la importancia del lenguaje y el pensamiento en la capacidad del ser humano para crecer, evolucionar, desarrollar su talento y potencial, entre otros son: Wittgenstein, Austin, Searle, Maturana, Flores, Chomsky, Echeverría y Olalla).

- **Inteligencia Transitiva-Espiritual:** Alude a la capacidad de toda persona para encontrar el sentido de su existencia, su "para qué" en la vida y en cada uno de los sistemas en los que habita. Tiene que ver con la capacidad para trascender la propia individualidad y ponerse al servicio del bien común o de aquello que es más grande que uno (autores que habla de la importancia de lo social, transitivo, sistémico y espiritual en la capacidad del ser humano para crecer, evolucionar,

desarrollar su talento y potencial, entre otros son: Wallon, Rogers, Csikzentmihalyi, Diener, Collings, Belf, Hellinger y Paoli).

- **Consciencia Energética:** Alude alude a la capacidad de toda persona para ser consciente de su dimensión energética como ser humano y de su capacidad para regular esta energía y utilizarla para vivir una vida con mayor sentido y bienestar (individual y colectivamente), lo que supone ser capaz de crear entornos (internos y externos) energéticamente limpios en los que la energía fluye para dar mejores resultados con menor nivel de esfuerzo (autores que hablan de la importancia de la regulación-conexión energética en la capacidad del ser humano para crecer, evolucionar, desarrollar su talento y potencial, entre otros son: Rogers, Yela, Csikszentmihalyi, Rubira, McCraty, Pérez López, Locke y Latham, Arranz Basagoiti, Diener, Viswas, Rong su, Jackson y Ekson, Ryff, De la Torre y Moraes, Paoli, Stelling, Moraes, R. Cohen, Renzulli, Torre, Ruiz, Barbarro, Carroll y Chapman).

	V19 (Self Drive Direction)	**V20 (Consciencia Energética)**	**V21 (Energía total)**
Mínimo	15	23	38
Máximo	75	115	190
Media	55.28	88.72	142.01
Desv. Estándar	10.56	13.49	22.40

Cuadro 2.15: Análisis numérico de Inteligencia energética.

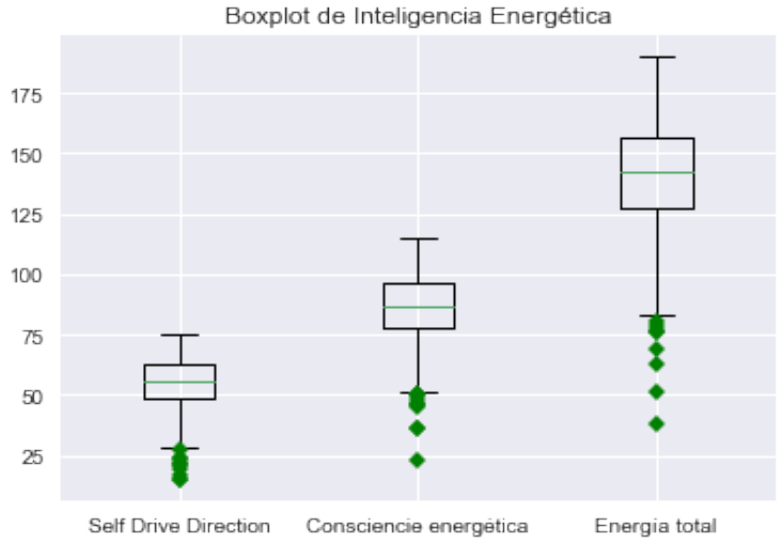

Figura 2.27: Boxplot de Inteligencia energética.

El rango real de **Self Drive Direction** es de 15 a 75, mientras que el rango de la **Consciencia energética** será de 23 a 115. Siendo la energía total la suma de ambas variables, obteniendo así un rango real de 38 a 190. Como puede verse en la tabla 2.15 y boxplot 2.27, dichos rangos reales son seguidos por el análisis llevado a cabo. Para estas variables representando la energía e inteligencia energética también analizaremos la última variable de la misma **V21** (Energía total). Su rango está entre los valores 38 y 190, siendo de los rangos más amplios vistos hasta ahora, es por ello que la oscilación que existe entre su rango efectivo es bastante amplio, llegando a los 20 puntos. Oscilando alrededor de los 140 puntos para esta variable. Los valores se encuentran concentrados en la parte alta de nuestra gráfica, cerca de los valores máximos, mientras que los valores de los encuestados cercanos al mínimo serán considerados como outliers. Sin embargo, serán estos los interesantes a analizar para esta variable.

2.3.10. Calidad personal y Organizacional

La **Valoración de la Calidad Personal y Organizacional** proporciona un perfil de calidad del ambiente de trabajo de su organización en términos de los principales factores que mejoran o disminuyen el desempeño laboral efectivo.

El VCPO-R4 ofrece información sobre cuatro factores: la vitalidad emocional, el estrés emocional, el estrés organizacional y el estrés físico. Cada uno de los cuatro factores forma parte de una escala sólida, estadísticamente

válida y fiable, y afectan directamente al desempeño laboral de los empleados.

La **Vitalidad Emocional** tiene un efecto positivo o de mejora en la salud, el bienestar y el rendimiento, y los otros tres factores **-Estrés organizacional, Estrés Emocional y Estrés físico-** tienen efectos negativos o nocivos sobre la salud, el bienestar y el desempeño.

Factores positivos:

1. **Escala de Vitalidad emocional:** Energía emocional positiva que enriquece la experiencia de vida y mejora la salud y el bienestar. Esta escala es una medida general del grado en que usted siente una energía emocional positiva que permite una experiencia de vida optimista y satisfactoria. Las puntuaciones bajas en esta escala sugieren que es probable que tenga bajos niveles de vitalidad emocional, y, por lo tanto, pueden tener energía emocional limitada disponible para invertir.

   a) **Subescala de Fortaleza emocional:** Las puntuaciones bajas en esta subescala indican que es probable que usted sienta que tienen bajos niveles de energía emocional disponibles para invertir en su trabajo y en su vida personal.

   b) **Subescala de Satisfacción emocional:** Las puntuaciones bajas en esta subescala sugieren que usted puede estar sintiendo bajos niveles de satisfacción y paz interior en relación con su vida, tanto en su trabajo como fuera de él.

Factores negativos:

1. **Escala de estrés organizacional:** Muestra impedimentos organizacionales y discordia relacional que perjudican el desempeño laboral, reducen la satisfacción laboral y aumentan la rotación.

   a) **Presiones de vida:** Las puntuaciones altas en esta subescala indican que puede sentirse abrumado por la presión del trabajo y por el resto de demandas de tiempo y energía en su vida personal.

   b) **Tensión relacional:** Las puntuaciones altas en esta subescala indican que puede sentirse presionado por la baja calidad de las relaciones y el conflicto con compañeros de trabajo, lo que crea un entorno estresante en el lugar de trabajo.

   c) **Estrés:** Las puntuaciones altas en esta subescala sugieren que puede sentirse abrumado por varias fuentes de estrés en el global de su vida.

2. **Escala de estrés emocional:** Discordia emocional que reduce la calidad de la experiencia de vida y pone en peligro la salud y el bienestar.

3. **Estrés físico:** Síntomas físicos de fatiga y mala salud que reflejan el estrés general que experimenta un profesional.

   a) **Fatiga:** Las puntuaciones altas en esta subescala indican que puede experimentar cansancio, fatiga y agotamiento físico.

   b) **Síntomas de salud:** Las puntuaciones altas en esta subescala indican que puede experimentar tensión física, achaques y dolores en general, malestar estomacal, aceleración del ritmo cardiaco y dolores de cabeza.

4. **Intención de abandono:** Las puntuaciones altas en esta subescala deben ser una señal de alarma para la gerencia. Indican que una proporción notables de empleados se sienten suficientemente insatisfechos con su ambiente de trabajo como para pensar en dejar la organización.

La integridad psicométrica de la reorganización de los 49 ítems en estos cuatro factores se verificó empíricamente mediante una validez y confiabilidad del estudio de medición realizado en la base de datos POQA-R existente de 5971 trabajadores de la salud. Se realizaron dos análisis estadísticos para verificar la validez y fiabilidad de las mediciones. En el primero, las siete escalas y sus subescalas asociadas fueron sometidas a un análisis de la consistencia interna de la medición utilizando el coeficiente α de Cronbach. Los resultados de las cuatro escalas primarias mostraron que todos los constructos superaron el criterio de adecuación técnica ($\alpha > 0.75$): los coeficientes α oscilaron entre 0.76, para el estrés organizacional, y 0.92 para la vitalidad emocional y el estrés emocional. Los resultados obtenidos para las nueve subescalas multi-items mostraron que, con una medición técnicamente adecuada. En las otras ocho subescalas multi-item el coeficiente α osciló entre 0.76, para síntomas de salud, a 0.9, para fortaleza emocional, intención de dejar la organización y ansiedad/depresión.

	V22 (Vitalidad emocional)	**V23 (Estrés organizacional)**	**V24 (Estrés emocional)**	**V25 (Estrés físico)**	**V26 (Abandono)**	**V27 (Riesgo Salud)**	**V28 (Calidad personal y organizacional total)**
Mínimo	14	8	15	10	2	2	50
Máximo	98	56	92	60	14	14	290
Media	70.56	28.33	42.25	31.87	5.25	10.29	186.19
Desv. Estándar	12.97	9.42	13.82	9.36	3.20	2.36	24.63

Cuadro 2.16: Análisis numérico de Calidad personal y Organizacional.

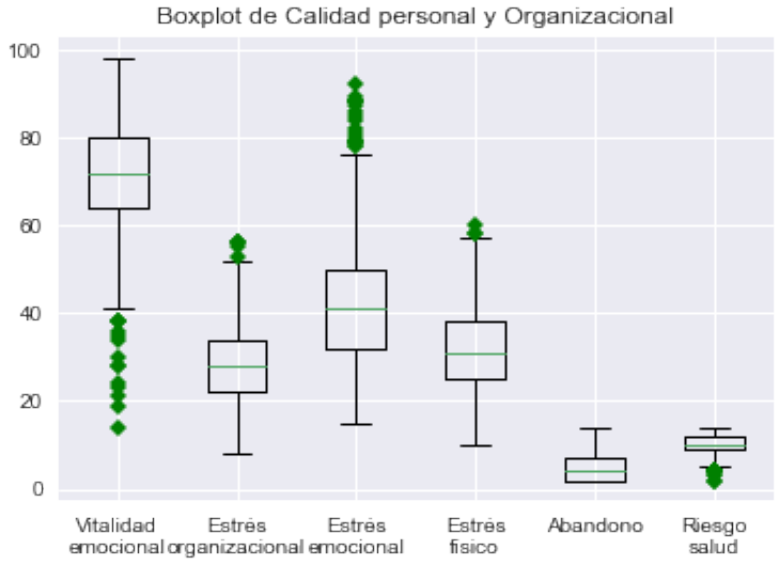

Figura 2.28: Boxplot de Calidad personal y Organizacional.

El rango real para la variable de vitalidad emocional es de 14 a 98, como nos muestran nuestros datos. Para la variable **estrés organizacional** es de 8 a 56. El **estrés emocional** tiene un rango de 15 a 105, por lo que podemos observar que no tenemos en nuestros datos a ningún individuo que haya alcanzado el máximo en cuanto a estrés emocional se refiere. Para el **estrés físico** el rango real va de 10 a 70, mientras que nuestro máximo únicamente alcanza 60. Mientras que para **Abandono y Riesgo de salud**, ambos tienen un rango real de 2 a 14. Como se muestra en la tabla con nuestros datos 2.16 y boxplot 2.28. Lo representado por cada una de estas variables ya ha sido mostrado en la tabla realizada al principio de esta sección, todas estas variables han sido sumadas y restadas (según fueran positivas o negativas) para obtener una variable total que las englobe, la variable **V28 (Calidad Personal y Organizacional Total)**.

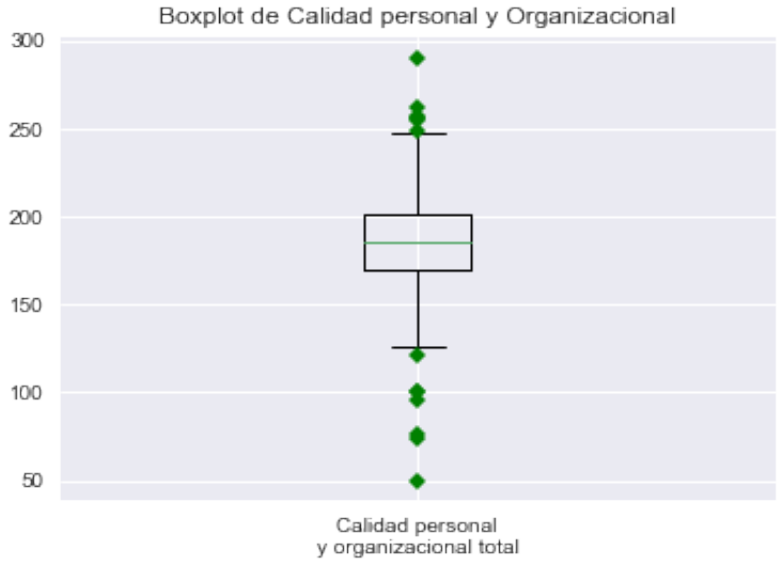

Figura 2.29: Boxplot de Calidad personal y Organizacional total.

Esta variable **V28 (Calidad Personal y Organizacional Total)**, representando el total de las variables de esta sección de Calidad personal y Organizacional, tiene valores en el rango de 50 a 290, como se observa en el boxplot 2.29. La mayoría de los encuestados están concentrados en torno al valor 186 con una oscilación de 24 puntos, la cual no es demasiado grande teniendo en cuenta el gran rango en el que trabaja esta variable. La puntuación está bastante centrada, teniendo esta vez tanto outliers máximos como mínimos. Siendo la calidad personal y organizacional de los encuestados no tan buena como otras variables analizadas en el test.

Capítulo 3

Modelos de Machine Learning usados

En este capítulo hablaremos brevemente sobre todos los modelos y algoritmos usados en cada uno de los apartados de este problema. Realizaremos una pequeña introducción teórica a cada uno de ellos para que el lector tenga un conocimiento mínimo sobre ellos. Cada sección de este capítulo estará dedicada por tanto a cada uno de los modelos y algoritmos mencionados en este documento.

Dichos algoritmos serán usados en la parte de clasificación, usando por tanto dos variantes en cada uno de ellos, para un problema de regresión y para un problema de clasificación. Todos los algoritmos mencionados se han implementado con Python gracias a la librería scikit-learn[24].

3.1. Regresión Lineal

Se ha usado este modelo lineal de regresión[31] como base para los demás modelos de regresión, será el mínimo que han de obtener como resultado. Este modelo se basa en una lógica simple. Lo que buscará será reducir el error residual por el método de los cuadrados, pero únicamente usando para ello una fórmula lineal. Es un método muy simple y que no tendrá parámetros a analizar o mejorar. Se basa en trazar una línea que reduzca lo máximo posible el error entre los puntos de entrenamiento a analizar, dicha línea formada será usada posteriormente para predicción, siendo la siguiente imagen un ejemplo de cómo funcionará este modelo.

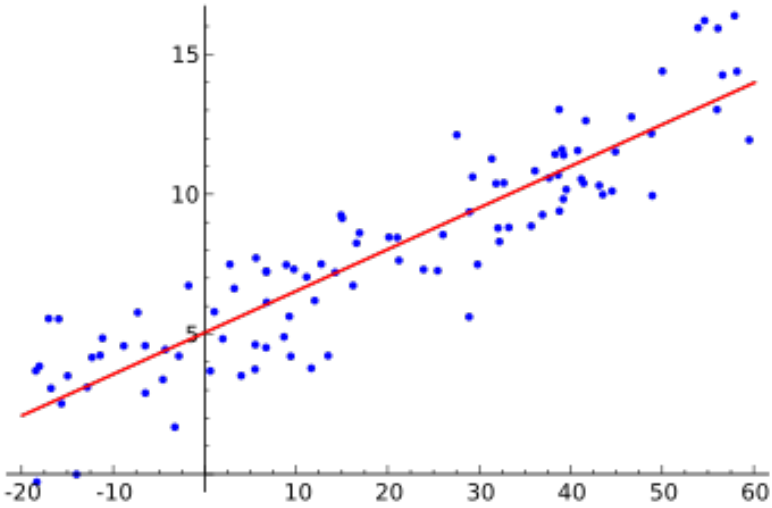

Figura 3.1: Ejemplo de regresión lineal.

3.2. Nearest Neighbour

Implementación del famoso **Nearest Neighbour**[12], algoritmo basado en la distancia existente entre los distintos datos de entrada para predecir la clase. Este algoritmo se fijará en el número de vecinos dado por el usuario para ver los vecinos más cercanos. Observando las instancias más cercanas de la fase de entrenamiento, se quedará con la clase predominante en ellas, la cual asignará posteriormente a esa nueva instancia que queremos clasificar.

3.3. SVM

Implementación del famoso algoritmo **SVM**[33]. Este algoritmo se basa en la separación de las distintas clases de nuestro problema, aunque estas no sean separables, para ello hará uso de hiperplanos. Llevándose las distintas instancias de nuestro problema a otras dimensiones espaciales. En dichas dimensiones se separarán las clases y luego volveremos a nuestro plano. Debido a la gran cantidad de cómputo que necesita este modelo, el tiempo necesario para el entrenamiento del mismo es muy elevado.

3.4. Decision Tree

Implementación de otro clasificador estado del arte, como es el árbol de decisión[26]. Ampliamente usado debido a la gran información que nos proporciona sobre los datos que tenemos y la relación que existe entre las variables de entrada y las clases.

Sobre la altura del árbol y la profundidad del mismo para las hojas, usaremos los parámetros por defecto.

3.5. Random Forest

Agrupación de árboles de decisión[18], es uno de los clasificadores más usados a día de hoy debido a su gran rendimiento tanto en resultados como en coste computacional. Se basa en la agrupación de muchos árboles de decisión que han sido construidos con una pequeña parte de las variables de entrada, posteriormente se realizará algún tipo de operación de agrupamiento para obtener el resultado final para cada instancia del problema a predecir. Normalmente se suele hacer una votación para obtener la clase final.

3.6. Multi-Layer Perceptron

Este modelo será una red neuronal de tipo Feedforward con neuronas Perceptron[9], usaremos un modelo con varias capas y varias neuronas por capa, jugaremos con dichos parámetros para obtener la mejor arquitectura de red para nuestro problema. El entrenamiento de esta red se basará en la propagación hacia atrás, es decir, se creará una red con pesos aleatorios y se irán modificando dichos pesos para corregir el error de salida de nuestro modelo. Por ello, el coste computacional de este modelo es elevado, puesto que son necesarias varias iteraciones para que los pesos de nuestra red puedan converger hasta un punto óptimo para nuestros resultados. Además de tener un gran problema con el sobreentrenamiento, por lo que deberemos de usar un conjunto de validación, para impedir que nuestra red sobreentrene.

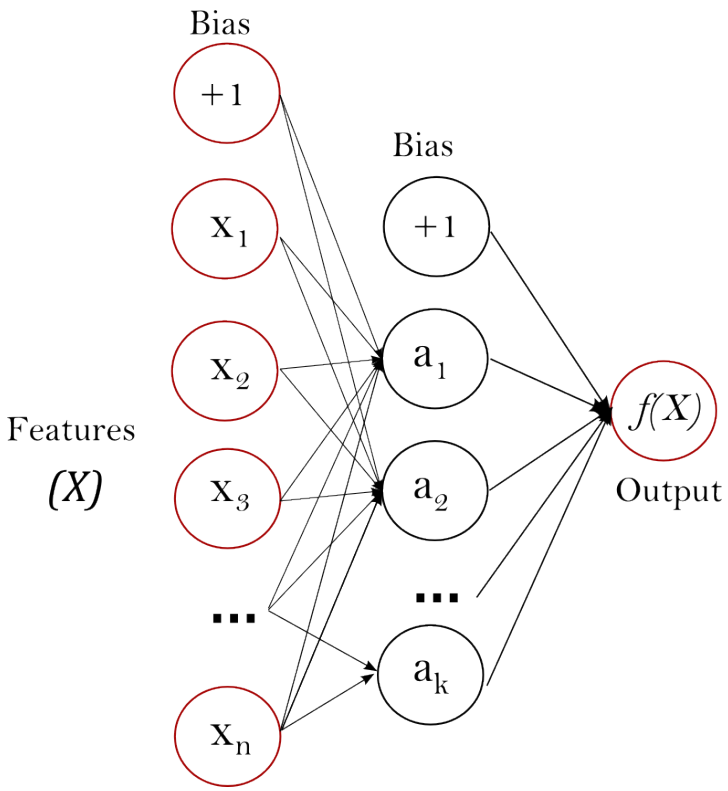

Figura 3.2: Ejemplo de red neuronal.

3.7. Principal Component Analysis (PCA)

Principal Component Analysis (PCA)[14], consistirá en reducir las entradas de nuestro problema a un número representable en un plano o gráfico. Utilizando para ello la combinación de todas nuestras variables, representadas por dos variables principales que nos darán una información resumida sobre todas las variables de nuestro problema. Con este método podremos visualizar claramente cómo están repartidas todas las instancias de nuestro problema y ver la distancia que existen entre ellas.

3.8. K-Means

K-Means o K-medias[11], intentará separar nuestros datos en un número de K clusters especificados por el usuario. Para ello, usará la distancia existente entre las instancias de nuestro problema. Se basará en la inicialización aleatoria de un número K de centroides, dichos centroides representarán el centro de nuestros clusters. En cada iteración de este método, una nueva

instancia del problema se irá añadiendo en cada uno de los clusters por cercanía, siguiendo la siguiente fórmula:

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2) \tag{3.1}$$

A cada iteración del algoritmo también se calculará el centro de cada clúster, siendo este la media de todas las instancias pertenecientes al clúster. De este modo se irán introduciendo todas las instancias en cada uno de los clusters definidos. Minimizando la distancia intra-cluster y maximizando la distancia inter-cluster.

3.9. Hierarchical Clustering

Es una familia de algoritmos de clustering[13] basados en jerarquías, es decir, clusters anidados que se van anidando y separando según el nivel que tomemos en nuestra jerarquía. También está basado en la distancia que existe entre las distintas instancias de nuestro problema. La forma que tiene este método para agrupar o desagrupar los clusters es bastante importante a la hora de la creación de los mismos, existen cuatro formas principales:

- **Ward:** Minimiza la suma de diferencias cuadradas entre todos los clusters.
- **Complete linkage:** Minimiza la distancia máxima entre dos clusters. Escogiendo las dos instancias más alejadas de los dos clusters.
- **Average linkage:** Minimiza la media de las distancias de todas las instancias de dos clusters.
- **Single linkage:** Minimiza la distancia entre las dos instancias más cercanas de dos clusters distintos.

Con estos métodos se irán formando los distintos clusters entre nuestros datos, pudiendo ser nosotros los que cortemos al nivel preferido, para quedarnos con el número de clusters deseado.

3.10. DBSCAN

Los clusters en este método[4] serán creados por áreas de alta densidad separadas unas de otras por áreas de baja densidad de instancias. Los clusters pueden tener cualquier forma, al contrario de lo que pasa con K-means, que los clusters han de tener una forma convexa. El componente central de

este método son los *core samples*, esto serán las instancias que están dentro de las regiones de alta densidad. Para que un cluster sea declarado como tal, debe de existir un mínimo de instancias en dicha área de alta densidad que no superen entre ellas una distancia máxima marcada por el usuario.

Capítulo 4

Relaciones entre variables

4.1. Correlación entre variables y *heatmap*

Para ver la correlación entre variable se usará la siguiente fórmula[3] entre cada una de las variables. Gracias a dicha fórmula podremos saber cómo de relacionadas están dos variables entre sí, pudiendo ser esta correlación tanto positiva como negativa.

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2} \cdot \sqrt{E(Y^2) - E(Y)^2}} \tag{4.1}$$

Para el estudio de las relaciones entre las distintas variables se mostrará a continuación un *heatmap*, mostrando la correlación existente entre unas y otras. Para ello se han quitado las variables que son introducidas por el usuario al no poder modelarse totalmente para nuestro problema, además de no aportar apenas información sobre el mismo. La siguiente figura nos muestra las relaciones existentes entre todas las variables del problema:

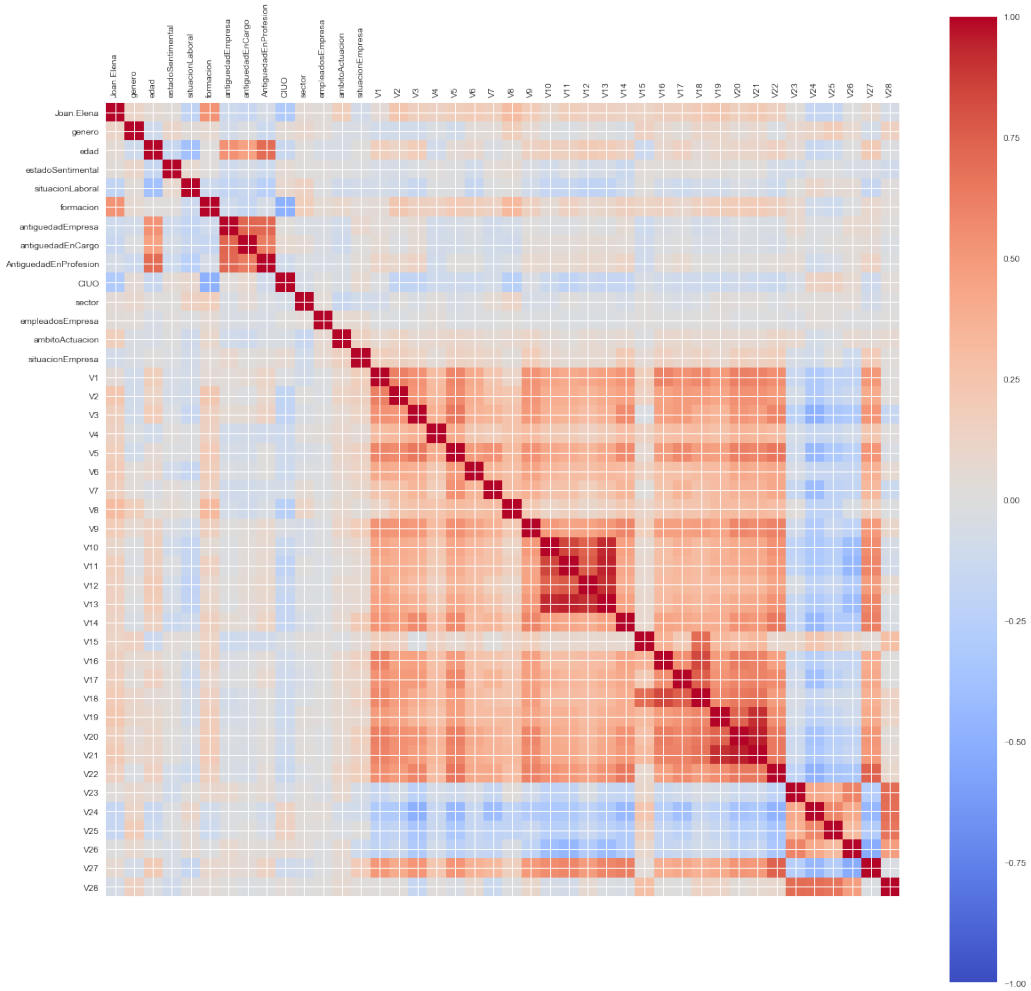

Figura 4.1: Heatmap con todas las variables.

Debido a la gran cantidad de variables a analizar en este gráfico 4.1, este se ve demasiado pequeño y es bastante difícil de analizar por lo que en las siguientes imágenes se ha realizado una división de las variables. Un hecho que puede verse bien a simple vista es que no existe apenas relación entre las variables de tipo personal con las variables psicológicas. Mientras que las variables psicológicas entre ellas si tienen un gran grado de correlación. Debido a que en esta gráfica no se observa demasiado bien las relaciones existentes debido al tamaño de la misma, en los siguiente apartados se mostrarán las variables a predecir de inteligencia energética, realizando para ello dos grupos de variables distintos. Separando para ello las variables personales de las variables psicológicas, excepto las que miden el grado de inteligencia energética.

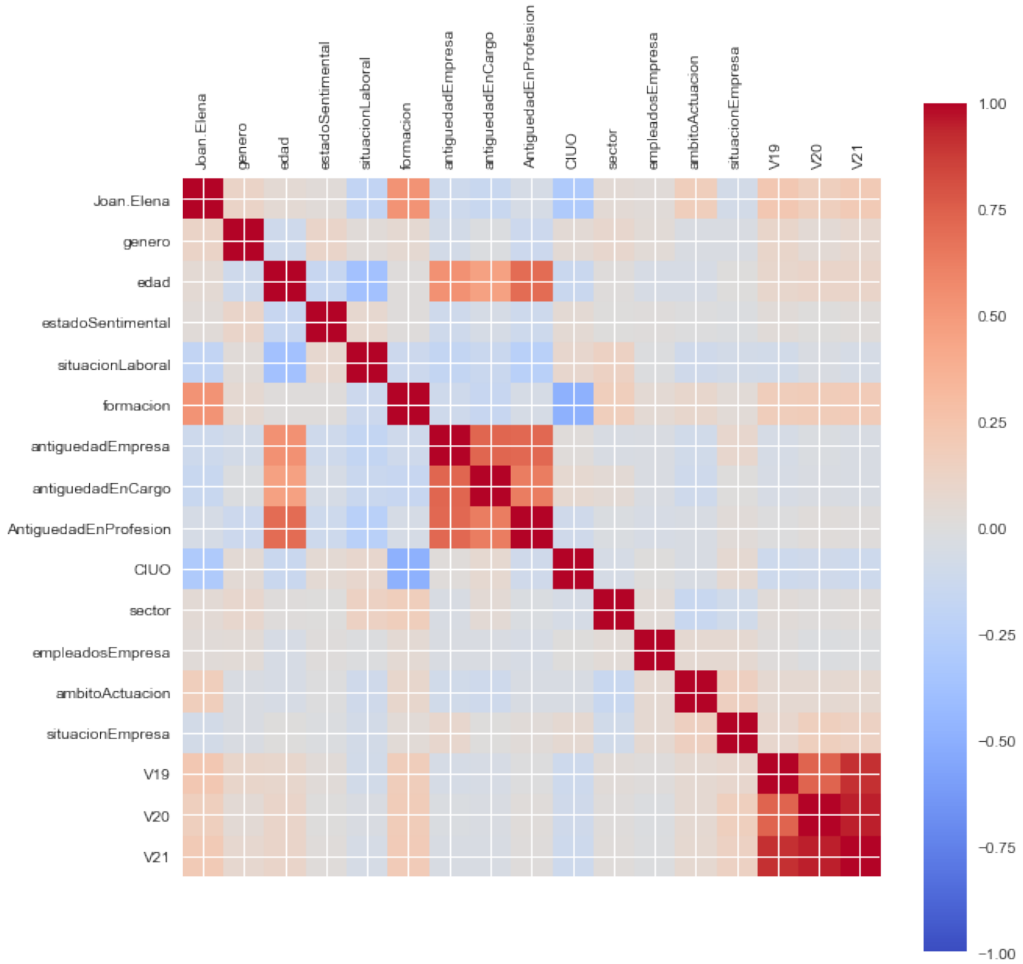

Figura 4.2: Heatmap con las variables sociales.

En esta figura 4.2 puede verse que apenas existe una correlación fuerte entre las variables a predecir, marcadas como **V19** (Self Drive Direction), **V20** (Consciencia energética) y **V21** (Energía total). Con las variables personales y de empresa de nuestros datos. Aunque las que tienen una mayor correlación será la formación de las personas y la situación de la empresa. Siendo esta correlación bastante baja pero más alta que para todas las demás variables. La variable de si la encuesta ha sido realizada por Joan o Elena también tiene correlación con dichas variables de carácter energético. Aunque esto puede ser debido a la alta correlación que tiene con la formación de los sujetos. Una razón para esta correlación puede explicarse debido a la división del trabajo de campo realizado donde cada uno de los entrevistadores ha atacado a sectores distintos de la población para no solapar las respuestas y obtener el mayor grado de información posible. Conforme a

las correlaciones negativas entre nuestras variables, podemos observar una clara correlación entre la formación de nuestros encuestados y el CIUO. Esto puede deberse a que los primeros grupos del CIUO serán los que requieran una mayor formación, siendo esto lo inverso a lo ocurrido con la formación. Ocurre una situación parecida con la edad y la situación laboral de las personas, estando estas correlacionadas negativamente, puesto que a mayor edad la situación laboral deberá de figurar como empleado. A continuación, mostraremos otra figura con todas las variables psicológicas, para ver cuales tienen una mayor relación con nuestra inteligencia energética.

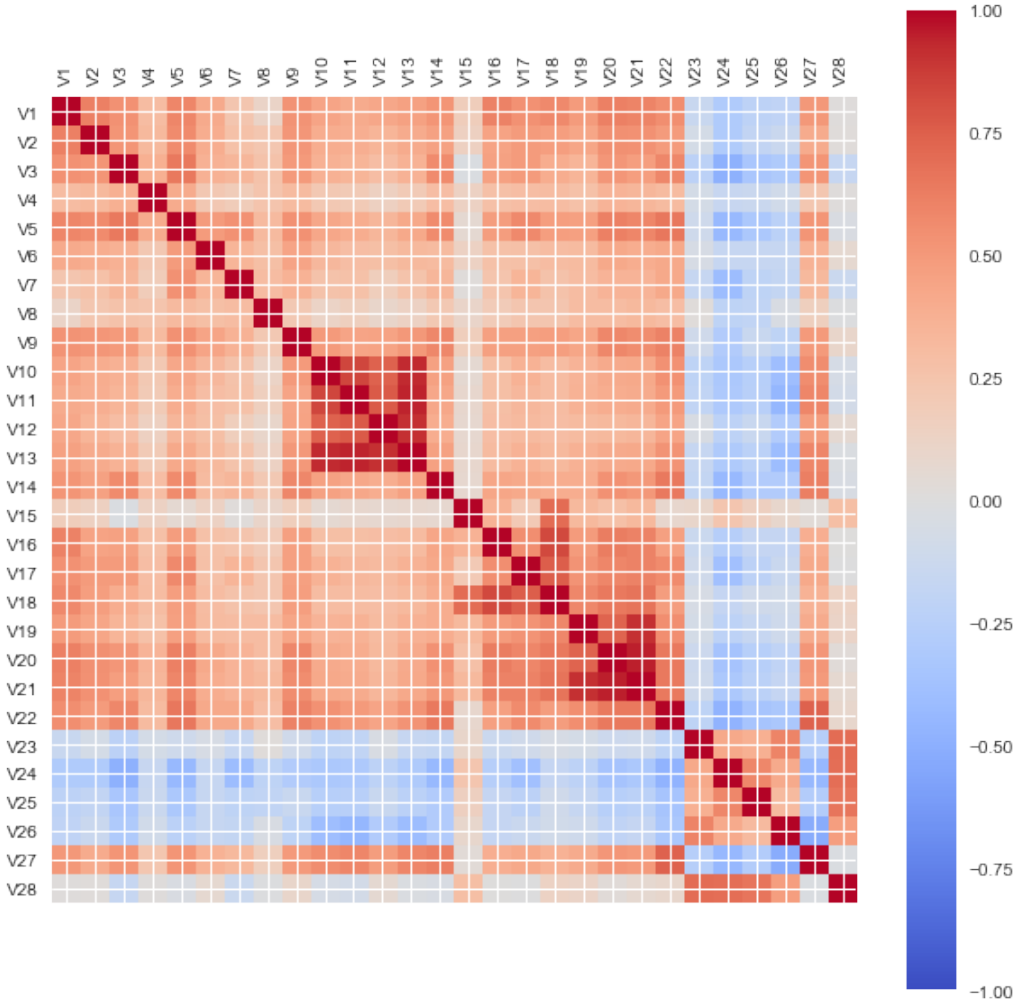

Figura 4.3: Heatmap con las variables psicológicas.

En esta figura 4.3 si se puede apreciar mejor como entre las variables de carácter psicológico existe una mayor correlación, siendo la mayoría de

estas correlaciones positivas. Existiendo un grupo que tiene correlaciones negativas con todas las demás variables, siendo estas variables **V23** (Estrés organizacional), **V24** (Estrés emocional), **V25** (Estrés físico) y **V26** (Abandono), correspondiendo todas estas variables a estados negativos de la persona analizada, siendo estas las únicas variables consideradas negativas de nuestro estudio, por ello es lógico suponer (y se puede observar) que serán relacionadas inversamente con las demás. Siguiendo con nuestro pequeño estudio de estas correlaciones, puede observarse la existencia de pequeños clusters en este gráfico. Por ejemplo, el existente entre las variables de **V10 a V13** (Vigor, Dedicación, Absorción y Engagement total), variables que se usan para calcular el **engagement** de una persona. Otro grupo claro es el formado por las variables de **V15 a V18** (Percepción, Comprensión, Regulación e Inteligencia emocional total), las cuales son variables para calcular la **Inteligencia Emocional** de una persona. Y por último el grupo formado por las variables interesantes de nuestro problema desde **V19 a V21** (Self Drive Direction, Consciencia energética y Energía total). Todos estos grupos comentados son grupos ya predefinidos en nuestro problema, debido a que miden partes distintas de la psicología asociada a los individuos a analizar y todas las variables de un mismo grupo representarán una parte importante de cada uno de los sujetos. Para las variables a predecir que nos interesan, podemos ver que desde **V16 a V18** (Comprensión, Regulación e Inteligencia emocional total) existe bastante correlación con las variables de inteligencia energética. Esto puede deberse a que la inteligencia emocional puede estar bastante ligada a la inteligencia energética, puesto que con un alto rango en ambas somos capaces de sentir mejor nuestro propio cuerpo y nuestras emociones, así como las energías que rodean al mismo, es por tanto considerable que dichas variables estén correladas. Otras variables que también tienen una gran correlación con nuestras variables a predecir serán **V1** (Fluir), **V2** (Autoeficacia), **V3** (Autoestima), **V5** (Estabilidad emocional) y **V9** (Florecimiento). Sin embargo, tanto V1 (Fluir) como V9 (Florecimiento) serán parte de las variables a estudiar en este informe, siendo parte de las variables energéticas. También tendrá una alta correlación negativa la variable **V24** (Estrés emocional).

4.2. Correlación usando Random Forest

En este apartado usaremos el clasificador *Random Forest* para obtener una correlación entre las variables. Esto es gracias a la naturaleza de este clasificador. Como se sabe el clasificador Random Forest consiste en un conjunto de árboles de decisión. Dichos árboles de decisión serán creados con un número reducido de las variables del problema. Lo que haremos para obtener el grado de correlación entre las variables a predecir y las variables

de entrada será ver el número de veces que estas variables de entrada han sido seleccionadas como nodo raíz. Siendo este nodo el más importante para nuestro clasificador. Puesto que dicho nodo será el que tiene una mayor capacidad de división para nuestras instancias en las distintas clases del problema.

Para este apartado, hemos dividido el problema en varios, tratando de predecir las siguientes variables V1 (Flow-Fluir), V9 (Flourishing-Florecimiento), V19 (Self Drive Direction), V20 (Consciencia energética), V21 (Energía total), usando para ello como variables de entrada todas las demás variables existentes en nuestro problema. Todos los problemas se han realizado usando la implementación de Random Forest de scikit-learn, usando para ello 1000 árboles en cada clasificador, para cerciorarnos de que los datos obtenidos son reales.

Los resultados obtenidos serán mostrados y comentados a continuación:

4.2.1. V1 (Flow-Fluir)

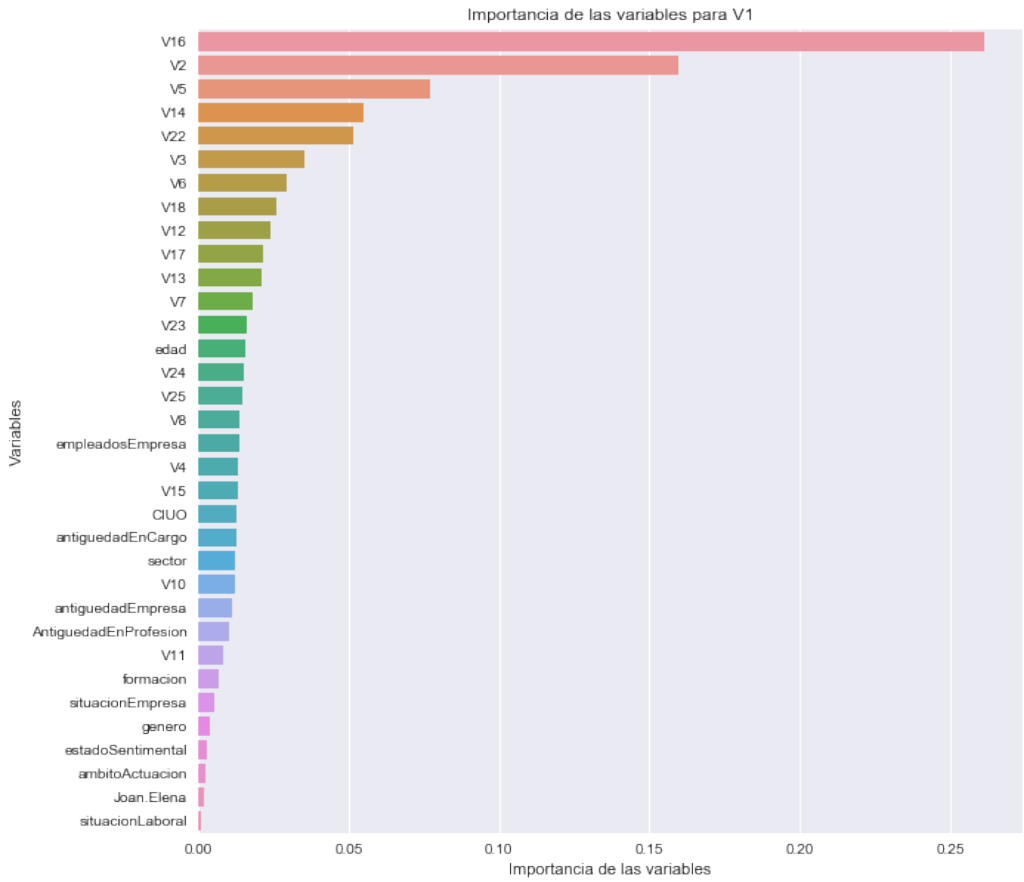

Figura 4.4: Importancia de las variables para V1 según Random Forest.

Como podemos ver en la gráfica 4.4 creada para la variable **V1** (Flow-Fluir), las tres variables más relevantes para este problema de predicción han sido **V16** (Comprensión), **V2** (Autoeficacia) y **V5** (Estabilidad emocional). Podemos ver que estas variables tienen relación con el estado de fluir, puesto que necesitas tener una estabilidad emocional y ser bastante eficiente para entrar en dicho estado. Lo curioso de nuestro problema es que las variables de carácter personal y de empresa apenas tienen repercusión en el mismo, como podremos ver en las siguientes situaciones. La que más relevancia tendrá es el tamaño de la empresa, que puede influir en otras variables como la estabilidad de la misma. Puede verse también que la variable con un menor impacto sobre nuestra variable a predecir será la situación laboral, siendo esto extraño puesto que la situación laboral de una persona puede afectar en la carga psicológica de la misma.

4.2.2. V9 (Flourishing-Florecimiento)

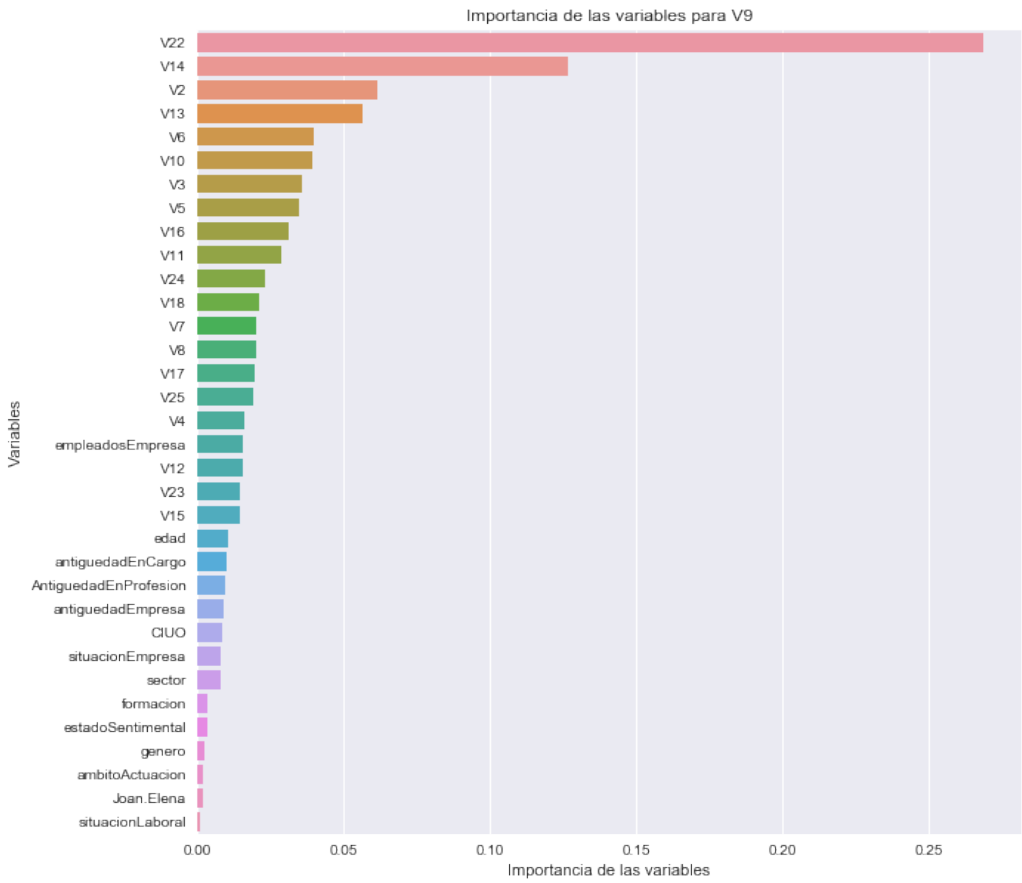

Figura 4.5: Importancia de las variables para V9 según Random Forest.

Para esta variable **V9** (Florecimiento) podemos observar la gráfica 4.5, las tres variables que más han influido en la predicción de la misma son: **V22** (Vitalidad emocional), **V14** (Satisfacción con la vida) y **V2** (Autoeficacia). Volvemos a ver la autoeficacia influyendo en la predicción de una nueva variable. Puede verse otra vez que las variables personales apenas tienen importancia para este problema. Siendo otra vez la más influyente el número de personas en la empresa. Vemos lo ya repetido para la gráfica anterior, una variable muy influyente con casi un 25 % de importancia, por encima de variables que apenas tienen importancia como la situación laboral.

4.2.3. V19 (Self Drive Direction)

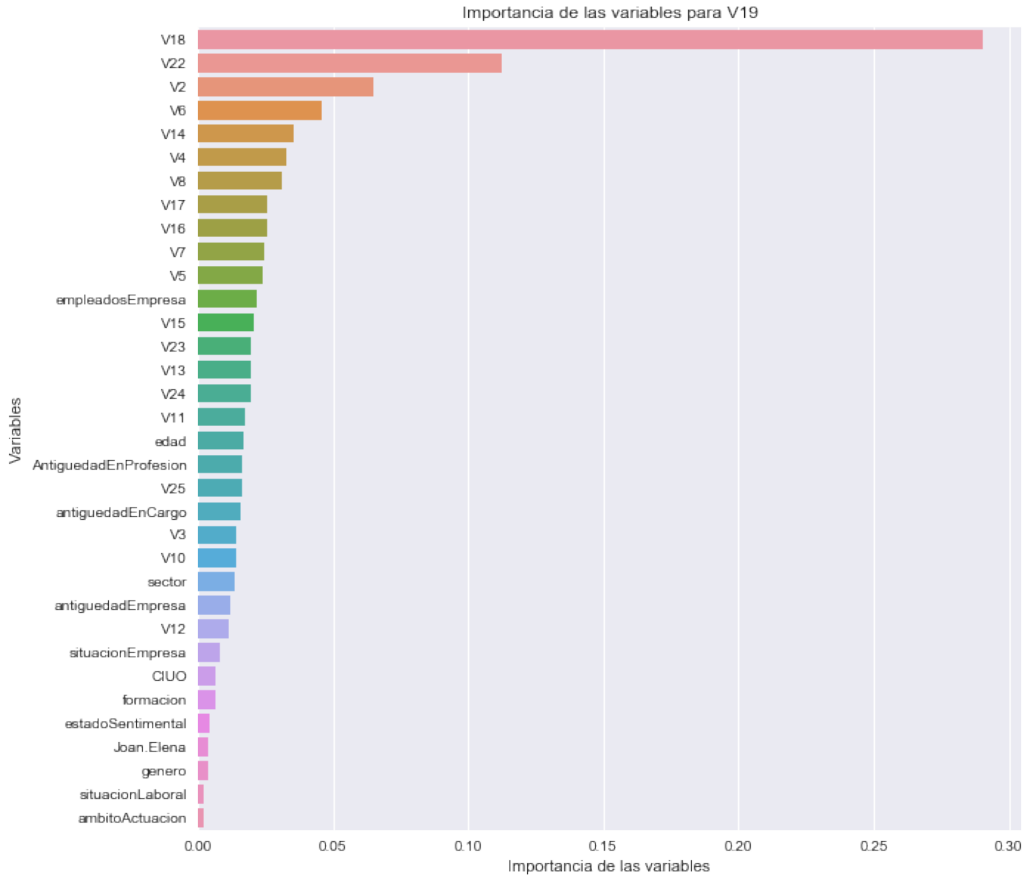

Figura 4.6: Importancia de las variables para V19 según Random Forest.

Empezando con las variables de Inteligencia energética, tenemos a **V19** (Self Drive Direction), la gráfica 4.6 nos muestra sus relaciones con otras variables, la cual hace referencia a una parte de nuestra inteligencia energética. Según podemos ver en esta gráfica de barras, las variables influyentes para esta variable energética a predecir han cambiado con respecto a las anteriores. Siendo esta vez, las tres más influyentes las siguientes: **V18** (Inteligencia Emocional total), **V22** (Vitalidad Emocional) y **V2** (Autoeficacia). Aunque V22 (Vitalidad emocional) y V2 (Autoeficacia) ya se habían visto antes para las variables a predecir anteriores, podemos concluir por tanto que serán variables relevantes para nuestro problema en general y estarán relacionados con todos los aspectos psicológicos estudiados. Viendo otra vez como la variable más influyente para los aspectos personales estudiados será el tamaño de la empresa, como hemos visto para todas las variables a predecir ante-

riormente estudiadas. Observando esta vez que la variable menos influyente en nuestro problema será el ámbito de actuación de la empresa. Siendo esta una vez más, una de las variables personales recogidas.

4.2.4. V20 (Consciencia energética)

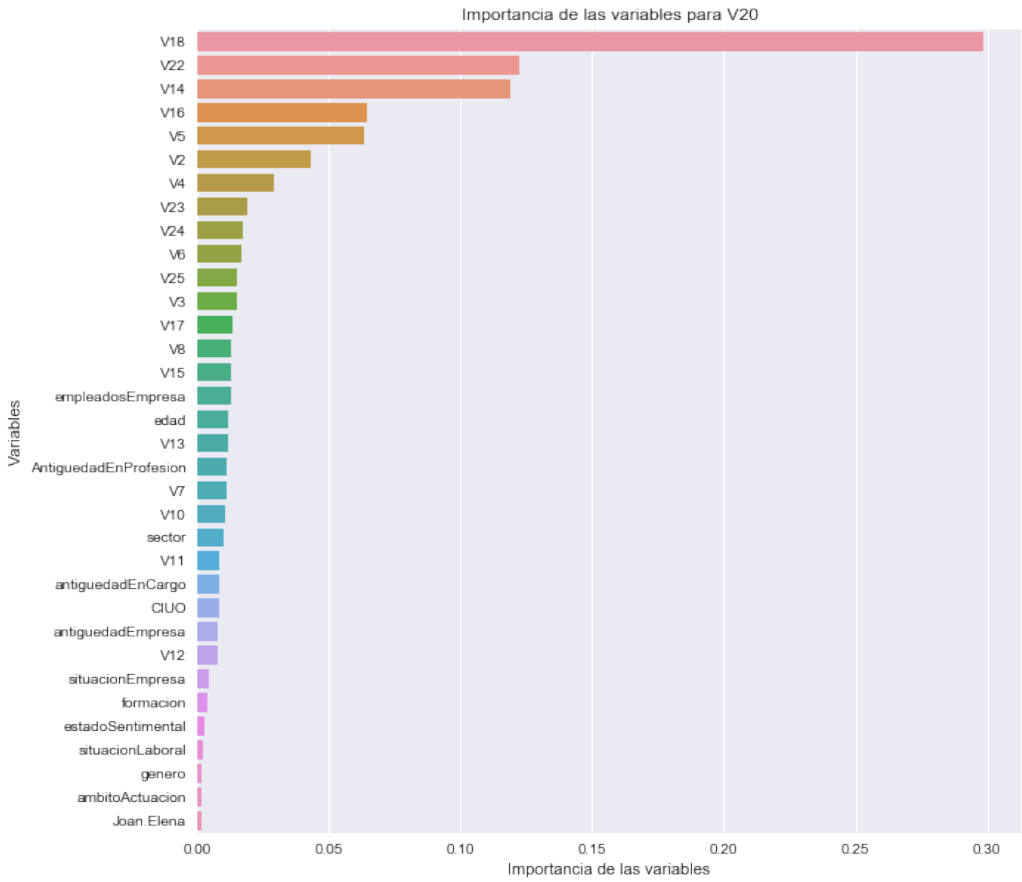

Figura 4.7: Importancia de las variables para V20 según Random Forest.

En esta variable **V20** (Consciencia energética), también del apartado de inteligencia energética y cuya gráfica 4.7 se puede observar arriba, vemos relaciones con otras variables que hemos visto en el apartado anterior. Concretamente con **V22** (Vitalidad emocional) y **V18** (Inteligencia total emocional). Alcanzando el top 3 en relación con nuestra variable, la variable **V14** (Satisfacción con la vida). Puede verse que las variables energéticas están bastante relacionadas con las puntuaciones en el apartado de inteligencia emocional, autoeficacia y satisfacción con la vida. Mientras que las variables personales apenas tienen influencia, da igual el ámbito del que

venga la persona, sabiendo sus puntuaciones en las variables de inteligencia emocional puede predecirse cómo será su inteligencia energética.

4.2.5. V21 (Energía total)

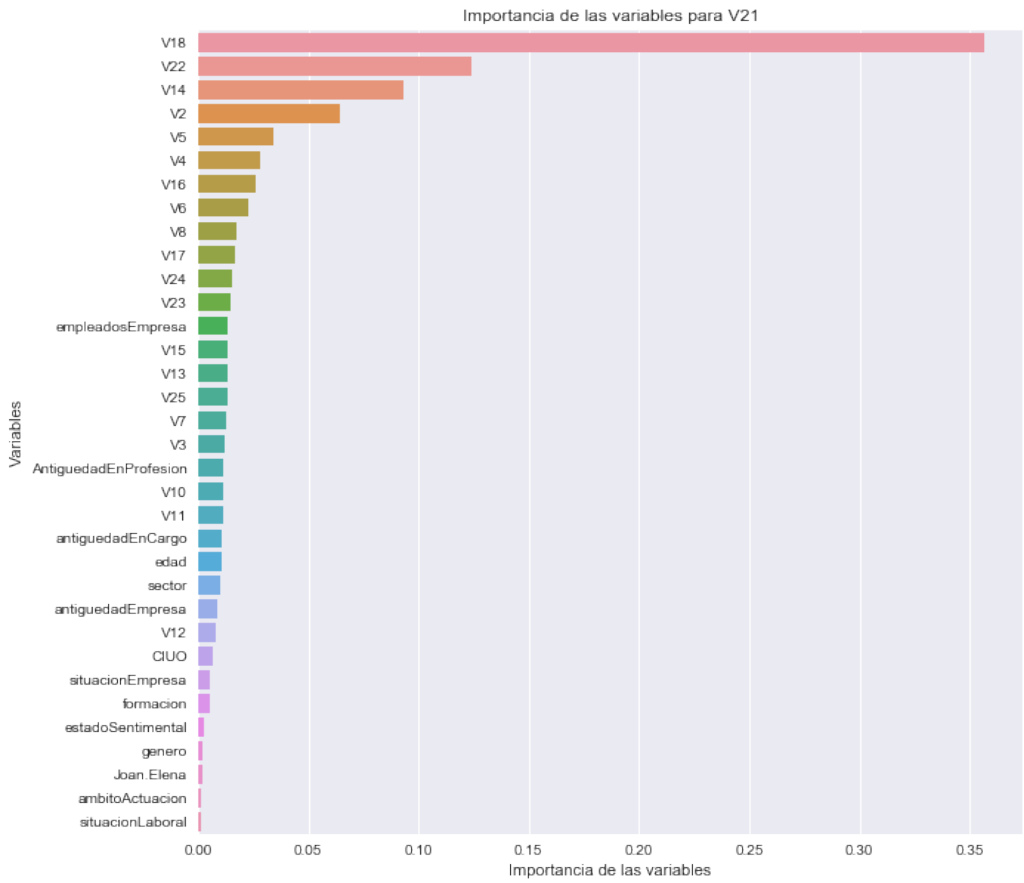

Figura 4.8: Importancia de las variables para V21 según Random Forest.

V21 representa la variable total de la inteligencia energética, esta será la suma de las dos anteriores, por lo que es de suponer que las variables más relacionadas con esta sean bastante parecidas a las que hemos visto en los dos apartados anteriores. Esta gráfica 4.8 es prácticamente idéntica a la obtenida para la variable **V20**. Con esta gráfica también se da el top 3 de relaciones mencionado anteriormente, siendo estas **V18** (Inteligencia emocional total), **V22** (Vitalidad emocional) y **V14** (Satisfacción con la vida).

4.2.6. Conclusiones

Como gran conclusión de estos estudios con el clasificador Random Forest, se ha obtenido la alta relación que existe entre las variables psicológicas que han medido la inteligencia emocional con nuestras variables a analizar para este problema, las variables de inteligencia energética. Podemos presumir que dichas inteligencias están bastante relacionadas y son bastante parecidas, puesto que ambas se basan en la propia consciencia del cuerpo y de lo que le rodea. Por lo tanto, puede implicar que tener una alta capacidad de inteligencia emocional llegue a dar lugar a una gran capacidad en el aspecto de inteligencia energética. Otros aspectos importantes que son bastante influyentes para estas variables a analizar, como hemos podido ver en los apartados anteriores, tienen que ver con la capacidad de autoeficacia y autorrealización de uno mismo, las cuales están directamente relacionadas con la satisfacción con la vida. Puesto que si somos autoeficaces y vemos que tenemos el control de nuestra vida, esto concluye en una mayor satisfacción con la misma. Dicho grado de control de nuestra vida también estará implícitamente relacionado con el control que tenemos de nuestro cuerpo, así como de las energías y sentimientos que contenemos y nos rodean.

Capítulo 5

Procedimiento de los experimentos a realizar

Con el fin de obtener unos resultados lo más replicables y fiables posibles, para todos los experimentos realizados se ha seguido la configuración que describiremos a continuación. Como puede verse en la siguiente figura, esta ha sido la configuración establecida para los experimentos:

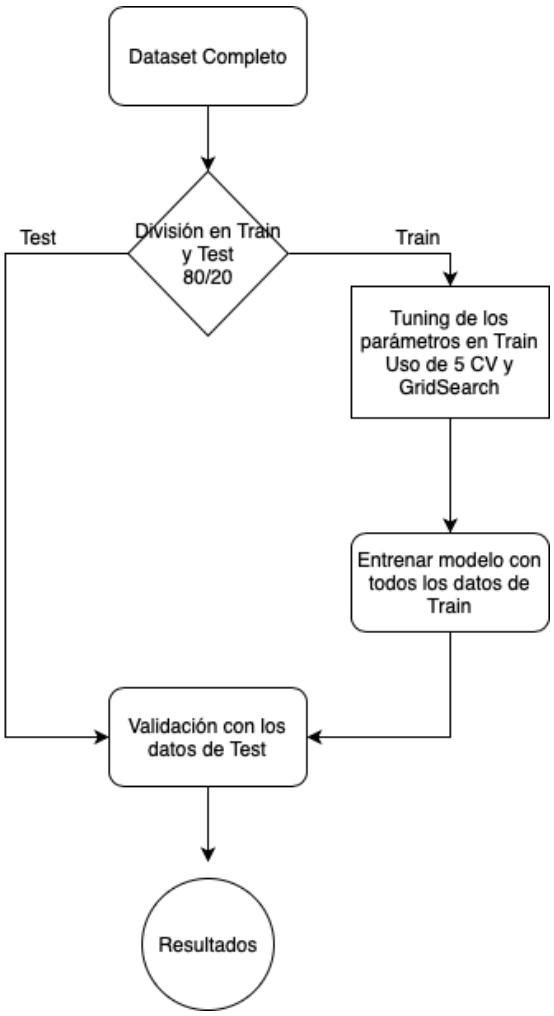

Figura 5.1: Diagrama del proceso del experimento.

Siguiendo este diagrama5.1, partiremos del conjunto de datos completo con todas las variables necesarias para cada uno de los problemas a resolver (se crearán experimentos diferentes para cada una de las variables a predecir del problema). Una vez cargado dicho dataset completo, se dividirá usando un 80 % de datos para entrenamiento y un 20 % de datos para Test[35]. Dichos datos para Test no se volverán a ver hasta el final del experimento para obtener los resultados finales de validación. Con el fin de poder comparar todos los modelos y experimentos, esta división se realizará antes de la ejecución de los experimentos y se guardarán los datos en disco, haciendo que todos los modelos obtengan el mismo conjunto de entrenamiento y de test. Sobre este conjunto de entrenamiento se realizará el ajuste de los parámetros del modelo con una búsqueda en rejilla. Dicha búsqueda de parámetros tendrá en cuenta los resultados obtenidos en una validación cruzada de 5

particiones[16]. Al final de cada ronda de la validación cruzada, obtendremos los resultados para los parámetros probados. Al final de la comprobación de todos los parámetros, nos quedaremos con los que mejores resultados nos hayan dado de media en la validación cruzada. Una vez obtenidos los mejores parámetros para los datos de entrenamiento. Se entranará el modelo final, usando para ello los mejores parámetros obtenidos hasta el momento y todo el conjunto de entrenamiento (el 80 % de los datos totales). Cuando este modelo haya completado su entrenamiento lo haremos predecir el 20 % de los datos de Test restantes, obteniendo así los resultados esperados para nuestro experimento. Se realizará este experimento para cada uno de los modelos propuestos y cada uno de los problemas a predecir con nuestros datos.

Se realizará la partición para cada uno de los problemas a predecir, siendo estos los siguientes: (Hay que hacer notar que estos problemas de predicción fueron diseñados por los expertos psicólogos).

Atributos del problema	Variables a predecir
Dp 1.- Género Dp 2.- Edad Dp 3.- Estado civil Dp 4.- Situación Dp 5.- Nivel de formación académica Dp 7.- Antigüedad en la empresa-institución Dp 9.- Número total de profesionales/empleados a su cargo Dp 10.- Antigüedad en el cargo Dp 11.- Antigüedad en esta profesión V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción. V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V1.- Flow-Fluir**

Dp 1.- Género Dp 2.- Edad Dp 3.- Estado civil Dp 4.- Situación Dp 5.- Nivel de formación académica Dp 7.- Antigüedad en la empresa-institución Dp 9.- Número total de profesionales/empleados a su cargo Dp 10.- Antigüedad en el cargo Dp 11.- Antigüedad en esta profesión V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V9.- Flourishing-Florecimiento**
Dp 1.- Género Dp 2.- Edad Dp 3.- Estado civil Dp 4.- Situación Dp 5.- Nivel de formación académica Dp 7.- Antigüedad en la empresa-institución Dp 9.- Número total de profesionales/empleados a su cargo Dp 10.- Antigüedad en el cargo Dp 11.- Antigüedad en esta profesión V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V19.- Energía.1. Self Drive Direction.**

Dp 1.- Género Dp 2.- Edad Dp 3.- Estado civil Dp 4.- Situación Dp 5.- Nivel de formación académica Dp 7.- Antigüedad en la empresa-institución Dp 9.- Número total de profesionales/empleados a su cargo Dp 10.- Antigüedad en el cargo Dp 11.- Antigüedad en esta profesión V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V20. Energía. Consciencia energética.**
Dp 1.- Género Dp 2.- Edad Dp 3.- Estado civil Dp 4.- Situación Dp 5.- Nivel de formación académica Dp 7.- Antigüedad en la empresa-institución Dp 9.- Número total de profesionales/empleados a su cargo Dp 10.- Antigüedad en el cargo Dp 11.- Antigüedad en esta profesión V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V21. Energía Total.**

Cuadro 5.1: Variables de los experimentos.

Se han normalizado todos los valores de nuestros datos para que al aplicar modelos basados en distancias, estos no se vean afectados por el rango dispar de las variables. Además, al tener todas las variables en el mismo rango, será más sencilla la clasificación y predicción para nuestros modelos, así como la interpretación de los resultados obtenidos.

Capítulo 6

Uso de modelos de regresión para los distintos problemas

Como hemos podido observar en lo que llevamos de documento y de análisis del problema. A la hora de predecir las variables a estudiar, la mejor forma de hacerlo será con modelos de regresión. Puesto que dichas variables a predecir constan de un rango amplio de valores, es decir, nuestro problema debe de ser categorizado como un problema de regresión y no como un problema de clasificación multiclase. Aunque posteriormente trataremos de etiquetar a cada uno de nuestros sujetos, para convertir dichas variables de regresión en variables multiclase.

En este apartado veremos los distintos pasos a seguir para cada uno de los modelos escogidos en el problema de regresión, así como de las métricas usadas para comparar dichos modelos. Todos los modelos serán entrenados y construidos siguiendo los pasos propuestos en la anterior sección.

Como métrica principal para obtener los mejores parámetros a usar en nuestro modelo, usaremos la métrica: *MSE*.

6.1. Métricas usadas

A continuación, hablaremos de las métricas usadas para este apartado, para la comparación de los distintos modelos.

6.1.1. Mean Squared Error (MSE)

El error cuadrático medio[37] es de las métricas más utilizadas para evaluar los modelos de regresión, tanto esta métrica como su variante, la raíz del error cuadrático medio. Son las dos métricas más usadas en todos

los problemas de regresión. Se basa en ver el error cometido por cada una de las instancias predichas conforme al valor real de las mismas, se realizará la media de dicho error y este será el error cometido por nuestro modelo. La fórmula es la siguiente:

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2. \tag{6.1}$$

6.1.2. Explained Variance Score

Esta métrica[22] nos dirá la proporción en la que el modelo tiene en cuenta la dispersión de nuestros datos. Esta variación será cuantificada por la varianza. El mejor resultado posible para esta métrica será 1. La fórmula para computarla puede verse como sigue:

$$explained_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}} \tag{6.2}$$

6.1.3. R2 Score

Esta métrica es el coeficiente de determinación[20]. El cual representa la proporción de varianza explicada por las variables independientes del modelo. Nos proporciona una indicación sobre cómo de bueno será nuestro modelo para predecir instancias no vistas por el mismo. Es una medida que depende del conjunto de datos sobre el que se aplica, puesto que mira la varianza de nuestros datos. Sin embargo, al tener todos los problemas en el mismo conjunto de datos y con las mismas instancias, excepto los valores de la variable a predecir, en nuestro caso sí será comparable. Será una métrica a mejorar, siendo la máxima puntuación posible 1.0. Pudiendo ser esta métrica negativa, debido a que el modelo puede ser peor que realizar las predicciones con la varianza.

La fórmula para obtener esta métrica es la siguiente:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.3}$$

Donde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$

6.2. Modelos y parámetros escogidos

En esta sección mencionaremos los modelos que han sido usados para este análisis del problema en regresión, además hablaremos de los distintos parámetros a mejorar para cada uno de estos modelos.

6.2.1. Linear Regression

No es necesaria la mejora de parámetros para este modelo.

6.2.2. Nearest Neighbor Regressor

Los parámetros a mejorar serán los siguientes:

- **n_neighbors:** Número de vecinos
- **algorithm:** Algoritmo usado para computar los vecinos más cercanos.

Y los parámetros que probaremos en la fase de entrenamiento de nuestro modelo están listados en la siguiente tabla:

n_neighbors	{1...20}
algorithm	auto

Cuadro 6.1: Parámetros a mejorar para kNN Regressor.

6.2.3. SVM Regressor

Los parámetros a mejorar serán los siguientes:

- **kernel:** Especifica el tipo de kernel que usaremos para el clasificador.
- **gamma:** El coeficiente del kernel.
- **C:** Parámetro para penalizar los errores.

kernel	{rbf, poly, linear, sigmoid}
gamma	auto
C	{0.001, 0.01, 0.1, 0.5, 1, 10}

Cuadro 6.2: Parámetros a mejorar para SVM Regressor.

6.2.4. Decision Tree Regressor

Los parámetros a mejorar serán los siguientes:

- **criterion:** Medida a usar por el árbol para realizar los distintos cortes y separaciones en el árbol.

criterion	{mse, friedman_mse, mae}

Cuadro 6.3: Parámetros a mejorar para Decision Tree Regressor.

6.2.5. Random Forest Regressor

Los parámetros a mejorar serán los siguientes:

- **n_estimators:** El número de árboles que queremos en nuestro modelo.
- **criterion:** Como hemos visto en el apartado del árbol de decisión, será la medida a seguir para realizar los cortes y separaciones en nuestros árboles.

n_estimators	{10, 50, 100, 500, 1000}
criterion	{mse, friedman_mse, mae}

Cuadro 6.4: Parámetros a mejorar para Random Forest.

6.2.6. Multi-Layer Perceptron Regressor

Los parámetros a mejorar en este modelo son los siguientes:

- **hidden_layer_sizes:** Nos dará información sobre el número de capas, así como el número de neuronas en cada una de las capas.
- **activation:** Tipo de activación que tendrán las neuronas de nuestra red neuronal.
- **solver:** Optimizador interno para nuestro modelo de red neuronal.

Para este modelo si tendremos bastantes configuraciones debido a la gran cantidad de arquitecturas de red que nos permite el escoger las distintas neuronas por capa de nuestro modelo, así como el número de capas. Probaremos tanto con redes neuronales de una sola capa y muchas neuronas, como con redes con varias capas y pocas neuronas o muchas en cada una de ellas. Para la obtención de los mejores parámetros y hacer que nuestra red no sobreentrene sobre el proceso de entrenamiento se usará un 20 % de los datos de entrenamiento como validación para la red neuronal, sabiendo esta el mejor momento para parar de entrenar (early-stopping).

hidden_layer_sizes	1x(2), 1x(5), 1x(10), 1x(20), 1x(50), 1x(100) 2x(2), 2x(5), 2x(10), 2x(20), 2x(50), 2x(100) 3x(2), 3x(5), 3x(10), 3x(20), 3x(50), 3x(100) 5x(2), 5x(5), 5x(10), 5x(20), 5x(50), 5x(100)
activation	{identity, logistic, tank, relee}
solver	{lifts, sgd, adam}
early_stopping	True
validation_fraction	0.2

Cuadro 6.5: Parámetros a mejorar para la Red Neuronal.

*Para entender esta nomenclatura, tomaremos el primer número como el número de capas ocultas en nuestra red, mientras que el número entre paréntesis será el número de neuronas en cada una de las capas ocultas. Por ejemplo, 3x(5) representa la existencia de 3 capas ocultas en nuestro modelo en la que cada una de las capas ocultas tendrá 5 neuronas. Siendo un esquema gráfico el siguiente:

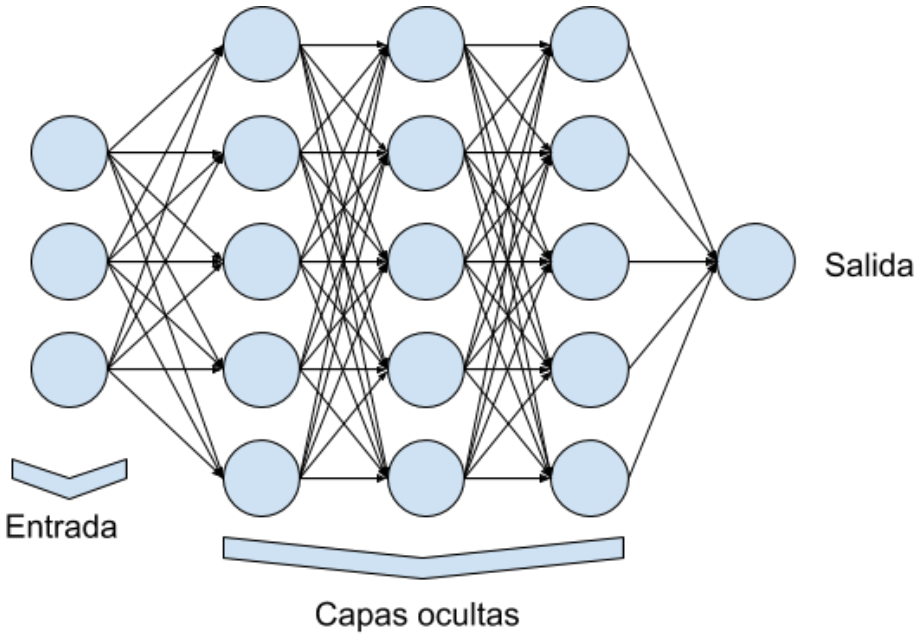

Figura 6.1: Ejemplo de red neuronal con nomenclatura.

6.3. Resultados obtenidos

En esta sección mostraremos los resultados obtenidos para cada uno de los modelos en cada uno de los problemas analizados. Se mostrará la media obtenida para el 5-fold, esta será la mejor media obtenida para los datos de Train, así como las medidas obtenidas para los datos de Test, los cuales no han sido vistos por el modelo hasta el momento de la predicción. Se mostrarán también la mejor configuración de parámetros obtenida para cada uno de los problemas y modelos.

En este apartado se dividirán los resultados obtenidos según los distintos problemas a analizar, se expondrán todos en una tabla, marcando en negrita el modelo que haya obtenido el mejor resultado para cada una de las medidas.

Los resultados obtenidos por los modelos en los que se han usado todas las variables del problema como variables de entrada podrán consultarse en el apéndice, sin embargo, hé aquí las conclusiones finales de los mismos.

6.3.1. Conclusiones

Aunque se han ido comentando los resultados en cada uno de los problemas analizados, aquí se mostrarán algunas conclusiones generales obtenidas de la observación de todos los resultados. Como primera observación puede verse claramente que los dos mejores modelos para estos problemas han sido **Random Forest** y **MLP Regressor**, siendo este una red neuronal. Aunque los mejores resultados en la fase de Test, que es la más interesante, casi siempre han sido obtenidos por Random Forest. Cabe destacar también que el entrenamiento de una red neuronal es bastante complejo, puesto que las mismas pueden sobreentrenarse con facilidad. Esta es la razón que puede estar asociada a que haya obtenido los mejores resultados para la fase de entrenamiento pero no para la de test. Sin embargo, no se ha quedado muy atrás de nuestro conjunto de árboles. Aunque el error cometido tenga un rango bastante bajo en la mayoría de los modelos, puede observarse en casi todas las variables que los modelos han obtenido resultados un poco mejores que el modelo de regresión lineal, sin embargo, esta pequeña mejora no es suficiente para nuestro problema, puesto que este es el modelo más simple de los implementados y debería de darnos un corte para nuestros resultados en las distintas variables. No podemos decir lo mismo de los otros modelos: **Nearest Neighbor Regressor**, **SVM Regressor** y **Decision Tree Regressor**. Los cuales a pesar de tener un coste computacional mayor y ser más complejos que la regresión lineal, han obtenido peores resultados, lo cual es un poco inesperado para estos modelos.

Como conclusión final, para estos problemas escogería el modelo de **Random Forest**, debido a que es un modelo con un bajo coste computacional,

puesto que su clasificador base es un árbol de decisión. Pese a su poco coste computacional ha sido el que mejores resultados ha obtenido para todos estos problemas. Aún estando casi a la par la red neuronal, pero esta es bastante más compleja tanto computacionalmente como en entendimiento. Además, en el siguiente apartado, utilizaremos únicamente las variables de ámbito psicológico para la construcción de nuestros modelos. Pues como vimos en el apartado de la correlación entre variables, las variables de carácter personal apenas aportan información a nuestro problema. Con este nuevo enfoque intentaremos mejorar los resultados obtenidos hasta el momento.

6.4. Repetición de los experimentos sin las variables personales

Debido a los bajos resultados obtenidos en nuestros experimentos y como comentamos en los apartados anteriores sobre la correlación de las variables, volveremos a repetir estos experimentos pero ignorando las variables de carácter personal de nuestros datos para comprobar si estas han influido en la mala predicción de nuestros modelos. A continuación mostraremos los resultados obtenidos para esta parte de los experimentos, esperando que mejoren los resultados obtenidos en la sección anterior.

Atributos del problema	Variables a predecir
V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción. V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V1.- Flow-Fluir**
V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V9.- Flourishing-Florecimiento**

V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V19.- Energía.1. Self Drive Direction.**
V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V20. Energía. Consciencia energética.**
V2.- Autoeficacia V3.- Autoestima V4.- Extraversión V5.- Estabilidad emocional V6.- Responsabilidad V7.- Amabilidad V8.- Apertura a la experiencia V10.- Engagement 1. Vigor V11.- Engagement 2. Dedicación V12.- Engagement 3. Absorción V13.- Engagement total V14.- Satisfacción con la vida V15.- Inteligencia emocional 1. Percepción V16.- Inteligencia emocional 2. Comprensión V17.- Inteligencia emocional 3. Regulación V18.- Inteligencia emocional total V22.- VCPO 1. Vitalidad emocional V23.- VCPO 2. Estrés organizacional V24.- VCPO 3. Estrés emocional V25.- VCPO 4. Estrés físico	**V21. Energía Total.**

Cuadro 6.6: Experimentos con únicamente variables psicológicas.

6.4.1. V1 (Flow-Fluir)

Resultados obtenidos para la variable V1, utilizando únicamente las variables de carácter psicológico:

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,007367	0,582401	0,584896	0,009369	0,545863	0,546311	None
Nearest Neighbor Regression	0,009068	0,488040	0,492397	0,011183	0,457967	0,460508	n_neighbors: 6
SVM Regressor	0,007360	0,582267	0,585413	0,009777	0,526100	0,533549	Kernel: Linear, C: 0.1
Decision Tree Regressor	0,015990	0,090382	0,108656	0,014925	0,276581	0,286595	Criterion: mse
Random Forest Regressor	0,007545	0,572750	0,576149	**0,008295**	**0,597940**	**0,601470**	Criterion: friedman_mse, n_estimators: 500
MLP Regressor	0,007199	0,590843	0,593517	0,009512	0,538940	0,541743	Hidden_layers: 1x5, Activation: relu, Solver: lbfgs

Cuadro 6.7: Resultados de regresión para V1 sin variables personales.

Para este problema, como vimos en el apartado anterior. El mejor modelo ha sido **Random Forest**. Los resultados son mejores que los que vimos en el apartado anterior pero la mejora no es apenas significativa.

6.4.2. V9 (Flourishing-Florecimiento)

Resultados obtenidos para la variable V9, utilizando únicamente las variables de carácter psicológico:

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,008651	0,517296	0,518195	0,012155	0,330014	0,330217	None
Nearest Neighbor Regression	0,009293	0,487135	0,488901	0,011988	0,339180	0,341621	n_neighbors: 6
SVM Regressor	0,008000	0,551869	0,554791	0,011525	0,364719	0,364948	Kernel: rbf, C: 10
Decision Tree Regressor	0,015617	0,110082	0,114266	0,016673	0,080938	0,083772	Criterion: friedman_mse
Random Forest Regressor	0,008336	0,533882	0,536776	**0,010788**	**0,405348**	**0,405360**	Criterion: mae, n_estimators: 500
MLP Regressor	0,008189	0,542105	0,544398	0,011649	0,357852	0,359819	Hidden_layers: 2x5, Activation: Logistic, Solver: lbgfs

Cuadro 6.8: Resultados de regresión para V9 sin variables personales.

En esta variable V9, volvemos a ver como Random Forest queda como el mejor modelo para los datos de test. Obteniendo además las mejores métricas en todas las medidas. Sin embargo, la separación con el modelo de regresión lineal es apenas tangible. Aunque se hayan mejorado los resultados estos no son demasiado significativos.

6.4.3. V19 (Self Drive Direction)

Resultados obtenidos para la variable V19, utilizando únicamente las variables de carácter psicológico:

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,018029	0,428203	0,433408	0,016578	0,397951	0,402385	None
Nearest Neighbor Regression	0,019516	0,383023	0,390486	0,018341	0,333917	0,334953	n_neighbors: 10
SVM Regressor	0,017766	0,438165	0,442720	0,017025	0,381717	0,381899	Kernel: rbf, C: 1
Decision Tree Regressor	0,036713	-0,186308	-0,181692	0,028309	-0,028063	0,025826	Criterion: friedman_mse
Random Forest Regressor	0,018457	0,415450	0,419312	**0,016301**	**0,408024**	**0,409902**	Criterion: mae, n_estimators: 1000
MLP Regressor	0,017960	0,429890	0,435978	0,017602	0,360764	0,362264	Hidden_layers: 2x5, Activation: Logistic, Solver: lbgfs

Cuadro 6.9: Resultados de regresión para V19 sin variables personales.

Para esta variable V19, volvemos a ver a Random Forest como el mejor modelo en la parte de test. Incluso superando a la red neuronal, la cual se ha creado con 2 capas ocultas y 100 neuronas de tipo logística en cada una de ellas.

6.4.4. V20 (Consciencia energética)

Resultados obtenidos para la variable V20, utilizando únicamente las variables de carácter psicológico:

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,008138	0,618810	0,619340	0,010551	0,497817	0,501066	None
Nearest Neighbor Regression	0,009452	0,557061	0,563347	0,012421	0,408827	0,409514	n_neighbors: 8
SVM Regressor	0,008197	0,616437	0,616542	0,010507	0,499918	0,504682	Kernel: rbf, C: 0.1
Decision Tree Regressor	0,015967	0,249128	0,263865	0,021879	-0,041376	-0,028724	Criterion: mae
Random Forest Regressor	0,007824	0,634142	0,634865	0,010866	0,482836	0,486231	Criterion: mae, n_estimators: 1000
MLP Regressor	0,007941	0,628703	0,629120	**0,010537**	**0,498466**	**0,501684**	Hidden_layers: 3x2, Activation: Identity, Solver: lbgfs

Cuadro 6.10: Resultados de regresión para V20 sin variables personales.

En esta variable obtenemos los mejores resultados con la **red neuronal**, siendo esta bastante sencilla. De tres capas ocultas con dos neuronas en cada

una de ellas, además estas neuronas son de Identidad. Podemos ver en esta tabla que los resultados obtenidos incluso por el modelo **Random Forest**, que es el que nos ha dado los mejores resultados en otras variables queda por detrás de la regresión lineal. Lo que nos indica que tenemos bastantes dificultades para predecir esta variable.

6.4.5. V21 (Energía total)

Resultados obtenidos para la variable V21, utilizando únicamente las variables de carácter psicológico:

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,008583	0,599678	0,601921	0,009636	0,525836	0,529621	None
Nearest Neighbor Regression	0,009973	0,536047	0,547349	0,011356	0,441173	0,442264	n_neighbors: 9
SVM Regressor	0,008665	0,598494	0,600099	0,010061	0,504910	0,509355	Kernel: rbf, C: 0,5
Decision Tree Regressor	0,017885	0,161896	0,167702	0,017015	0,162739	0,167311	Criterion: friedman_mse
Random Forest Regressor	0,008521	0,603568	0,605339	**0,009718**	**0,521804**	**0,524046**	Criterion: mae, n_estimators: 1000
MLP Regressor	0,008489	0,605744	0,611369	0,009979	0,508956	0,510049	Hidden_layers: 5x50, Activation: Identity, Solver: Adam

Cuadro 6.11: Resultados de regresión para V21 sin variables personales.

Para esta variable observamos algo bastante sorprendente y es que el modelo de regresión lineal será el que ha obtenido los mejores resultados de entre todos los modelos propuestos. Esto es bastante extraño puesto que dicho modelo es el más simple de todos los modelos con los que hemos experimentado y es el que nos debería de dar los peores resultados. Podemos decir que esta variable sigue una distribución lineal en sus resultados.

6.4.6. Conclusiones

Se ha visto cómo se han obtenido mejores resultados que utilizando todas las variables del problema, pero este cambio en los resultados es apenas significativo para nosotros, puesto que no hay apenas una mejora notable en los resultados. La mayoría de los modelos siguen estando por debajo del corte de la regresión lineal y estos que la superan es por pocas centésimas o milésimas. Lo cual nos indica que estamos ante un problema bastante complejo de resolver. Sin embargo, con este enfoque hemos mejorado un poco los resultados. Algo muy curioso ha ocurrido con la última variable analizada, V21 (Energía total), y es que el mejor modelo ha sido la regresión lineal,

cuando en un primer momento incluimos dicho modelo debido a su simplicidad para que nos sirviera de corte en nuestros datos. Algo particularmente sorprendente para un modelo tan simple como este. Dándonos a entender que esta variable sigue una representación lineal en los datos.

Capítulo 7

Aplicación de técnicas de aprendizaje no supervisado

En esta sección aplicaremos técnicas de aprendizaje no supervisado a nuestros datos, con el fin de convertir nuestro problema de regresión en un problema de clasificación, creando clusters entre nuestros datos. Para ello aplicaremos dichas técnicas a todo el conjunto de datos (excepto las variables a predecir de nuestro problema) y después crearemos los distintos intervalos para la clasificación de nuestros datos. Estos clusters serán examinados, con el fin de ver la similitud que existan entre los individuos que lo forman. Una vez creados dichos clusters podremos pasar al siguiente paso de la clasificación, la aplicación de técnicas de aprendizaje supervisado multiclase a nuestro problema, sobre los clusters creados con un menor error. Se seleccionarán un total de 5 clusters para los algoritmos que necesiten saber este dato de antemano. Tras la realización de algunas pruebas, se ha decidido eliminar la variable género de nuestros datos para este problema debido a la creación de dos clústeres prácticamente inseparables por culpa de esta variable. Esto podrá verse en la primera sección de este apartado donde se realizará un PCA (Principal Component Analysis) y podrá apreciarse el resultado obtenido con dicha variable inclusive o sin ésta.

7.1. Visualización de los datos utilizando Principal Component Analysis (PCA)

Como ya explicamos en la sección 3 donde vimos los métodos utilizados en todo nuestro documento. Aplicaremos el método PCA (Principal Component Analysis) para poder mostrar nuestros datos en una representación 2D de los mismos. Para ello hemos escogido todas las instancias seleccionadas y las variables usadas en el anterior apartado para los métodos de regre-

sión. Hemos aplicado esta técnica y visualizando los resultados, obteniendo la siguiente imagen 7.1:

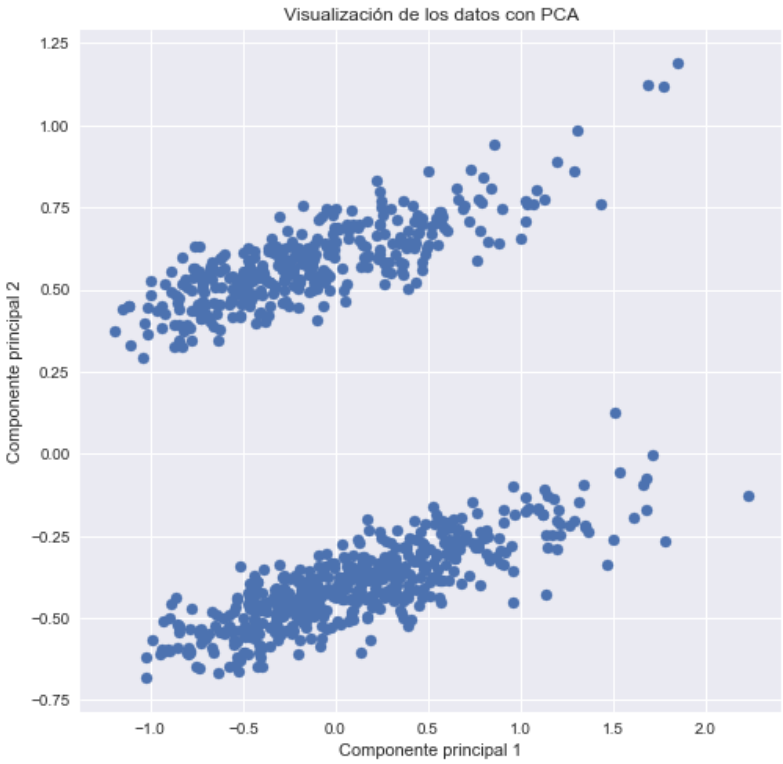

Figura 7.1: Visualización de los datos con PCA.

Como puede verse en esta imagen que nos muestra todas las instancias de nuestro problema, previamente transformadas a un plano 2D con una conversión PCA. Se pueden ver claramente 2 clusters en nuestros datos, así como algunas instancias más dispersas. Pero la división entre estos dos clusters es muy clara. Posteriormente analizaremos cuales son las instancias pertenecientes a cada cluster, para saber más de esta división en nuestros datos.

Quitando la variable género para que no se creen esos dos clusters, se ha obtenido la siguiente gráfica 7.2:

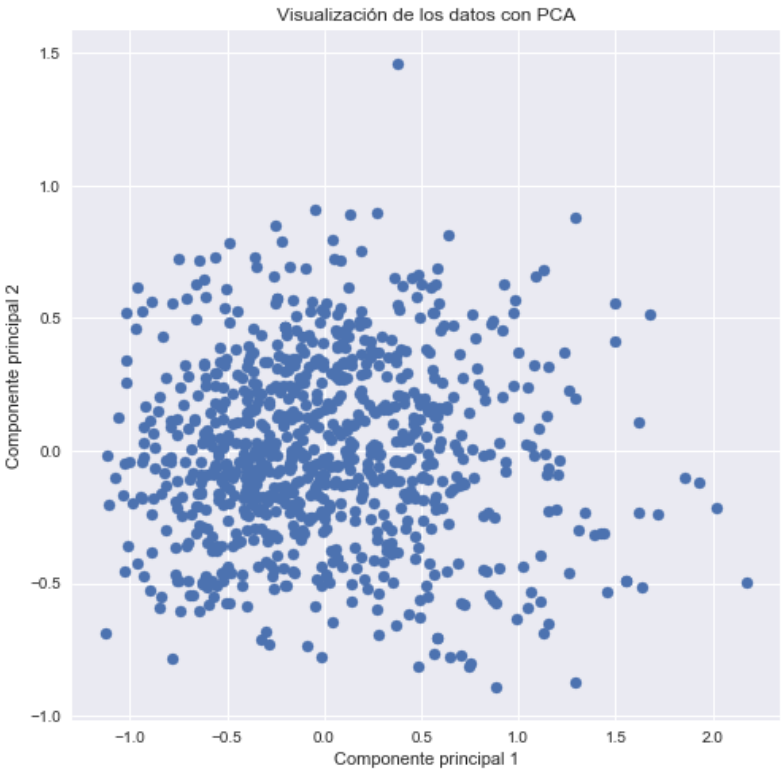

Figura 7.2: Visualización de los datos con PCA sin la variable Género.

Observando las dos gráficas, puede verse que la variable género influye mucho en nuestro análisis de componentes principales, siendo esta casi una de las mismas, debido a que todas las instancias de nuestros datos tienen o un género u otro. Por lo tanto, se crearán dos clusters, sin embargo, para nuestro análisis del problema dichos clusters no nos son de utilidad puesto que no nos aportan información adicional a nuestro problema. Puesto que una división por género es muy específica para los datos y lo que deseamos con estos problemas será realizar separaciones en base a los valores de las variables psicológicas.

Por ello y siguiendo con la lógica vista hasta ahora con el problema de regresión, se usarán únicamente las variables psicológicas en este análisis, como ya hicimos para la sección anterior, puesto que los resultados obtenidos con estas son mejores que usando todas las variables. El gráfico de nuestros datos aplicando PCA únicamente a las variables psicológicas puede verse en la siguiente imagen 7.3:

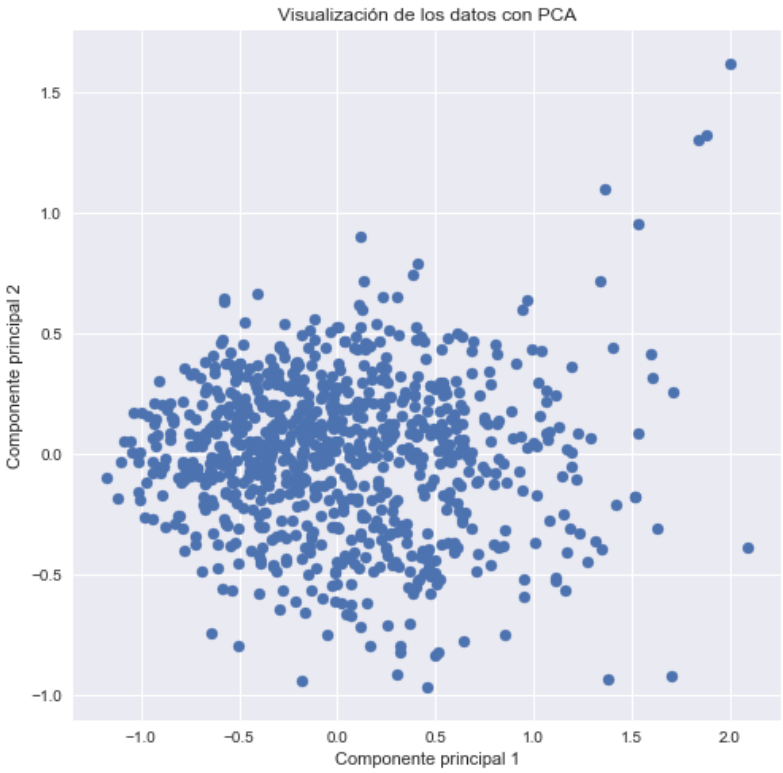

Figura 7.3: Visualización de los datos con PCA sin las variables sociales.

7.2. Métricas utilizadas para medir nuestros clusters

He utilizado las métricas que se definen a continuación al ser las únicas implementadas en la librería de scikit-learn que no necesitan de los valores reales de la clase a predecir. Puesto que para nosotros se trata de un problema no supervisado real, vamos a suponer que no tenemos los valores a los que pertenece cada instancia. Aunque se podrían comparar con los intervalos de dichas variables.

7.2.1. Silhouette Coefficient

Cuando no se conocen las clases reales, tenemos que evaluar el modelo en sí. Esta métrica[1] examinará las distancias medias que existen entre cada una de las instancias con las instancias del mismo clúster. Así como las distancias que existen con instancias de otro clúster. Comprobando que se está minimizando la distancia intra-cluster y maximizando la distancia

inter-cluster. Siendo la fórmula para la misma como sigue:

$$s = \frac{b - a}{max(a, b)} \tag{7.1}$$

Siendo:

- **a:** La distancia media entre una instancia y las instancias del mismo clúster.
- **b:** La distancia media entre una instancia y las instancias del clúster más próximo.

Será una medida a maximizar, siendo 1 su mayor puntuación posible.

7.2.2. Davies-Bouldin Index

Esta métrica[1] está enfocada a comprobar la separación que existe entre los distintos clusters. Para ello, se realizará el cálculo de la siguiente forma:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{7.2}$$

Siendo:

- s_i: La distancia media entre cada punto del cluster i y el centroide de cada cluster.
- d_{ij}: La distancia entre los centroides de los clusters i y j.

Definiendo entonces Davies-Bouldin Index:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \operatorname*{máx}_{i \neq j} R_{ij} \tag{7.3}$$

Es una medida a minimizar, siendo 0 el mejor valor posible para esta medida, indicándonos que los clusters están perfectamente separados.

7.3. Modelos escogidos y resultados

Los modelos fueron explicados en la sección 3 de este informe. Por lo que se pasará directamente a exponer los resultados obtenidos por los mismos.

7.3.1. K-Means

El resultado de los clusters obtenidos con este método es el siguiente:

Figura 7.4: Visualización de los clusters formados por KMeans.

Se puede observar en la gráfica 7.4 la distribución de los 5 clusters creados con este método. Siendo el clúster 1 el que tiene un número mayor de instancias y estando las mismas más concentradas. Sin embargo, el clúster 2 ha recogido las instancias más dispersas y alejadas de la aglomeración principal. Los clusters 4, 3 y 5 tienen bastantes instancias que se entrelazan entre sí por lo que dichos tres clusters podrían considerarse como uno solo. Aunque estas consideraciones son observando dicha gráfica, deberían de validarse conforme a las variables a predecir en nuestros datos. Sin embargo, puesto que estamos tratándolo como un problema no supervisado y en cuanto a los análisis realizados con anterioridad. En el clúster más disperso y más alejado del centro deben de concentrarse las instancias que han obtenido los valores más pequeños para las variables psicológicas, puesto que estas eran las más minoritarias y consideradas outliers por nuestro análisis numérico. Mientras que el cluster 1, el cual tiene el mayor número de instancias y están más cercanas y concentradas entre sí, deben de pertenecer los valores

altos de las variables, al ser un mayor número pero se mueven en un rango menor. Mientras que los otros clusters representan los valores medios de las variables.

7.3.2. Hierarchical Clustering

A continuación mostraremos las gráficas obtenidas para cada uno de los *linkage* distintos:

Ward En este gráfico 7.5 podemos apreciar una separación parecida a la comentada en el **KMeans**, sin embargo existe más solapamiento entre los clusters intermedios. Mientras que el cluster 4 tiene menos instancias que las vistas en KMeans para dicho clúster. Las instancias alejadas siguen perteneciendo al mismo clúster.

Figura 7.5: Visualización de los clusters formados por Hierarchical Clustering (Ward).

Complete linkage Se observa en la imagen 7.6 unos clusters más robustos frente a un par de clusters que apenas tienen instancias. Sigue existiendo el clúster de los outliers y el clúster 4, que creemos hace referencia

a los valores más altos de las variables. Mientras que los clusters que existían en medio se han visto devorados por el cluster 4 y unidos en uno solo, siendo este el clúster 1. Además de existir instancias que se solapan con otros clusters como el clúster 2. Siguiendo este esquema, seguramente existirían 3 tipos de clases en nuestros datos.

Figura 7.6: Visualización de los clusters formados por Hierarchical Clustering (Complete linkage).

Average linkage En este gráfico 7.7 puede verse la existencia clara de dos únicos clusters, uniendo el clúster 1 y 2. Los cuales tienen pocas instancias cada uno de ellos. Mientras que el clúster que concentraba la mayoría de las instancias ha tomado casi todo el poder convirtiéndose así en el clúster 3, siendo este el más grande.

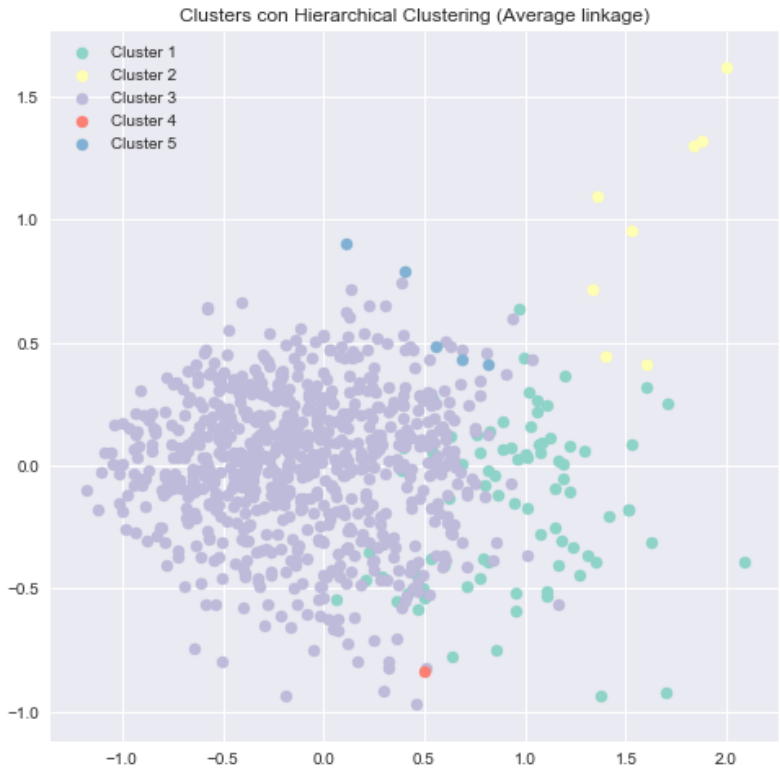

Figura 7.7: Visualización de los clusters formados por Hierarchical Clustering (Average linkage).

Single linkage Observamos en la imagen 7.8 que se ha tomado como consideración un único clúster en nuestros datos, puesto que los demás serán los centroides. Viendo esta representación no tendremos en cuenta esta opción debido a lo obvio. Aunque cabe preguntarse si no existirá únicamente un único clúster en nuestros datos o estos serán difícilmente separables.

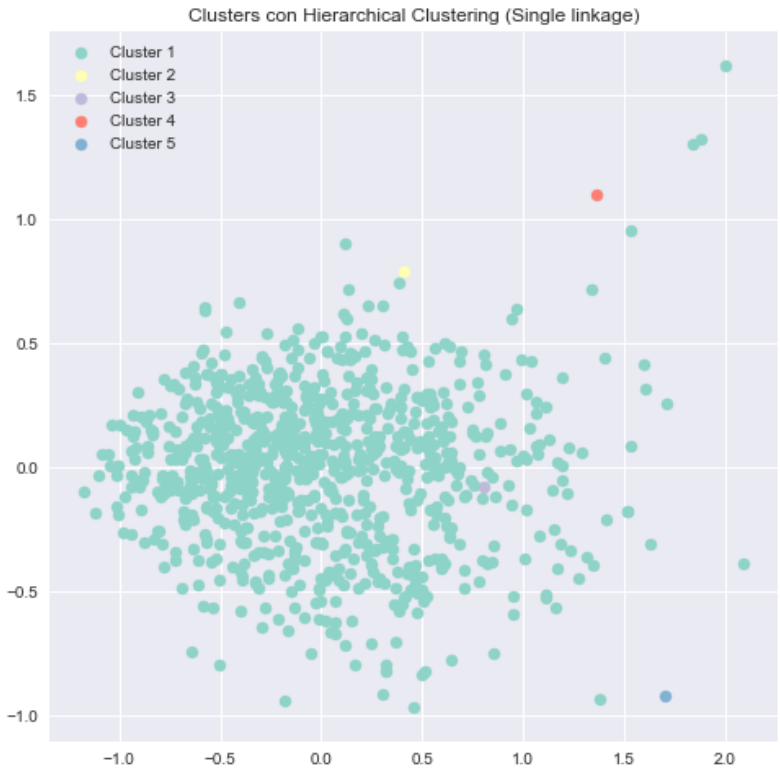

Figura 7.8: Visualización de los clusters formados por Hierarchical Clustering (Single linkage).

7.3.3. DBSCAN

A continuación se muestran en la figura 7.9 los clusters creados por este método para nuestros datos:

Figura 7.9: Visualización de los clusters formados por DBSCAN.

Este método que buscará él mismo los clusters a crear usando para ello las fronteras de baja densidad de nuestras instancias, nos ha dado como resultado la creación de únicamente 3 clusters en nuestros datos, siendo nosotros los que escogimos 5 para tener una mayor variedad a la hora de clasificar. Sin embargo, observando la imagen puede verse que el clúster considerado es el que hablábamos al principio de la sección como el clúster más centrado y con instancias más relacionadas. Mientras que ha tomado todas las demás instancias como outliers. Si algo podemos sacar en claro de esta aproximación es la existencia de una gran relación entre las instancias que consideramos con un alto grado en todas las variables psicológicas, mientras que las demás serán más difíciles de predecir.

7.4. Conclusiones

Como punto final de este apartado veremos los resultados de las distintas medidas para cada uno de los métodos aplicados. Dichas medidas se han obtenido para todos los clusters que hemos ido viendo en este apartado. Para nuestros modelos hemos decidido seleccionar la creación de 5 clusters. Siendo las medidas mostradas a continuación las relacionadas con dichas creaciones:

	Silhouette Score ↑	**Davies-Bouldin Index ↓**
K-Means	0,125416	2,022405
DBSCAN	0,026838	2,334281
Hierarchical Clustering (Ward)	0,067756	2,436834
Hierarchical Clustering (Complete)	0,158427	1,640709
Hierarchical Clustering (Average)	**0,198196**	1,260211
Hierarchical Clustering (Single)	0,191764	**0,533501**

Cuadro 7.1: Resultados obtenidos por los métodos no supervisados.

Comentaremos a continuación los resultados de la tabla 7.1 obtenidos para cada uno de los modelos usados en la creación de clusters. Sobre la métrica de Silhouette Score, la cual será a maximizar, obtenemos que el mejor modelo ha sido **Hierarchical Clustering (Average)**, esto se puede explicar debido a que como vimos en el gráfico mostrado para este método se creaban prácticamente dos clusters. Uno muy grande que agrupaba la mayoría de las instancias y otros dos más pequeños, los cuales agrupaban los outliers de nuestro problema o las instancias más alejadas del centro. Por otro lado, para la métrica de Davies-Bouldin la cual será a minimizar, el mejor modelo ha sido el de **Hierarchical Clustering** con el modo **Single**, mirando hacia atrás pudimos ver que esta metodología de creación de clusters también nos creó un único cluster y los demás eran únicamente los centroides de los demás clusters. Como conclusión en este apartado, aunque las métricas nos indiquen que los datos no son realmente separables debido a las malas puntuaciones obtenidas en ambas. Además, siendo los métodos que han considerado prácticamente un único clúster en nuestros datos los que han obtenido mejores resultados en estas métricas. Volviendo atrás y observando las gráficas obtenidas por los modelos **KMeans** y **Hierarchical clustering (Ward)**, puede verse esa pequeña distribución de nuestras instancias como comentamos en cada una de las secciones. Aunque esta distribución no ha sido apoyada por los resultados obtenidos, visualmente son las que tienen un mayor sentido. Lo que nos vienen a decir las métricas es que este será un problema muy difícil de clasificar y separar, puesto que la mayoría de las instancias están muy juntas entre sí. Por lo tanto para el apartado de clasificación será muy difícil crear barreras entre los datos para distinguirlos en las diferentes clases.

Capítulo 8

Uso de modelos de clasificación multiclase

8.1. Creación de intervalos en nuestros datos

Debido al fracaso obtenido por los modelos no supervisados para la creación de clusters en nuestros datos, en cuanto a la transformación de nuestro problema de regresión en uno de clasificación multiclase, lo que haremos a continuación será dividir nuestras variables a predecir de regresión en intervalos, creando con ello un problema de clasificación multiclase. Primeramente, mostraremos unos histogramas para cada una de las variables para ver la distribución de las clases en estos problemas. La división se realizará de la forma mostrada en la tabla 8.1, ya que tenemos los datos normalizados, la división de los mismos se podrá realizar de una forma más sencilla.

Rango	**Valor**
0-0.2	0 (Muy bajo)
0.2-0-4	1 (Bajo)
0.4-0.6	2 (Medio)
0.6-0.8	3 (Alto)
0.8-1	4 (Muy alto)

Cuadro 8.1: Intervalos y valores para las clases creadas.

De esta forma tendremos las variables psicológicas a predecir dividas por intervalos que tienen sentido lógicamente hablando, puesto que las mismas miden la intensidad de dichas características del individuo y a mayor valor en dicha variable, mayor será la intensidad de ese campo para dicho individuo.

8.1.1. Visualización de la distribución de clases por variable

Una vez realizada la conversión de las variables en intervalos, mostraremos a continuación histogramas para cada una de las variables a predecir, lo cual nos mostrará la distribución de las distintas clases en nuestro problema.

V1 (Flow-Fluir)

A continuación mostraremos la distribución de las clases para la variable a predecir V1, la cual representa el Flow o Fluir de una persona.

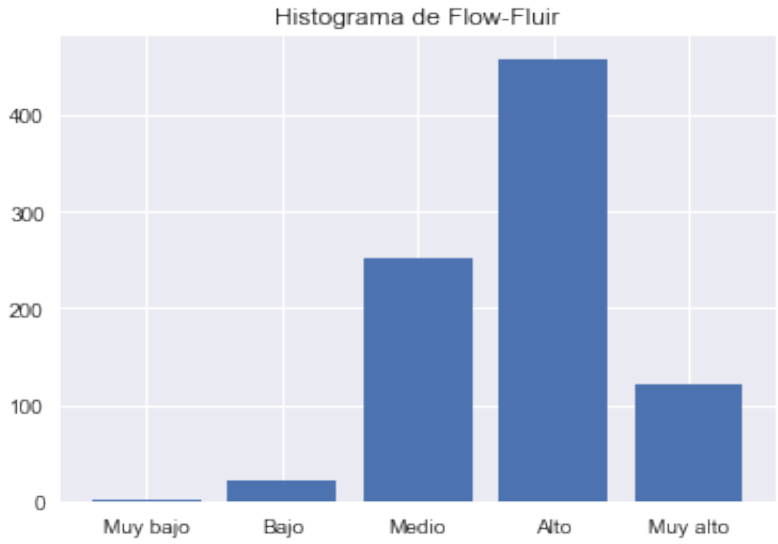

Figura 8.1: Histograma de las clases asignadas a Flow-Fluir.

Puede observarse en el histograma 8.1 según la distribución anterior que estamos ante un problema un tanto desbalanceado, puesto que para las clases 0 (Muy bajo) y 1 (Bajo), apenas hay representación en nuestros datos. La mayoría de la representación está contenida en las clases 2 (Medio) y 3 (Alto). Observando que todos los individuos encuestados tienen una alta capacidad en esta variable psicológica.

A continuación lo mostraremos representado en un mapa 2D gracias a un análisis por PCA, como ya hicimos en la sección del análisis no supervisado. Usando únicamente las variables psicológicas:

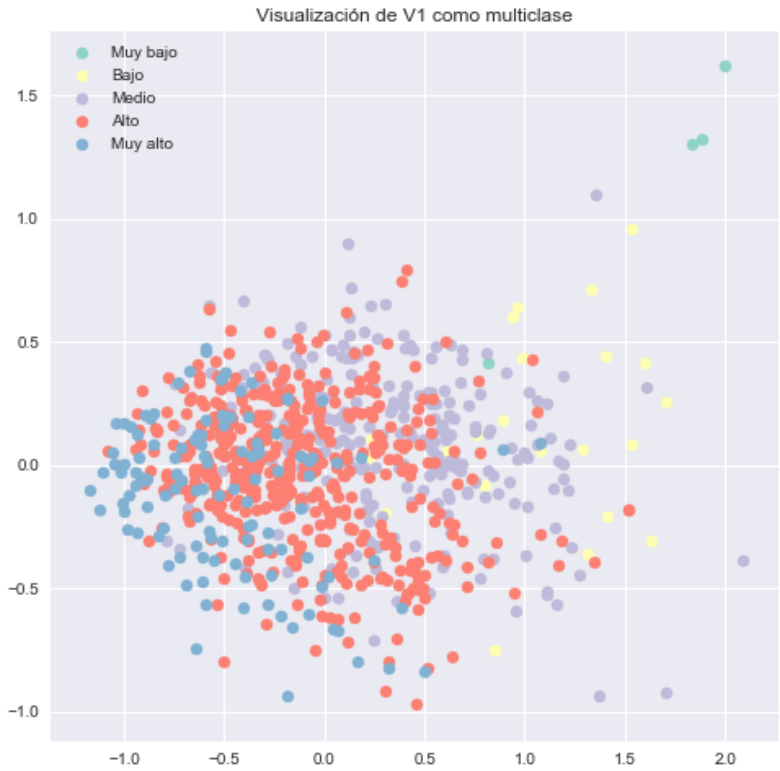

Figura 8.2: Visualización PCA sobre Flow-Fluir multiclase.

Como se observa en la anterior imagen 8.2 podemos confirmar lo comentado en los análisis no supervisados realizados. Observando como las instancias más alejadas del problema corresponden a los valores "Muy bajo" o "Bajo" de los mismos. Mientras que cuanto más a la izquierda vamos más suelen subir estos valores. Siendo en este caso el mayor grupo el de "Alto", el cual está entremezclado con los valores "Medio" y "Muy alto", aunque existen ciertas instancias que son claramente diferenciables usando únicamente las demás variables psicológicas.

8.1.2. V9 (Flurishing-Florecimiento)

A continuación hablaremos de la variable V9, mostrando su distribución de clases así como comentando la misma.

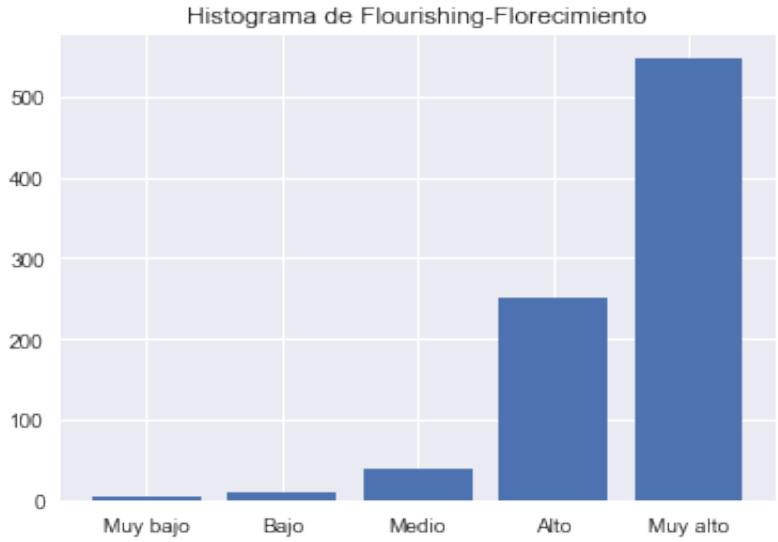

Figura 8.3: Histograma de las clases asignadas a Flourishing-Florecimiento.

Como puede observarse en el histograma anterior 8.3, estamos ante un problema claramente no balanceado que tiende a tener más representación en las clases más altas. Siendo la clase "Muy alto"la que tenga una mayor representación de casi más de 500 instancias. Mientras que las variables de las clases bajas apenas tienen instancias que las representen. Dándonos a entender que la gran mayoría de los individuos encuestados tienen una gran capacidad de Florecimiento.

A continuación el mapa usando PCA:

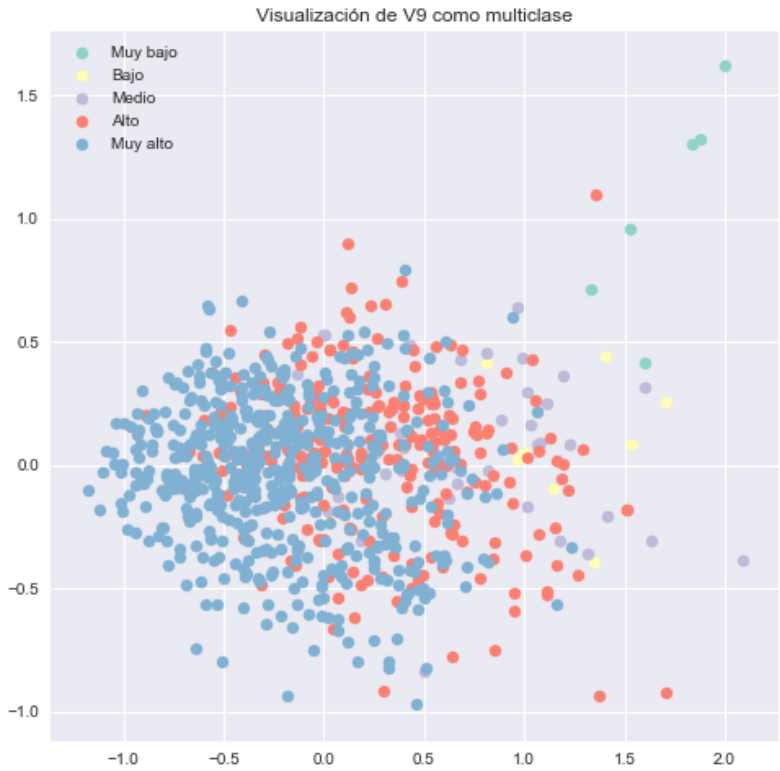

Figura 8.4: Visualización PCA sobre Flourishing-Florecimiento multiclase.

En este mapa 8.4 podemos observar como las instancias más representadas son las pertenecientes a la clase "Muy alto". Podría ser que debido a los rangos empleados, deberíamos de aumentar el rango mínimo de esta clase. Puesto que la mayoría de las instancias están concentradas en esa clase. También puede ser que sea una variable en la que la mayoría de los individuos hayan obtenido una muy buena puntuación.

8.1.3. V19 (Self Drive Direction)

A continuación hablaremos y mostraremos un histograma con la distribución de las clases para la variable V19 (Self Drive Direction), la cual es la primera variable sobre inteligencia energética a predecir en nuestros datos.

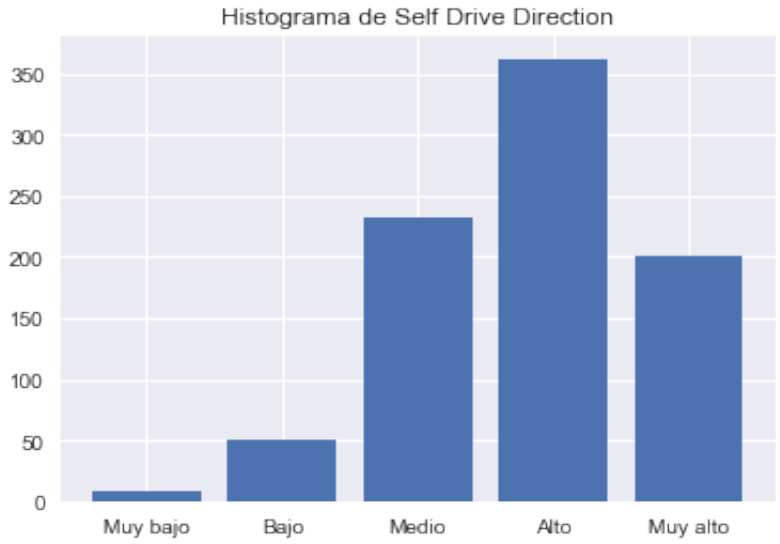

Figura 8.5: Histograma de las clases asignadas a Self Drive Direction.

Observando el histograma anterior 8.5, se puede ver lo dicho en las anteriores variables, estamos ante un problema no balanceado puesto que la mayoría de las instancias están repartidas entre las clases "Medio", .Altoz "Muy alto". Siendo la clase .Alto"la que tiene un mayor número de instancias. Sin embargo, la representación de las clases bajas en este problema es más alta que en las variables anteriores.

A continuación el mapa usando PCA:

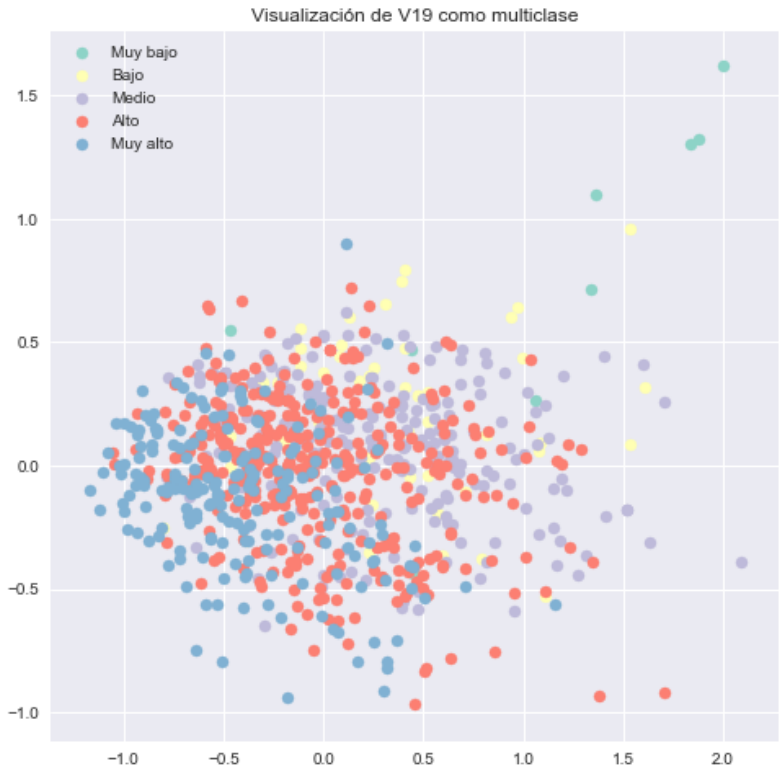

Figura 8.6: Visualización PCA sobre Self Drive Direction multiclase.

En este mapa 8.6 si existe una interacción mayor entre las clases, estando estas más mezcladas. Pero seguimos viendo la pauta que dijimos en el análisis no supervisado. Estando los valores altos de esta variable a la izquierda concentrados. Mientras que cuanto menor es el valor más a la derecha y dispersos están los mismos. Siendo el centro un aglomerado de clases "Alto" o "Medio". Por lo que resultará muy difícil de predecir para nuestro clasificador.

8.1.4. V20 (Consciencia energética)

A continuación hablaremos y comentaremos un poco sobre la distribución de las clases para la variable V20, segunda variable del apartado de inteligencia energética.

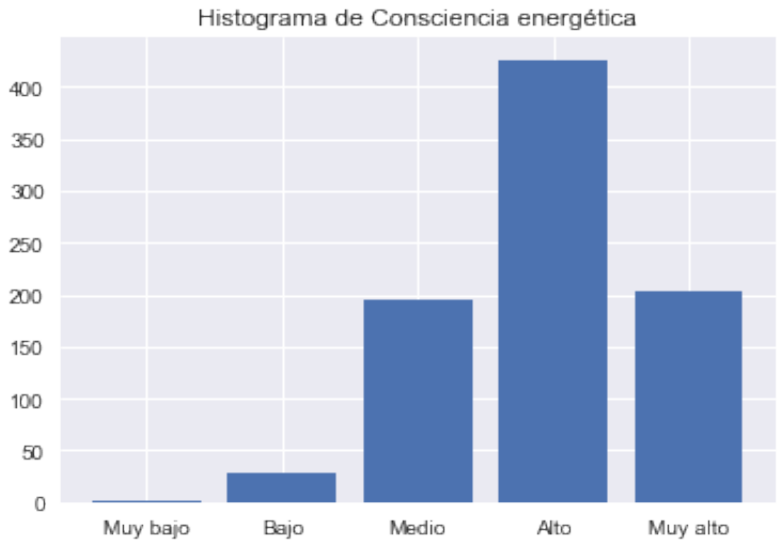

Figura 8.7: Histograma de las clases asignadas a Consciencia energética.

Esta distribución, como se aprecia en el histograma 8.7, sigue una pauta parecida a la vista en la variable V19, aunque existen más representación para la clase .Alto". Mientras que apenas existen instancias que nos muestran la clase "Muy bajo". La mayoría de las instancias están concentradas en altas capacidades para esta variable de consciencia energética entre los usuarios encuestados.

A continuación el mapa usando PCA:

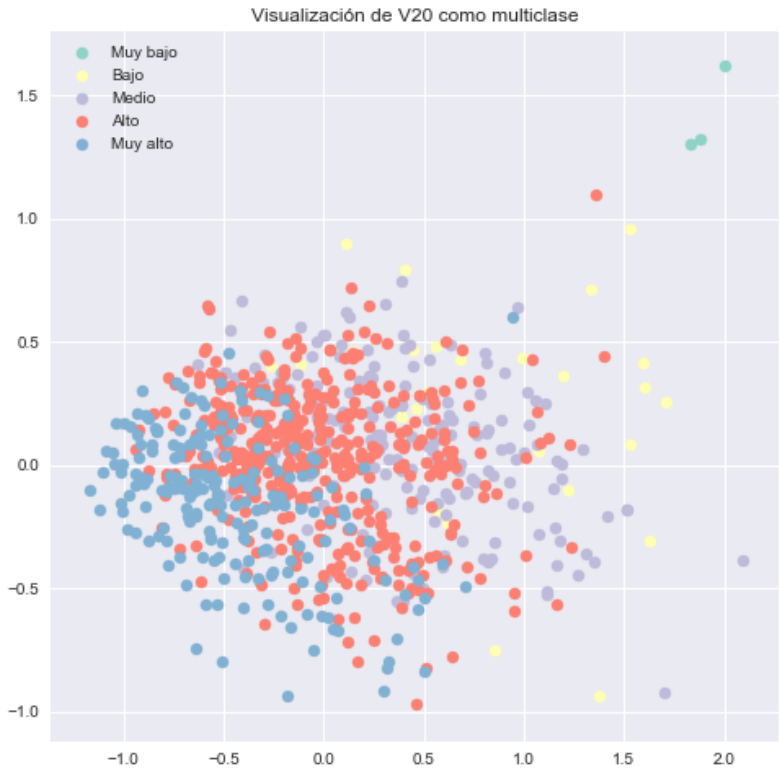

Figura 8.8: Visualización PCA sobre Consciencia energética multiclase.

Como se puede observar en esta imagen 8.8, para esta variable los datos si están más repartidos entre las clases. A excepción como siempre de las clases más bajas, las cuales están compuestas por las instancias más alejadas. Podemos ver como en el centro de nuestros datos están concentradas las clases "Medio" y "Alto", mientras que las instancias pertenecientes a la clase "Muy Alto" están más alejadas del centro hacia la izquierda. Este problema si tendrá una más fácil diferenciación que las variables anteriores.

8.1.5. V21 (Energía total)

Por último comentaremos la distribución para la variable V21 que representa la energía total del individuo. Esta distribución será igual o parecida a las dos anteriores, puesto que esta variable es la suma de ambas.

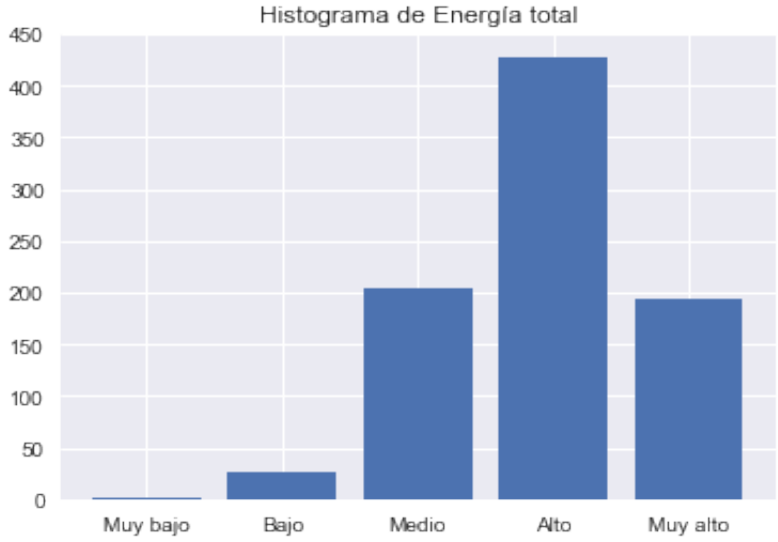

Figura 8.9: Histograma de las clases asignadas a Energía total.

Como puede verse en el histograma 8.9, ocurre lo mencionado anteriormente. La mayoría de la población está repartida entre las tres clase más altas, dejando a las clases bajas sin apenas representación como puede verse para las clases "Muy bajoz "Bajo".

A continuación el mapa usando PCA:

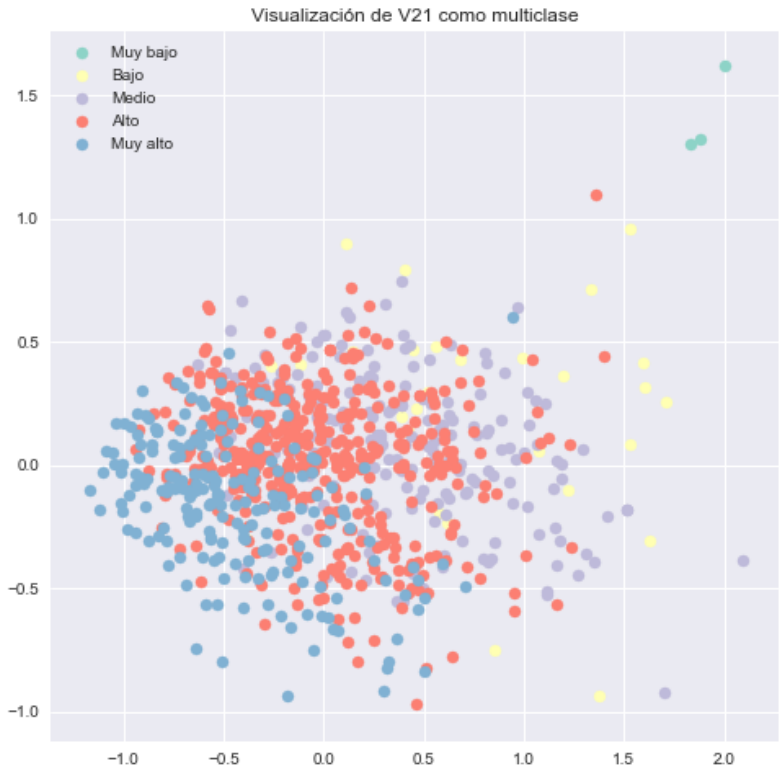

Figura 8.10: Visualización PCA sobre Energía total multiclase.

Esta imagen 8.10 representa prácticamente lo mismo que lo observado en las dos variables anteriores, puesto que será una agregación de las mismas. Podemos ver como la clase "Muy Alto" será bastante separable de la clase "Alto" . Mientras que existirán clases más entremezcladas entre sí como son las clases "Alto" y "Medio".

8.2. Aplicación de técnicas de aprendizaje supervisado (multiclase)

En este apartado, una vez cambiado nuestro problema de regresión a un problema de clasificación multiclase aplicaremos métodos multiclase para la predicción de nuestro problema. La forma de aplicar estos métodos será la misma que usamos en el apartado de regresión. Usando para ello el modelo de experimento comentado en la sección 5. A continuación hablaremos de los modelos escogidos así como de las métricas usadas para clasificar dichos modelos. Para estos experimentos también se realizarán los dos experimentos mencionados en el apartado de regresión. Usando todas las variables

mencionadas y luego quitando las variables de carácter personal de nuestros datos, para ver si quitando dichas variables nuestros modelos son capaces de obtener unos mejores resultados.

8.2.1. Modelos escogidos

Los modelos escogidos en este apartado han sido los mismos que los utilizados en la sección de regresión. Por lo tanto no serán explicados y solamente serán listados. La búsqueda de parámetros también será la misma para estos modelos, usando para ello la técnica de búsqueda en rejilla. Estos modelos son los listados a continuación:

- **Support Vector Machine (SVM)**
- **Nearest Neighbor (kNN)**
- **Decision Tree**
- **Random Forest**
- **Multilayer Perceptron**

8.2.2. Métricas usadas

Como vimos en el apartado anterior nos encontramos ante problemas de tipo no balanceado por lo que las métricas a usar deberán de representar esto, no nos fiaremos únicamente de la métrica *Accuracy*, si no que usaremos para la interpretación de los resultados otro tipo de métricas.

Accuracy Score

Nos devuelve la precisión (Accuracy)[32] de nuestros modelo, es decir, el porcentaje de aciertos de nuestro modelo. Observando para ello la clase real y la clase predicha. Lo malo de esta métrica es el sesgo que tiene hacia problemas no balanceados. La fórmula por la que se calcula esta métrica es la siguiente:

$$accuracy(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \tag{8.1}$$

Recall

Recall[10] es el ratio $TP/(TP + FN)$ donde TP es el número de verdaderos positivos y FN es el número de falsos negativos. Será la habilidad del

clasificador de encontrar todos los casos positivos. Para nuestro problema, al ser este multiclase, usaremos la aproximación *Macro*.

Hamming Loss

Hamming Loss[21] nos calculará la distancia hamming existente entre los valores reales de nuestros datos y los valores predichos por nuestro modelo. Dicha distancia hamming es calculada de la siguiente manera:

$$L_{Hamming}(y, \hat{y}) = \frac{1}{n_{labels}} \sum_{j=0}^{n_{labels}-1} 1(\hat{y}_j \neq y_j) \tag{8.2}$$

Precision

Precision[10] es el ratio $TP/(TP + FP)$ donde TP es el número de verdaderos positivos y FP es el número de falsos positivos. Medirá la capacidad que tiene nuestro modelo para no marcar una clase positiva como negativa. Usaremos la aproximación *Macro* en este problema multiclase.

8.2.3. Resultados con todas las variables del problema

En este apartado comentaremos los resultados obtenidos para cada una de las variables como ya hicimos en el apartado de regresión. En este apartado se han utilizado todas y cada una de las variables del problema. Se mostrarán tanto los resultados de entrenamiento como de test.

Los resultados obtenidos para nuestro problema con todas las variables podrán ser consultados en el apéndice de este informe.

8.2.4. Resultados sin usar las variables sociales

Como ya vimos en la sección anterior de regresión, volveremos a repetir los experimentos. Esta vez únicamente usando las variables psicológicas como entrada, es decir, sin usar las variables de tipo social o personal de los encuestados.

V1 (Flow-Fluir)

Los resultados obtenidos para esta variable sin las entradas sociales son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,652179	0,482755	0,561639	0,347821	0,627907	0,489416	0,747019	0,372093	kernel: linear, C: 10
kNN	0,175660	0,084199	0,118585	0,074260	0,587209	0,377096	0,544471	0,412791	k: 17
Decision Tree	0,154102	0,127478	0,146377	0,093898	0,540698	0,492626	0,593977	0,459302	Criterion: gini
Random Forest	0,681355	0,506213	0,623433	0,318645	**0,639535**	**0,456988**	**0,524155**	**0,360465**	Criterion: entropy, n_estimators: 1000
MLP	0,665210	0,457166	0,503261	0,334790	0,616279	0,461247	0,590450	0,383721	Hidden_layers: 5x100, Activation: tanh, Solver: Adam

Cuadro 8.2: Resultados multiclase para V1 sin las variables sociales.

Para la variable V1 como se puede observar en la tabla 8.2, hemos obtenido como mejor modelo en la fase de generalización a **Random Forest**. Aunque la mejor métrica obtenida en el apartado de *Precision* la ha dado **SVM**. Con lo cual podemos suponer que el modelo **SVM** ha clasificado mejor las clases minoritarias en nuestro problema. Además, puede verse también un descenso en los resultados de generalización, por lo que podemos suponer un poco de sobreentrenamiento en nuestros modelos. Siendo además la arquitectura de esta red neuronal una de las más complejas de las obtenidas hasta el momento, puesto que cuenta con 3 capas de 100 neuronas cada una, siendo estas de activación tanh.

V9 (Flourishing-Florecimiento)

Los resultados obtenidos para esta variable sin las entradas sociales son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,741209	0,346687	0,299426	0,258791	0,656977	0,319167	0,299426	0,343023	kernel: poly, C: 10
kNN	0,185939	0,116125	0,098955	0,063661	0,662791	0,318852	0,300067	0,337209	k: 15
Decision Tree	0,177682	0,115536	0,100250	0,071918	0,668605	0,312298	0,311123	0,331395	Criterion: gini
Random Forest	0,754390	0,462470	0,548427	0,245610	**0,703488**	**0,319134**	**0,462622**	**0,296512**	Criterion: gini, n_estimators: 500
MLP	0,723465	0,372216	0,366864	0,276535	0,668605	0,344022	0,553734	0,331395	Hidden_layers: 5x100, Activation: Identity, Solver: Adam

Cuadro 8.3: Resultados multiclase para V9 sin las variables sociales.

Para esta variable V9, como podemos observar en la tabla 8.3, hemos obtenido el mismo mejor modelo en ambas fases, de entrenamiento y de test, siendo este modelo **Random Forest**. A pesar de ser el mejor modelo en *Accuracy*, ha sido la **red neuronal**, esta vez de 5 capas ocultas con 100 neuronas en cada una de ellas de tipo Identidad, la que ha obtenido los mejores resultados en *Recall* y *Precision*, aunque estos resultados son bastante bajos para todos los modelos en esta variables. Por lo que se deberán de intentar mejorar en otros apartados del problema.

V19 (Self Drive Direction)

Los resultados obtenidos para esta variable sin las entradas sociales son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,536518	0,393306	0,433577	0,463482	0,552326	0,326275	0,355532	0,447674	kernel: rbf, C: 10
kNN	0,132905	0,083642	0,082899	0,115095	0,529070	0,314285	0,339751	0,470930	k: 19
Decision Tree	0,118015	0,107970	0,115413	0,131585	0,511628	0,352911	0,342790	0,488372	Criterion: gini
Random Forest	0,546726	0,412202	0,533958	0,453274	0,558140	0,331228	0,356853	0,441860	Criterion: gini, n_estimators: 500
MLP	0,495663	0,307858	0,341500	0,504337	**0,563953**	**0,345840**	**0,342798**	**0,436047**	Hidden_layers: 5x100, Activation: Identity, Solver: Adam

Cuadro 8.4: Resultados multiclase para V19 sin las variables sociales.

Para esta variable V19, como podemos observar en la tabla 8.4, hemos obtenido un cambio en los resultados y es que la **red neuronal** ha obtenido los mejores resultados en la fase de generalización. Siendo la arquitectura de esta de 5 capas ocultas con 100 neuronas de tipo tanh en ellas. Además, son interesantes estos resultados puesto que han mejorado a los obtenidos en la fase de entrenamiento. En esta variable los resultados han empeorado en comparación con la predicción usando todas las variables del problema.

V20 (Consciencia energética)

Los resultados obtenidos para esta variable sin las entradas sociales son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,647692	0,428673	0,451929	0,352308	0,616279	0,437144	0,475289	0,383721	kernel: poly, C: 10
kNN	0,175182	0,141323	0,167053	0,074418	0,587209	0,405146	0,449224	0,412791	k: 9
Decision Tree	0,142837	0,144793	0,169958	0,107083	0,622093	0,540029	0,540359	0,377907	Criterion: entropy
Random Forest	0,672542	0,551271	0,691608	0,327458	**0,656977**	**0,462491**	**0,507855**	**0,343023**	Criterion: gini, n_estimators: 1000
MLP	0,602351	0,429656	0,521212	0,397649	0,610465	0,481901	0,692955	0,389535	Hidden_layers: 5x100, Activation: tanh, Solver: Adam

Cuadro 8.5: Resultados multiclase para V20 sin las variables sociales.

Para esta variable V20, como podemos observar en la tabla 8.5, nos volvemos a encontrar con lo mencionado en las variables anteriores, volvemos a ver a **Random Forest** siendo el mejor tanto en la fase de entrenamiento como en la fase de Test. Aunque es el mejor en cuanto a *Accuracy*, hay otros modelos con mejores resultados en *Recall* y *Precision*. Los resultados obtenidos tanto en la fase de entrenamiento como en la fase de Test son bastante parecidos, por lo que no se aprecia un sobreentrenamiento para estas variables.

V21 (Energía total)

Los resultados obtenidos para esta variable sin las entradas sociales son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,640425	0,426733	0,460471	0,359575	0,662791	0,471259	0,517594	0,337209	kernel: poly, C: 10
kNN	0,162816	0,110833	0,128177	0,087104	0,616279	0,417864	0,488040	0,383721	k: 17
Decision Tree	0,144747	0,149641	0,147519	0,104853	0,575581	0,420340	0,411512	0,424419	Criterion: entropy
Random Forest	0,649098	0,504732	0,609018	0,350902	**0,691860**	**0,542893**	**0,784028**	**0,308140**	Criterion: entropy, n_estimators: 100
MLP	0,660842	0,518109	0,547735	0,339158	0,686047	0,504949	0,519287	0,313953	Hidden_layers: 5x20, Activation: tanh, Solver: Adam

Cuadro 8.6: Resultados multiclase para V21 sin las variables sociales.

V21 será la energía total del individuo, es decir es el resultado de la suma de las variables V19 y V20. Por lo que podemos considerarla como su resumen. Como se puede observar en la tabla 8.6, en la fase de test, se ha obtenido como mejor modelo **Random Forest**. Además también ha obtenido una *Precision* sobresaliente, superando el 75 %. Mientras que el *Accuracy* está cerca del 70 %, a pesar de estos buenos resultados, ha empeorado los resultados obtenidos por el experimento con todas las variables del problema.

Conclusiones

Como conclusión principal cabe destacar la mejoría en nuestros modelos cuando únicamente usamos las variables psicológicas como entradas para estos, siendo visible esta mejora únicamente en las variables V1 y V9, es decir, las variables que no tienen relación con la inteligencia energética. También cabe mencionar los buenos resultados obtenidos por el algoritmo **Random Forest**, el cual ha obtenido los mejores resultados en la mayoría de las variables, tanto en entrenamiento como en Test (los mejores resultados en Accuracy). Mientras que los mejores resultados en otras métricas han sido obtenidos por modelos más simples como el **Decision Tree**. La red neuronal siempre iba detrás de Random Forest en cuanto a rendimiento, siendo casi siempre el segundo mejor modelo para los distintos problemas afrontados. Obteniendo aún así grandes resultados para todos los problemas a pesar de tener configuraciones simples en algunos de ellos. Con este acercamiento a los datos en forma de una predicción multiclase hemos visto que es un problema bastante complicado de predecir, puesto que apenas se ha superado el 75 % de predicción en el mejor caso para los distintos problemas observados. Además de existir el problema mencionado anteriormente, en el que se comenta el desbalanceo de los datos con el sistema seguido por intervalos para transformar este problema de regresión en uno multiclase. Este pro-

blema ha podido ser aún más observable cuando hemos visto las distintas métricas mostradas en este apartado como *Recall* o *Precision*. En las que nuestros modelos han fracasado estrepitosamente. Debido probablemente a que obvian la clase minoritaria del problema y se centran únicamente en la clase mayoritaria. Obteniendo por ello los resultados vistos en las tablas de esta sección. Se deberá de intentar atajar el problema de otra forma, usando para ello una selección de características, así como una eliminación de posible ruido en nuestros datos. Todo esto podrá resultar en mejores resultados en las predicciones.

8.2.5. Representación de los árboles creados por Random Forest

A continuación mostraremos algunos de los árboles creados por el modelo de **Random Forest**, comentando para cada una de las variables el proceso que han seguido para clasificar sus instancias, así como las variables más importantes para cada uno de los problemas y algunas breves conclusiones obtenidas de la observación de estos árboles. Aunque solo se muestren algunos ejemplos de estos árboles, para la escritura de las conclusiones fueron consultados más de 5 árboles por variable.

V1 (Flow-Fluir)

A continuación mostraremos algunos ejemplos de los obtenidos por nuestro modelo Random Forest para la variable V1. Más de estos árboles han sido analizados para llegar a las siguientes conclusiones:

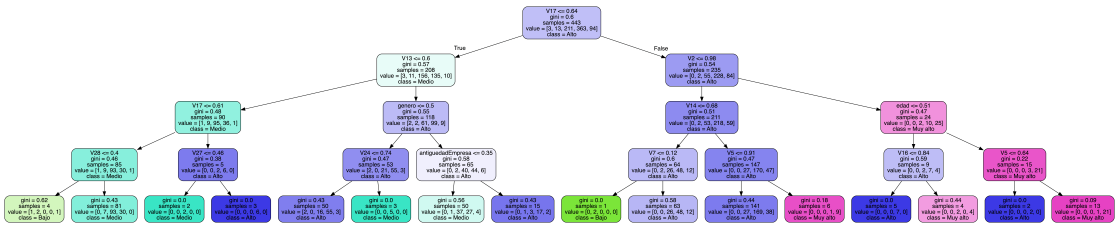

Figura 8.11: Ejemplo de Decision Tree para la variable V1 (1).

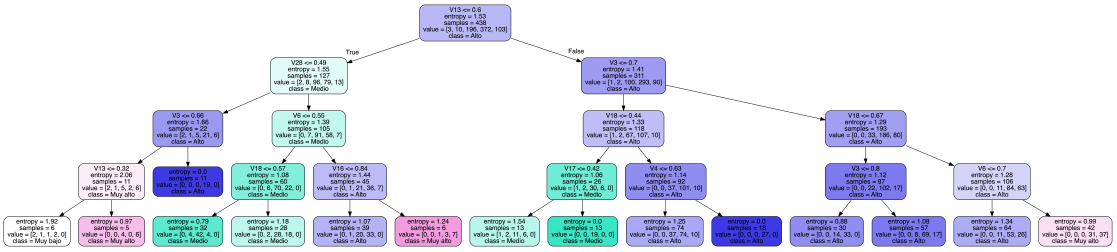

Figura 8.12: Ejemplo de Decision Tree para la variable V1 (2).

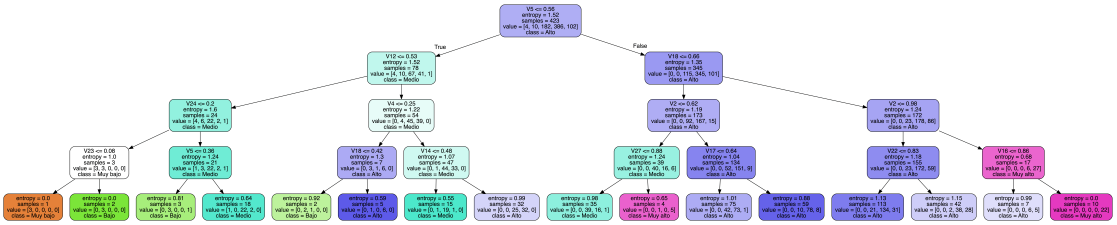

Figura 8.13: Ejemplo de Decision Tree para la variable V1 (3).

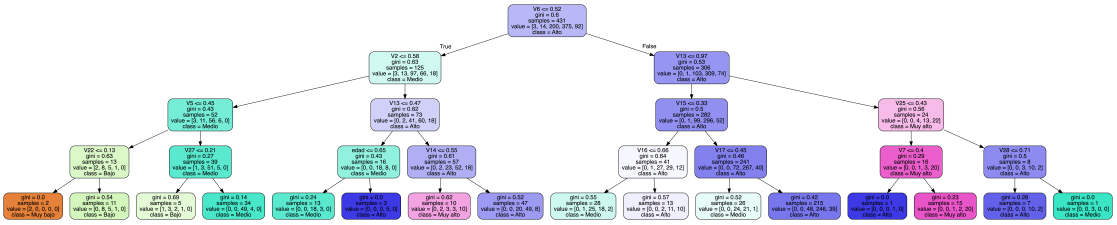

Figura 8.14: Ejemplo de Decision Tree para la variable V1 (4).

Como se puede observar en los árboles mostrados y los analizados 8.11, 8.12, 8.13 y 8.14. Para esta variable, normalmente se usan las variables V3 (Autoestima), V13 (Engagement total), V16 y V17 (variables pertenecientes a la inteligencia emocional). Como los primeros nodos para hacer las grandes separaciones entre las clases. Sin embargo, cuanto más bajamos en el árbol y sobretodo para los últimos nodos. Se utilizan las variables personales (edad o antigüedad en el cargo sobretodo), o las variables pertenecientes a la Calidad Personal y Organizacional de los encuestados. Variables que nos dan un vistazo mejor sobre individuo. Estas se usarán para hacer las últimas separaciones.

Por lo que podemos observar que esta variable está altamente relacionada con la autoestima y la percepción de la inteligencia emocional. Además cuando usa la edad, normalmente cuanta más edad suele llevar a las clases más altas de esta variable.

V9 (Flourishing-Florecimiento)

A continuación mostraremos algunos ejemplos de los obtenidos por nuestro modelo Random Forest para la variable V9. Más de estos árboles han sido analizados para llegar a las siguientes conclusiones:

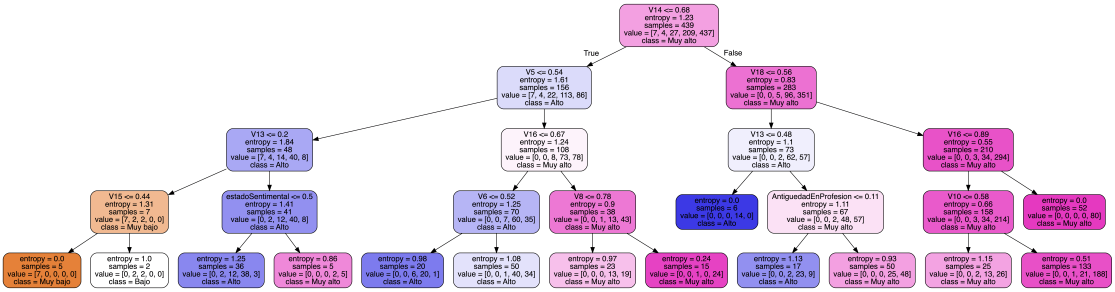

Figura 8.15: Ejemplo de Decision Tree para la variable V9 (1).

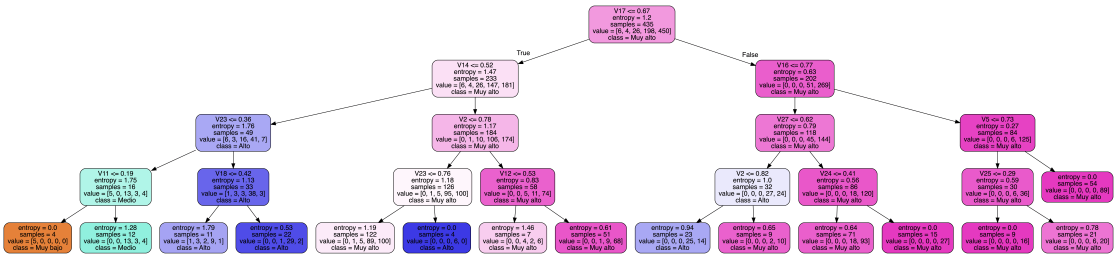

Figura 8.16: Ejemplo de Decision Tree para la variable V9 (2).

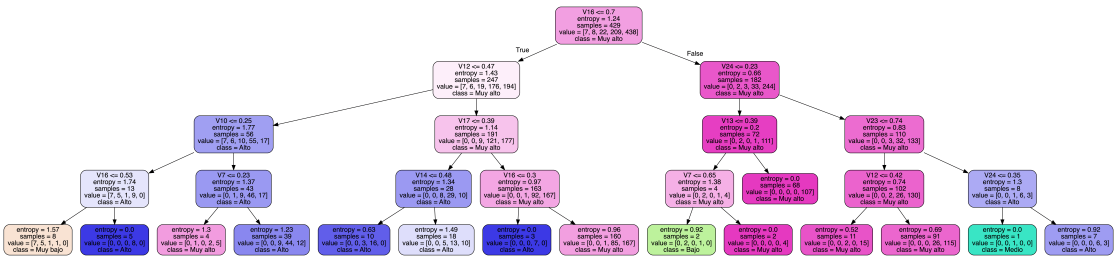

Figura 8.17: Ejemplo de Decision Tree para la variable V9 (3).

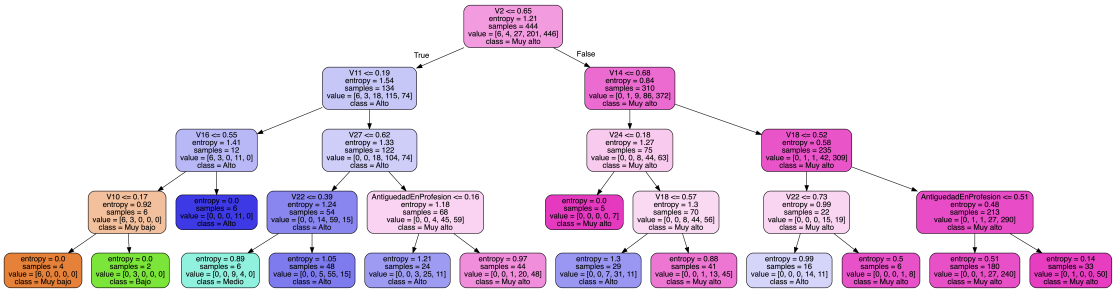

Figura 8.18: Ejemplo de Decision Tree para la variable V9 (4).

Como se puede observar en los árboles mostrados y los analizados 8.15, 8.16, 8.17 y 8.18. Para esta variable si hay una distinción más grande entre el problema con todas las variables o solo las psicológicas. Para todas las variables, de las variables que más se repiten para las grandes separaciones son V18 que es la inteligencia emocional total. Junto con V6 y V5, la responsabilidad y estabilidad emocional. Mientras que para hacer las distinciones más pequeñas y sobre todo entre la clase baja y muy baja, se suele usar la variable V8 .Apertura a la experiencia". Junto con variables sociales como la formación o la antigüedad en el cargo. También usando variables del tipo Calidad Personal y Organizacional.

Mientras que para el análisis del problema utilizando únicamente las variables psicológicas, las variables que nos aportan una mayor información a la hora de hacer grandes separaciones suelen ser V14 (Satisfacción con la vida) y las variables dedicadas a la inteligencia emocional o el engagement, las que se asocian a su percepción. Mientras que para las separaciones más pequeñas, el algoritmo sigue usando las variables dedicadas a la Calidad Personal y Organizacional y sobretodo V2 (Autoeficacia).

V21 (Energía total)

Debido a que la variable V21 resume a las variables V19 y V20, al ser la suma de ambas, mostraremos los árboles únicamente de esta variable y hablaremos sobre ella. A continuación mostraremos algunos ejemplos de los obtenidos por nuestro modelo Random Forest para la variable V21. Más de estos árboles han sido analizados para llegar a estas conclusiones:

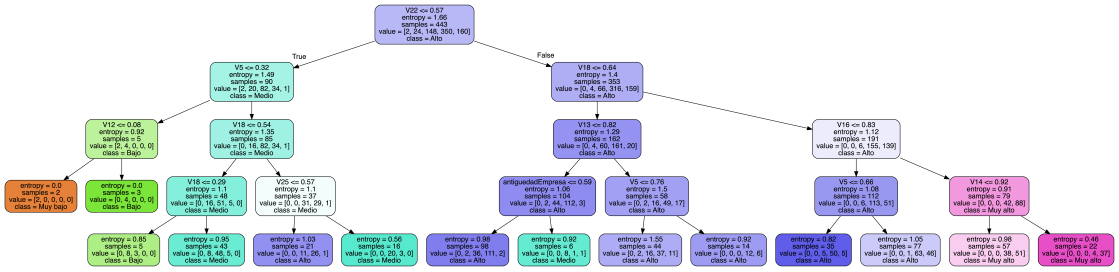

Figura 8.19: Ejemplo de Decision Tree para la variable V21 (1).

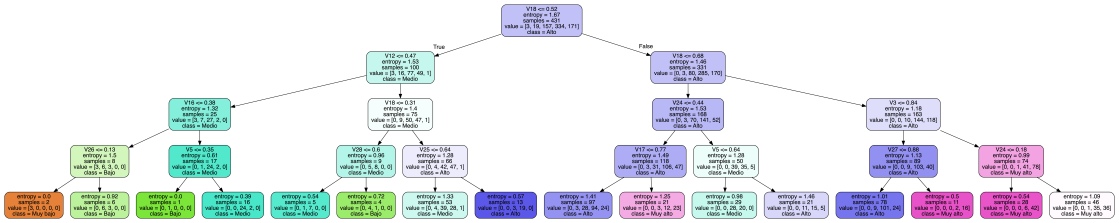

Figura 8.20: Ejemplo de Decision Tree para la variable V21 (2).

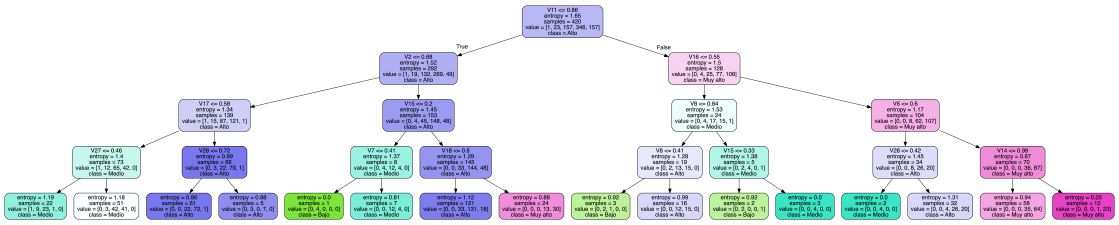

Figura 8.21: Ejemplo de Decision Tree para la variable V21 (3).

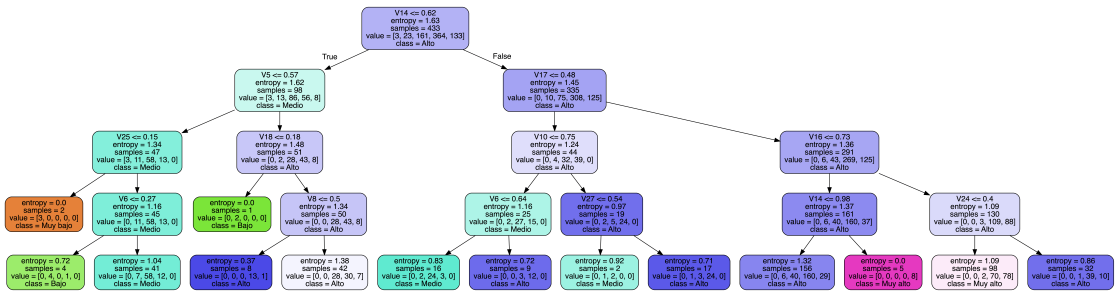

Figura 8.22: Ejemplo de Decision Tree para la variable V21 (4).

Como se puede observar en los árboles mostrados en 8.19, 8.20, 8.21 y 8.22. En este caso existen dos variables que deciden en casi todos los árboles

el primer corte, estas son V18 (Inteligencia emocional total) y V22 (Vitalidad emocional). Posteriormente para los cortes más pequeños se suele ver a la variable V5 (Estabilidad emocional) en casi todos los árboles. Observando los árboles con variables sociales también se ha visto como la formación y la antigüedad en la empresa tiene algún corte, encontrándonos incluso un corte realizado por la variable Género. Los demás estarán realizados como ya vimos en las variables anteriores por las variables pertenecientes a la Calidad Personal y Organizacional del individuo. Aunque en esta si se puede ver una mayor relación con las variables relacionadas con la inteligencia emocional, siendo V18 la más importante para la predicción de esta variables. Además, estos árboles si suelen ser más secuenciales en cuanto al orden de las clases se refiere.

Conclusiones para todas las variables

Lo que si se puede ver en todos los árboles estudiados es como se entremezclan las clases entre sí, puesto que no hay un orden en las ramas de género ascendente como deberían de ser las clases, si no que una variable podría separar las clases "Bajoz .Alto.aunque estas no sean consecutivas. Encontrándonos separaciones con las instancias de la clase "Muy alto.en lo que considerábamos la rama de la clase "Muy bajo".

Capítulo 9

Conclusiones y Trabajos Futuros

Como conclusión final de este trabajo, podemos hablar del gran análisis de datos realizado sobre estas encuestas de carácter psicológico. Este ha sido mi primer enfrentamiento con datos del mundo real, el cual ha ido cobrando sentido y sobre el que he podido ir aprendiendo conforme lo analizaba. Viendo como comentarios de los compañeros de la Universidad de Rovira y Virgili cobran sentido con el análisis. Debido a la clara conexión entre los resultados obtenidos por los individuos en cuanto a inteligencia emocional con la inteligencia a demostrar, que será la inteligencia energética. Otra aporte curiosa que hemos aprendido con este problema es que las variables sociales poco tienen que ver con los resultados obtenidos en cada una de las inteligencias estudiadas, cosa que antes de comenzar con este problema creía (en mi humilde opinión) tendría gran importancia. Se ha visto como estas variables serán bastante difíciles de predecir para futuros problemas y se deberá de dedicar un tiempo mayor al estudio de un modelo que pueda predecir con gran capacidad este problema, puesto que aquí se ha hecho un estudio de las mismas con varios modelos para comprobar su efectividad.

Como futuro trabajo o trabajo pendiente, me hubiera gustado probar otras técnicas para intentar mejorar los resultados de nuestros modelos. Como la utilización de un selector de características y técnicas de *undersampling* u *oversampling*. Así como la realización de unos intervalos más reales acorde al problema con la ayuda de nuestros compañeros en la Universidad de Rovira y Virgili.

Aunque los resultados obtenidos no son totalmente satisfactorios para un científico de datos y viendo los trabajos realizados en el máster, en los que obteníamos una gran precisión en ellos, quiero resaltar que lo que para nosotros pueden ser unos resultados no especialmente brillantes para expertos

de la psicología son verdaderamente significativos. Además del gran conocimiento ganado en la realización de este trabajo, sobretodo en cuanto a la estructuración de un problema real, así como en la visualización de datos.

Bibliografía

[1] S Aranganayagi y K Thangavel. "Clustering categorical data using silhouette coefficient as a relocating measure". En: *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2007)*. Vol. 2. IEEE. 2007, págs. 13-17.

[2] Albert Bandura. "Self-efficacy". En: *The Corsini encyclopedia of psychology* (2010), págs. 1-3.

[3] Nikhil Bansal, Avrim Blum y Shuchi Chawla. "Correlation clustering". En: *Machine learning* 56.1-3 (2004), págs. 89-113.

[4] Derya Birant y Alp Kut. "ST-DBSCAN: An algorithm for clustering spatial–temporal data". En: *Data & Knowledge Engineering* 60.1 (2007), págs. 208-221.

[5] Mihaly Csikszentmihalyi, Sami Abuhamdeh y Jeanne Nakamura. "Flow". En: *Flow and the foundations of positive psychology*. Springer, 2014, págs. 227-238.

[6] Edward Diener. *The science of well-being: The collected works of Ed Diener*. Vol. 37. Springer, 2009.

[7] ED Diener y col. "The satisfaction with life scale". En: *Journal of personality assessment* 49.1 (1985), págs. 71-75.

[8] Pablo Fernández-Berrocal, Natalio Extremera y Natalia Ramos. "Validity and reliability of the Spanish modified version of the Trait Meta-Mood Scale". En: *Psychological reports* 94.3 (2004), págs. 751-755.

[9] Matt W Gardner y S R Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". En: *Atmospheric environment* 32.14-15 (1998), págs. 2627-2636.

[10] Cyril Goutte y Eric Gaussier. "A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation". En: 2005, págs. 345-359. DOI: `10.1007/978-3-540-31865-1_25`. URL: `http://link.springer.com/10.1007/978-3-540-31865-1%7B%5C_%7D25`.

[11] John A Hartigan y Manchek A Wong. "Algorithm AS 136: A k-means clustering algorithm". En: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), págs. 100-108.

[12] Trevor Hastie y Robert Tibshirani. "Discriminant adaptive nearest neighbor classification and regression". En: *Advances in Neural Information Processing Systems.* 1996, págs. 409-415.

[13] Stephen C Johnson. "Hierarchical clustering schemes". En: *Psychometrika* 32.3 (1967), págs. 241-254.

[14] Ian Jolliffe. *Principal component analysis.* Springer, 2011.

[15] Todd B Kashdan, Robert Biswas-Diener y Laura A King. "Reconsidering happiness: The costs of distinguishing between hedonics and eudaimonia". En: *The Journal of Positive Psychology* 3.4 (2008), págs. 219-233.

[16] Ron Kohavi y col. "A study of cross-validation and bootstrap for accuracy estimation and model selection". En: *Ijcai.* Vol. 14. 2. Montreal, Canada. 1995, págs. 1137-1145.

[17] Randy J Larsen y Ed Diener. "A multitrait-multimethod examination of affect structure: Hedonic level and emotional intensity". En: *Personality and Individual Differences* 6.5 (1985), págs. 631-636.

[18] Andy Liaw, Matthew Wiener y col. "Classification and regression by randomForest". En: *R news* 2.3 (2002), págs. 18-22.

[19] John D Mayer y col. "Emotional development and emotional intelligence: Implications for educators". En: *What is emotional intelligence* (1997), págs. 3-31.

[20] Nico J D Nagelkerke y col. "A note on a general definition of the coefficient of determination". En: *Biometrika* 78.3 (1991), págs. 691-692.

[21] Mohammad Norouzi, David J Fleet y Ruslan R Salakhutdinov. "Hamming Distance Metric Learning". En: *Advances in Neural Information Processing Systems 25.* Ed. por F Pereira y col. Curran Associates, Inc., 2012, págs. 1061-1069. URL: `http://papers.nips.cc/paper/4808-hamming-distance-metric-learning.pdf`.

[22] Kevin E O'Grady. "Measures of explained variance: Cautions and limitations." En: *Psychological Bulletin* 92.3 (1982), pág. 766.

[23] William Pavot y col. "Further validation of the Satisfaction with Life Scale: Evidence for the cross-method convergence of well-being measures". En: *Journal of personality assessment* 57.1 (1991), págs. 149-161.

[24] Fabian Pedregosa y col. "Scikit-learn: Machine learning in Python". En: *Journal of machine learning research* 12.Oct (2011), págs. 2825-2830.

[25] Morris Rosenberg. "Rosenberg self-esteem scale (SES)". En: *Society and the adolescent self-image* (1965).

[26] S Rasoul Safavian y David Landgrebe. "A survey of decision tree classifier methodology". En: *IEEE transactions on systems, man, and cybernetics* 21.3 (1991), págs. 660-674.

[27] P Sanjuán, AM Pérez y J Bermúdez. *Escala de Autoeficacia General: Datos psicométricos de la adaptación para población Española [The General Self-efficacy Scale: Psychometric data from the Spanish adaptation]. Psicothema, 12 (supl. 2), 509-513.* 2000.

[28] Wilmar B Schaufeli, Arnold B Bakker y Marisa Salanova. "The measurement of work engagement with a short questionnaire: A cross-national study". En: *Educational and psychological measurement* 66.4 (2006), págs. 701-716.

[29] Wilmar B Schaufeli y col. "The measurement of engagement and burnout: A two sample confirmatory factor analytic approach". En: *Journal of Happiness studies* 3.1 (2002), págs. 71-92.

[30] Ralf Schwarzer y Judith Baessler. "Evaluación de la autoeficacia: Adaptación española de la escala de Autoeficacia General". En: *Ansiedad y estrés* 2.1 (1996), págs. 1-8.

[31] George A F Seber y Alan J Lee. *Linear regression analysis.* Vol. 329. John Wiley & Sons, 2012.

[32] Marina Sokolova, Nathalie Japkowicz y Stan Szpakowicz. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation". En: 2006, págs. 1015-1021. DOI: `10.1007/11941439_114`. URL: `http://link.springer.com/10.1007/11941439%7B%5C_%7D114`.

[33] Johan A K Suykens y Joos Vandewalle. "Least squares support vector machine classifiers". En: *Neural processing letters* 9.3 (1999), págs. 293-300.

[34] Robert J Vallerand y col. "Construction et validation de l'échelle de motivation en éducation (EME)." En: *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 21.3 (1989), pág. 323.

[35] Aki Vehtari, Andrew Gelman y Jonah Gabry. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". En: *Statistics and computing* 27.5 (2017), págs. 1413-1432.

[36] Andreu Vigil-Colet y col. "Development and validation of the overall personality assessment scale (OPERAS)". En: *Psicothema* 25.1 (2013), págs. 100-106.

[37] Zhou Wang y Alan C Bovik. "Mean squared error: Love it or leave it? A new look at signal fidelity measures". En: *IEEE signal processing magazine* 26.1 (2009), págs. 98-117.

[38] Barry J Zimmerman, Albert Bandura y Manuel Martinez-Pons. "Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting". En: *American educational research journal* 29.3 (1992), págs. 663-676.

Apéndice A

Resultados obtenidos con todas las variables para regresión

A.1. V1 (Flow-Fluir)

Para la predicción de la variable V1 que corresponde con la variable psicológica de Flow-Fluir, se usaron las variables de entrada mostradas en la tabla que anteriormente describimos en la Sección 6. Los resultados tanto para la fase de entrenamiento como para la fase de Test de cada uno de los modelos serán mostrados a continuación, así como la mejor configuración de los parámetros encontrada para este problema.

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,010828	0,422080	0,425670	0,009654	0,532053	0,535191	None
Nearest Neighbor Regression	0,009693	0,444362	0,451446	0,011597	0,437921	0,441383	n_neighbors: 3
SVM Regressor	0,007273	0,587107	0,590626	0,009766	0,526660	0,530612	Kernel: linear, C:0,1
Decision Tree Regressor	0,015766	0,109240	0,118324	0,020542	0,004348	0,028075	Criterion: mae
Random Forest Regressor	0,007459	0,577947	0,581862	**0,008174**	**0,603825**	**0,607812**	Criterion: friedman_mse, n_estimators: 500
MLP Regressor	0,007013	0,602034	0,605707	0,009755	0,527171	0,530223	Hidden_layers: 1x20, Activation: logistic, solver: lifts

Cuadro A.1: Resultados de regresión para V1.

Para esta variable podemos observar la tabla A.1 en la que el mejor modelo en la fase de test ha sido **Random Forest**, estando este muy seguido de la red neuronal tanto en entrenamiento como en test. Otra cosa curio-

sa a destacar es el bajo error en MSE que propone el modelo de regresión lineal, además de ser este el modelo más simple de los aplicados. Para los parámetros de la red neuronal, nuestra búsqueda en rejilla ha escogido una red con una sola capa y 20 neuronas de tipo Logística como activación. Puede observarse la gran diferencia existente en R2 y EVC para los modelos de Random Forest y MLP, siendo este el segundo. Y no solamente con dicho modelo, sino con todos los demás modelos comparados. Siendo esta diferencia de aproximadamente 8 centésimas. Además siendo Random Forest uno de los modelos con menos coste computacional debido a la simplicidad de su clasificador base, un árbol de decisión. También cabe destacar la dificultad del problema que nos encontramos, puesto que apenas hemos podido mejorar los resultados obtenidos por el regresor lineal.

A.2. V9 (Flourishing-Florecimiento)

Ahora mostraremos los resultados obtenidos por nuestros modelos para esta variable V9, representando el florecimiento. Mostraremos primero los resultados obtenidos para el proceso de entrenamiento y posteriormente para test.

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,008749	0,511276	0,512950	0,011881	0,345077	0,348708	None
Nearest Neighbor Regression	0,010296	0,435053	0,440425	0,012785	0,295255	0,309145	n_neighbors: 11
SVM Regressor	0,008067	0,543154	0,543653	0,012172	0,329078	0,333216	Kernel: rbf, C: 10
Decision Tree Regressor	0,015149	0,148923	0,151669	0,017578	0,031085	0,032180	Criterion: mae
Random Forest Regressor	0,008445	0,527409	0,530428	**0,010870**	**0,400813**	**0,400814**	Criterion: friedman_mse, n_estimators: 500
MLP Regressor	0,007889	0,559888	0,561780	0,012889	0,289545	0,293494	Hidden_layers: 1x5, Activation: Logistic, Solver: lbfgs

Cuadro A.2: Resultados de regresión para V9.

Como podemos ver en la tabla A.2, para esta variable se repite lo visto en la variable anterior. Todos los modelos obtienen un error muy bajo, incluso el modelo de regresión linear. Sin embargo, el mejor modelo, observando la fase de test ha sido otra vez **Random Forest**. Esta vez la mejor configuración para nuestra red neuronal ha sido de una capa oculta con 5 neuronas en ella, siendo las neuronas de tipo logística. Aunque los errores sean bajos, distan mucho del mínimo que nos gustaría alcanzar con nuestros modelos. Para esta variable el error ha sido bastante más alto que el cometido en la predicción de la variable V1 (Flow-Fluir).

A.3. V19 (Self Drive Direction)

Esta es la primera variable de tipo energético a analizar V19 (Self Drive Direction), estas variables tienen un mayor rango y dispersión de los datos que las variables a predecir anteriormente vistas, por lo que los resultados se verán afectados por este hecho. Siendo los errores más altos para todos los modelos que en las variables anteriores. A continuación, se muestran los resultados obtenidos por nuestros modelos tanto para la fase de entrenamiento como la de Test.

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,018320	0,416992	0,421799	0,017857	0,351529	0,351612	None
Nearest Neighbor Regression	0,020562	0,348017	0,356447	0,019826	0,280012	0,280414	n_neighbors: 16
SVM Regressor	0,017824	0,433942	0,439247	0,018743	0,319318	0,319340	Kernel: rbf, C: 1
Decision Tree Regressor	0,035764	-0,159866	-0,151989	0,028564	-0,037330	-0,032171	Criterion: friedman_mse
Random Forest Regressor	0,018441	0,414920	0,419673	**0,016461**	**0,402210**	**0,407383**	Criterion: mae, n_estimators: 100
MLP Regressor	0,018038	0,426476	0,431405	0,017676	0,358092	0,358392	Hidden_layers: 3x10, Activation: Identity, Solver: lbfgs

Cuadro A.3: Resultados de regresión para V19.

Para esta variable, también vemos en la tabla A.3 como sigue siendo **Random Forest** el mejor modelo, sin embargo, esta vez el modelo SVM también ha obtenido una buena puntuación. Lo cual nos indica que para esta variable, las instancias serán más separables entre ellas. En esta variable el modelo de Regresión lineal llega incluso a superar a otros modelos para las métricas de EVC y R2, lo cual nos indica que debemos de mejorar mucho más nuestros modelos debido a los bajos resultados obtenidos por todos ellos.

A.4. V20 (Consciencia energética)

Esta variable también es de inteligencia energética, concretamente representando la Consciencia energética. Y como vimos en las anteriores secciones sobre las relaciones entre variables, los resultados obtenidos para esta variable debería de ser parecidos a los obtenidos para la variable anterior, mostrados a continuación:

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,008156	0,617687	0,618386	0,010882	0,482062	0,483924	None
Nearest Neighbor Regression	0,010359	0,517785	0,522341	0,012853	0,388265	0,388745	n_neighbors: 16
SVM Regressor	0,008065	0,623019	0,623255	**0,010619**	**0,494594**	**0,496678**	Kernel: rbf, C: 1
Decision Tree Regressor	0,015868	0,256354	0,263399	0,018504	0,119283	0,121339	Criterion: mae
Random Forest Regressor	0,007831	0,633803	0,634541	0,010677	0,491794	0,495893	Criterion: mae, n_estimators: 500
MLP Regressor	0,008073	0,622317	0,622970	0,010673	0,492000	0,494429	Hidden_layers: 5x2, Activation: Identity, Solver: lbfgs

Cuadro A.4: Resultados de regresión para V20.

Para esta variable si tenemos un cambio, como podemos observar en la tabla A.4, y es que el modelo **SVM** ha obtenido los mejores resultados en la predicción de esta variable. Aunque los modelos SVM, Random Forest y MLP tienen unas puntuaciones casi idénticas, finalmente ha sido SVM el que ha obtenido una mejor predicción en el conjunto de test. Lo que puede venir a indicar la existencia de barreras en nuestros planos que podrán hacer de estos una separación más cómoda en lo que se refiere a esta variable. La configuración escogida esta vez para esta Red Neuronal ha sido de cinco capas ocultas con dos neuronas en cada una de ellas, siendo estas de tipo identidad. Mientras que SVM ha escogido un Kernel rbf y una C de 1.

A.5. V21 (Energía total)

Esta es la última de las variables pertenecientes a la inteligencia energética, representando el total de inteligencia energética que posee un individuo. Por lo tanto, los resultados serán bastante parecidos a los obtenidos en las dos variables anteriores. Puesto que esta es la suma de ambas, seguirá la misma distribución. Los resultados obtenidos por nuestros modelos son los siguientes:

	Train			Test			
	MSE ↓	**R2 ↑**	**EVC ↑**	**MSE ↓**	**R2 ↑**	**EVC ↑**	**Params**
Linear Regression	0,008478	0,604443	0,608108	0,010314	0,492449	0,493395	None
Nearest Neighbor Regression	0,010809	0,504114	0,510620	0,012304	0,394555	0,394796	n_neighbors: 10
SVM Regressor	0,008460	0,607358	0,609927	0,010368	0,489823	0,491073	Kernel: rbf, C: 1
Decision Tree Regressor	0,017556	0,173524	0,176278	0,019163	0,057026	0,065202	Criterion: mae
Random Forest Regressor	0,008421	0,609131	0,611884	**0,009711**	**0,522119**	**0,525499**	Criterion: mae, n_estimators: 500
MLP Regressor	0,008334	0,611917	0,614761	0,009982	0,508823	0,510500	Hidden_layers: 5x2, Activation: Identity, Solver: lbfgs

Cuadro A.5: Resultados de regresión para V21.

Como puede observarse en la tabla A.5, volvemos a ver a **Random Forest**, sin tener que usar totalmente su capacidad, puesto que usa solamente 100 de los 1000 árboles propuestos. Estando esta vez los resultados bastante reñidos en cuanto a MSE, sin embargo si miramos las métricas de EVC y R2, puede verse una clara diferencia entre los modelos.

Apéndice B

Resultados obtenidos con todas las variables en clasificación

B.1. V1 (Flow-Fluir)

Los resultados obtenidos para esta variable son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,652136	0,437440	0,429529	0,347864	0,552326	0,333974	0,278764	0,447674	kernel: sigmoid, C: 10
kNN	0,173124	0,092219	0,143155	0,076796	0,540698	0,324895	0,522222	0,459302	k: 19
Decision Tree	0,145600	0,126297	0,135349	0,104000	0,581395	0,482451	0,688351	0,418605	Criterion: entropy
Random Forest	0,687248	0,514833	0,636826	0,312752	**0,622093**	**0,442482**	**0,502137**	**0,377907**	Criterion: entropy, n_estimators: 100
MLP	0,638826	0,412333	0,438038	0,361174	0,575581	0,409638	0,421797	0,424419	Hidden_layers: 5x100, Activation: relu, Solver: Adam

Cuadro B.1: Resultados multiclase para la variable V1.

Como se puede observar en la tabla B.1, para la variable V1, las mejores puntuaciones han sido obtenidas por **Random Forest**. Hemos seleccionado al mejor modelo conforme al *Accuracy*, sin embargo, puede observarse que otros modelos obtienen una mejor puntuación en otras métricas tales como *Precision* o *Recall*. Puede verse como otros modelos más simples como el Decision Tree, ha obtenido unos resultados mejores para las métricas citadas. Otra observación será la arquitectura tan compleja que ha escogido la red neuronal para solucionar este problema, siendo esta de 5 capas ocultas con 100 neuronas en cada una de ellas.

B.2. V9 (Flourishing-Florecimiento)

Los resultados obtenidos para esta variable son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,730990	0,421443	0,424530	0,269010	0,656977	0,253247	0,240391	0,343023	kernel: linear, C: 1
kNN	0,185577	0,118243	0,130771	0,064023	0,686047	0,350902	0,564456	0,313953	k: 9
Decision Tree	0,178079	0,119043	0,118117	0,071521	0,656977	0,306542	0,312097	0,343023	Criterion: gini
Random Forest	0,747016	0,441069	0,498010	0,252984	**0,697674**	**0,290652**	**0,458667**	**0,302326**	Criterion: entropy, n_estimators: 100
MLP	0,711861	0,308716	0,306105	0,288139	0,662791	0,318852	0,302105	0,337209	Hidden_layers: 5x100, Activation: tanh, Solver: Adam

Cuadro B.2: Resultados multiclase para la variable V9.

Para esta variable V9, como podemos observar en la tabla B.2, también obtenemos como mejor modelo en cuanto a *Accuracy* a **Random Forest**. En esta ocasión también existen modelos que han obtenido mejores resultados tanto para *Recall* como para *Precision*. Además es observable la diferencia que existe entre los resultados de entrenamiento y los de Test. Observando la pobre generalización que tiene nuestro modelo en cuanto a los resultados de Test se refiere.

B.3. V19 (Self Drive Direction)

Los resultados obtenidos para esta variable son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,527780	0,323758	0,378810	0,472220	0,505814	0,283708	0,343750	0,494186	kernel: linear, C: 0.1
kNN	0,122126	0,099536	0,142592	0,127474	0,523256	0,337007	0,386309	0,476744	k: 10
Decision Tree	0,107971	0,100827	0,096464	0,140029	0,447674	0,313040	0,299101	0,552326	Criterion: entropy
Random Forest	0,552587	0,415135	0,558234	0,447413	**0,587209**	**0,346033**	**0,384470**	**0,412791**	Criterion: entropy, n_estimators: 500
MLP	0,501503	0,341982	0,367260	0,498497	0,441860	0,287289	0,448500	0,558140	Hidden_layers: 5x100, Activation: tanh, Solver: Adam

Cuadro B.3: Resultados multiclase para la variable V19.

En esta ocasión, para la variable V19, se ha obtenido como mejor modelo en la fase de generalización a **Random forest**, como puede observarse en la tabla B.3. Siendo los resultados en la fase de generalización mucho mejores que los obtenidos en la fase de entrenamiento. Aunque ambos resultados son bastantes pobres, puesto que ningún modelo ha superado el 60 % de *Accuracy* para esta variable. Lo que nos viene a indicar lo visto en el apartado de regresión. Una difícil tarea para clasificar esta variable.

B.4. V20 (Consciencia energética)

Los resultados obtenidos para esta variable son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,655013	0,587560	0,618092	0,344987	0,598837	0,431832	0,443420	0,401163	kernel: linear, C: 10
kNN	0,163620	0,105658	0,130609	0,085980	0,569767	0,372779	0,464344	0,430233	k: 16
Decision Tree	0,138320	0,100538	0,107012	0,111600	0,558140	0,424539	0,409024	0,441860	Criterion: gini
Random Forest	0,675472	0,523376	0,655974	0,324528	**0,662791**	**0,477173**	**0,502594**	**0,337209**	Criterion: entropy, n_estimators: 100
MLP	0,597950	0,472869	0,533313	0,402050	0,610465	0,521212	0,702685	0,389535	Hidden_layers: 5x100, Activation: tanh, Solver: Adam

Cuadro B.4: Resultados multiclase para la variable V20.

Para esta variable V20, como puede observarse en la tabla B.4, hemos obtenido el mismo mejor modelo tanto para la fase de entrenamiento como para la fase de Test, siendo este **Random Forest**. En esta ocasión la diferencia de rendimiento entre la fase de entrenamiento y la fase de Test es apenas visible en *Accuracy*, aunque si es notable en cuanto a *Precision* y *Recall*. Sin embargo, en esta ocasión **Random Forest** también ha obtenido los mejores resultados para estas métricas en la fase de Test.

B.5. V21 (Energía total)

Los resultados obtenidos para esta variable son los siguientes:

	Train				Test				
	Accuracy ↑	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Accuracy ↑**	**Recall ↑**	**Precision ↑**	**Hamming Loss ↓**	**Params**
SVM	0,630131	0,541914	0,586334	0,369869	0,651163	0,477664	0,497655	0,348837	kernel: linear, C: 10
kNN	0,156730	0,103381	0,115139	0,091270	0,569767	0,363127	0,442020	0,430233	k: 13
Decision Tree	0,145637	0,106548	0,106326	0,104283	0,552326	0,393313	0,395651	0,447674	Criterion: gini
Random Forest	0,662323	0,543416	0,655181	0,337677	**0,715116**	**0,581197**	**0,792815**	**0,284884**	Criterion: gini, n_estimators: 50
MLP	0,600848	0,388243	0,478861	0,399152	0,598837	0,440044	0,449196	0,401163	Hidden_layers: 5x100, Activation: tanh, Solver: Adam

Cuadro B.5: Resultados multiclase para la variable V21.

En esta última variable que representa el nivel de energía total del individuo, se han obtenido los mejores resultados otra vez con el algoritmo de **Random Forest**, como se puede observar en la tabla B.5. Siendo en esta ocasión los resultados mejores en la fase de generalización que en la de entrenamiento. Obteniendo además muy buenos resultados para la métrica de *Precision* y *Recall* por nuestro mejor modelo. Superando el 70 % en *Accuracy*.