Universitat Oberta de Catalunya

uoc.edu

ESTUDIO DE PREDICTORES DE FELICIDAD A NIVEL MUNDIAL

María Augusta Jimbo Granda

Máster universitario en Ciencia de Datos (Data Science)

Ciencia de Datos Aplicada a Salud

Director: José Luis Iglesias Allones

Co-Director: Liliana Elvira Enciso Quispe

Marzo de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio de predictores de felicidad a nivel mundial*
Nombre del autor:	*María Jimbo Granda*
Nombre del consultor/a:	*José Iglesias Allones* *Liliana Enciso Quispe*
Nombre del PRA:	*Àngels Ruis Gavidia*
Fecha de entrega (mm/aaaa):	*06/2019*
Titulación:	*Máster universitario en Ciencia de Datos (Data Science)*
Área del Trabajo Final:	*Ciencia de Datos aplicada a Salud*
Idioma del trabajo:	*Español*
Palabras clave:	*felicidad, factor, minería de datos*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El objetivo principal de desarrollar este proyecto es el de descubrir factores clave que hacen a la gente más feliz, para lo cual se ha diseñado y construido un modelo de minería de datos.

Los datos que se consideran en el desarrollo de este proyecto han sido obtenidos de encuestas históricas correspondientes a los años 2015, 2016 y 2017 en más de 100 países a nivel mundial. Los seis principales factores que constan en cada uno de los datasets y que sirven para evaluar la felicidad en cada país son: producción económica, apoyo social, esperanza de vida, libertad, ausencia de corrupción y generosidad.

Para el análisis de los datos del informe se utiliza técnicas de machine learning y minería de datos, la programación se ha desarrollado en el lenguaje R.

Según los resultados obtenidos, se concluye o da respuesta a las siguientes

preguntas clave:

- ✓ ¿Cuáles son los principales factores que contribuyen a la felicidad?
- ✓ ¿Existen diferencias importantes en dichos factores entre países?
- ✓ ¿Existen diferencias de felicidad en los tres años?
- ✓ ¿Existen relaciones entre las distintas regiones según el nivel de felicidad?
- ✓ ¿En qué región se encuentran los países más felices y menos felices del mundo?.

Abstract (in English, 250 words or less):

The main objective of developing this project is to discover key factors that make people happier, for which a data mining model has been designed and built.

The data considered in the development of this project have been obtained from historical surveys corresponding to the years 2015, 2016 and 2017 in more than 100 countries worldwide. The six main factors that appear in each of the datasets and that serve to assess happiness in each country are: economic production, social support, life expectancy, freedom, absence of corruption and generosity.

For the analysis of the data of the report, machine learning and data mining techniques are used, the programming has been developed in the R language.

According to the results obtained, the following key questions are concluded or answered:

- ✓ What are the main factors that contribute to happiness?
- ✓ Are there important differences in these factors between countries?
- ✓ Are there differences in happiness in the three years?
- ✓ Are there relations between the different regions according to the level of happiness?
- ✓ In which region are the happiest and least happy countries in the world?.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

DEDICATORIA

El presente trabajo lo dedico a Dios por ser siempre mi guía.

A mis hijos y esposo por ser fuente principal de inspiración.

A mis padres y hermanos por su apoyo incondicional en cada momento de mi vida.

CAPÍTULO 1. INTRODUCCIÓN

1.1 CONTEXTO Y JUSTIFICACIÓN DEL PROYECTO

La felicidad es un sentimiento que nos acerca a estados de bienestar. Es un conjunto de emociones que, si se mantienen en el tiempo, producen cambios en el cuerpo y en la mente, por lo tanto, se asocia con vivir plenamente y en buen estado de salud.

El informe de la felicidad mundial es una encuesta histórica del estado de la felicidad global. El primer informe se publicó en 2012, el segundo en 2013, el tercero en 2015 y el cuarto en la actualización de 2016. El informe de la felicidad del mundo 2017, que clasifica a 155 países según su nivel de felicidad, se lanzó en las Naciones Unidas en un evento que celebra el Día Internacional de la Felicidad el 20 de marzo. Es decir, existen 3 datasets con los cuales se realiza el trabajo y corresponden a los años 2015, 2016 y 2017; estos se encuentran publicados en el sitio web de kaggle [1]. En el proyecto no se utilizó data del año 2018, pues existen pocos registros de países (136), que no se pueden comparar con la cantidad de registros existentes en los años 2015, 2016 y 2017.

En los datasets se ha recopilado información de más de 150 países de todo el mundo, en que según los factores claves como son: producto interno bruto (PIB), ayuda social, esperanza de vida, libertad en el entorno social, generosidad y la no presencia o ausencia de corrupción [2], se los pueda catalogar como los más felices y los menos felices.

Un aspecto importante a tomar en cuenta es que si los migrantes que se encuentran en determinado país son felices, también los son y mucho más aquellos individuos originarios de ese país.

Un modelo de minería de datos "*es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones*" [3].

En el presente proyecto se desarrollará un modelo de minería de datos en el que se pueda analizar y por tanto determinar los principales factores que intervienen en la

felicidad; la idea también es que la implementación de este modelo puede ser usada en un futuro, para otros casos de estudio de predicciones.

1.1.1 MOTIVACIÓN

Mi motivación personal por desarrollar el presente trabajo es porque de lo que he aprendido hasta ahora en el máster me agrada más el estudio de las predicciones en las que se trabaja principalmente con contenidos de las materias de Estadística, lenguaje R y Visualización de datos.

Seleccioné el tema de felicidad porque necesito determinar la importancia de los factores clave que intervienen en este sentimiento de autorrealización y cumplimiento de nuestros deseos y aspiraciones. Al igual que el resto de personas me encanta sentirme feliz porque si yo lo estoy, puedo lograr que el resto también lo sea, es decir; me agrada saber que puedo aportar con un granito de arena para que al menos los que se encuentran a mi alrededor se sientan realizados y felices.

Adicional, porque quiero tener experiencia profesional y aplicar todos los contenidos aprendidos en mi entorno laboral.

1.2 OBJETIVOS DEL PROYECTO

1.2.1 Objetivo Principal

- Diseñar y construir un modelo de minería de datos para predecir la felicidad a nivel mundial.

1.2.2 Objetivos Específicos

- Determinar los principales factores que contribuyen a la felicidad.
- Descubrir diferencias de factores entre países.
- Determinar si existe diferencia de felicidad en los tres años.
- Determinar en qué región se encuentran los países más felices y menos felices del mundo.
- Analizar la evolución que ha existido en esta línea de investigación y su estado actual.
- Desarrollar y evaluar un modelo de minería a aplicar.

1.3 ENFOQUE Y MÉTODO

A continuación, se describe las fases del proceso de minería de datos y el tipo de modelo que se va a utilizar en el desarrollo del proyecto.

1.3.1 Enfoque

En este trabajo se construirá un modelo de minería de datos que sirva para determinar factores clave de felicidad en el mundo.

La minería de datos es la etapa de análisis de "Descubrimiento de conocimiento en bases de datos", adicional consiste en un campo de la estadística y las ciencias de la computación en donde se intenta descubrir patrones de grandes volúmenes de datos.

Un proceso de minería de datos consta de 6 fases que son:

1) **Seleccionar el conjunto de datos**: Se refiere a variables objetivo (las que se quiere predecir, calcular o inferir), a las variables independientes (las que sirven para hacer el cálculo o proceso) y a la muestra de los registros que se encuentren disponibles.
2) **Analizar las propiedades de los datos**: El análisis de las propiedades de los datos mediante el uso principal de histogramas, de diagramas de dispersión, presencia de valores atípicos (outliers) y ausencia de datos (valores nulos).
3) **Transformar el conjunto de datos de entrada:** Es decir, realizar un **preprocesamiento** de los datos, en donde se ejecute la transformación en función del análisis previo, esto con el objetivo de prepararlo para aplicar la técnica de minería que mejor se adapte a los datos.
4) **Seleccionar y aplicar la técnica de minería de datos**: Esta etapa abarca lo que es la construcción del modelo predictivo, de clasificación o de segmentación.
5) **Extraer el conocimiento**: Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, el mismo que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación que pudiesen existir entre dichas variables.
6) **Interpretar y evaluar los datos**: Una vez obtenido el modelo, se lo comprueba y valida, de esta evaluación se debe revisar que las conclusiones que se generan estén acordes al problema, si existen resultados inconsistentes

se debe seguir probando el modelo varias veces (volver a la etapa anterior), ajustando parámetros de entrada hasta obtener el mejor modelo.

El proceso se podría repetir desde el principio o si el experto lo considera oportuno, en el caso de que el modelo final no supere la evaluación. Esta retroalimentación se puede repetir n veces hasta obtener un modelo válido [4].

1.3.2 Método

Una metodología permite llevar a cabo el proceso de minería de datos en forma sistemática y ordenada. Para el desarrollo del proyecto, la metodología a usar es la de modelo en cascada.

El modelo en cascada es aquel que admite iteraciones, aunque sólo de una etapa a su inmediata anterior, por más que se represente como un simple modelo en forma de cascada al igual que un ciclo de vida secuencial como el lineal. Después de cada etapa se realiza una o varias revisiones para comprobar si se puede pasar a la siguiente. La necesidad de tener en claro los requerimientos al inicio del proyecto es primordial al optar por este modelo [5].

En la Ilustración 1, se observa las cinco etapas del modelo en cascada.

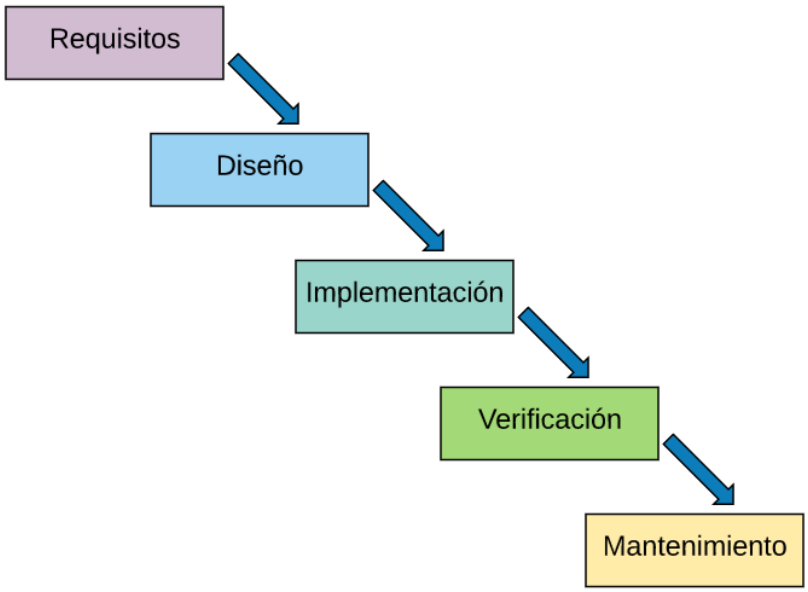


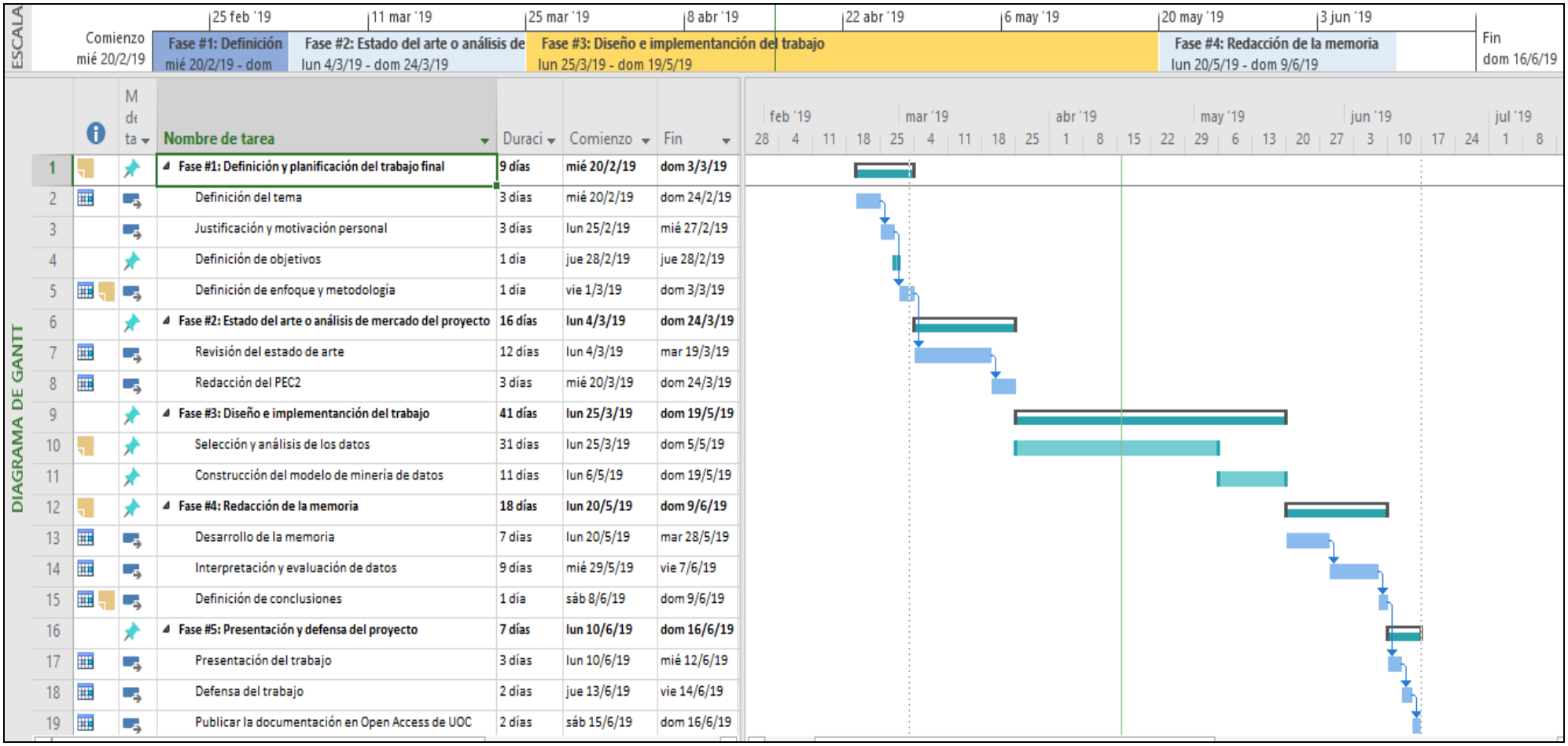
Ilustración 1. Etapas del modelo en cascada [5].

1.4 PLANIFICACIÓN DEL TRABAJO

El trabajo se planifica por fases, a continuación, se presenta un resumen:

Fase	**Fecha máxima de entrega**	**Descripción**
Fase 1	3 de marzo	Definición y planificación del trabajo final
Fase 2	24 de marzo	Estado del arte o análisis de mercado del proyecto
Fase 3	19 de mayo	Diseño e implementación del trabajo
Fase 4	9 de junio	Redacción de la memoria
Fase 5	16 de junio	Presentación y defensa del proyecto

A continuación, se detallan las tareas que incluiría cada una de las fases.


ESCALA
25 feb '19
11 mar '19
25 mar '19
8 abr '19
22 abr '19
6 may '19
20 may '19
3 jun '19
Comienzo mié 20/2/19
Fase #1: Definición mié 20/2/19 - dom
Fase #2: Estado del arte o análisis de lun 4/3/19 - dom 24/3/19
Fase #3: Diseño e implementación del trabajo lun 25/3/19 - dom 19/5/19
Fase #4: Redacción de la memoria lun 20/5/19 - dom 9/6/19
Fin dom 16/6/19
DIAGRAMA DE GANTT
Nombre de tarea
Duraci
Comienzo
Fin
1 Fase #1: Definición y planificación del trabajo final 9 días mié 20/2/19 dom 3/3/19
2 Definición del tema 3 días mié 20/2/19 dom 24/2/19
3 Justificación y motivación personal 3 días lun 25/2/19 mié 27/2/19
4 Definición de objetivos 1 día jue 28/2/19 jue 28/2/19
5 Definición de enfoque y metodología 1 día vie 1/3/19 dom 3/3/19
6 Fase #2: Estado del arte o análisis de mercado del proyecto 16 días lun 4/3/19 dom 24/3/19
7 Revisión del estado de arte 12 días lun 4/3/19 mar 19/3/19
8 Redacción del PEC2 3 días mié 20/3/19 dom 24/3/19
9 Fase #3: Diseño e implementación del trabajo 41 días lun 25/3/19 dom 19/5/19
10 Selección y análisis de los datos 31 días lun 25/3/19 dom 5/5/19
11 Construcción del modelo de minería de datos 11 días lun 6/5/19 dom 19/5/19
12 Fase #4: Redacción de la memoria 18 días lun 20/5/19 dom 9/6/19
13 Desarrollo de la memoria 7 días lun 20/5/19 mar 28/5/19
14 Interpretación y evaluación de datos 9 días mié 29/5/19 vie 7/6/19
15 Definición de conclusiones 1 día sáb 8/6/19 dom 9/6/19
16 Fase #5: Presentación y defensa del proyecto 7 días lun 10/6/19 dom 16/6/19
17 Presentación del trabajo 3 días lun 10/6/19 mié 12/6/19
18 Defensa del trabajo 2 días jue 13/6/19 vie 14/6/19
19 Publicar la documentación en Open Access de UOC 2 días sáb 15/6/19 dom 16/6/19
feb '19
mar '19
abr '19
may '19
jun '19
jul '19
28 4 11 18 25 4 11 18 25 1 8 15 22 29 6 13 20 27 3 10 17 24 1 8

CAPÍTULO 2. ESTADO DEL ARTE O ANÁLISIS DE MERCADO DEL PROYECTO

2.1 FELICIDAD

2.1.1 DEFINICIÓN

Felicidad es un estado afectivo de satisfacción plena que experimenta subjetivamente el individuo en posesión de un bien deseado. A partir de esta definición, se tiene los siguientes indicadores:

a) La felicidad significa sentimientos de satisfacción que vivencia la persona en su vida interior.
b) El hecho de ser un estado de la conducta, alude un carácter temporal de la felicidad, es decir; puede ser duradera, pero a la vez, también es perecible.
c) La felicidad supone la posesión de un bien, es decir, uno es feliz en tanto posee el bien u objeto.
d) El bien o bienes que generan la felicidad son de distinta naturaleza como: materiales, éticos, estéticos, psicológicos, religiosos, sociales, etc.

La felicidad es un estado y a la vez, un proceso dinámico, que es generado por la interacción de un amplio número de condiciones o variables que actúan sobre el individuo provocando respuestas terminales de naturaleza positiva. Estas variables pueden agruparse de diferente manera: Biológicas (género, salud, malformaciones), psicológicas (rasgos de personalidad, autoestima, valores, creencias, afectos) y socioculturales (matrimonio, ingreso económico, familia, marginación, etc.).

2.1.2 DÍA INTERNACIONAL DE LA FELICIDAD

El 20 de marzo se celebra el Día Internacional de la Felicidad.

La felicidad y el bienestar van de la mano por constituirse en objetivos y aspiraciones universales en las vidas de los seres humanos en todo el mundo.

La investigación sobre el bienestar, a veces llamada estudios de la felicidad, se puede encontrar en una amplia gama de campos que incluyen economía, negocios, psicología, sociología, ciencias políticas y educación.

2.1.3 INFORME DE LA FELICIDAD MUNDIAL

Es una encuesta histórica que se ha realizado desde el 2012 sobre el estado de la felicidad global que clasifica a 158 países por lo felices que se sienten sus ciudadanos. El informe es producido por la Red de Soluciones de Desarrollo Sostenible de las Naciones Unidas en asociación con la Fundación Ernesto Illy.

El bienestar de los ciudadanos de los distintos países se mide en base a variables estadísticas y según las opiniones individuales, las cuales son recogidas en el informe. La mayoría de gobiernos utilizan de estos datos e investigaciones para buscar políticas adecuadas que ayuden a sus ciudadanos a vivir mejor.

Existe una consistencia notable entre la felicidad de los inmigrantes y la de los que nacieron en aquel país. Las personas que se mudan a lugares más felices que su país natal "ganan", mientras que aquellas quienes se van a sitios menos felices "pierden" [6].

Según los autores del informe, el caso de América Latina muestra que la abundancia y la naturaleza de las relaciones interpersonales es un importante motor de la felicidad. La gente en la región tiende a generar relaciones interpersonales abundantes, pues son cercanas, cálidas y genuinas.

Una relación de buena calidad significa una relación en la que te sientes seguro, en la que puedes ser tú mismo. Claro que ninguna relación es ideal, pero esas son cualidades que hacen que la gente florezca.

En la Ilustración 2, se aprecia una imagen de niños felices, disfrutando en ese momento en un tobogán de piscina.

Ilustración 2. Niños diviertiéndose en un tobogan [7].

2.1.4 FACTORES QUE DETERMINAN LA FELICIDAD

Un predictor es una variable o factor.

En la investigación se va a analizar la importancia de cada uno de los factores que intervienen en la felicidad, así como también determinar la relación entre estos, identificar, entre las variables, cuál o cuáles son los mejores predictores de la felicidad.

La muestra estuvo integrada por 158 países del mundo entero.

Son fundamentalmente 6 los factores que influyen en la felicidad. A continuación, se detalla cada uno de ellos:

a) **Producto interno bruto per cápita**: Es la suma final de cantidades de bienes y servicios que se producen en un país, al valor monetario de un país de referencia.
b) **Apoyo social**: Es la ayuda de parte de familiares o amigos en caso de tener problemas.
c) **La esperanza de vida**: Es un índice con el que se determina cuánto se espera que viva una persona en un contexto social determinado.

d) **La libertad de tomar decisiones:** Es la facultad que tiene la persona para la toma de decisiones, es decir; que es responsable directamente de sus actos.
e) **La generosidad**: Es la virtud de dar y compartir por sobre el propio interés o utilidad.
f) **La percepción de la corrupción**: Es el concepto que los ciudadanos tienen respecto al gobierno y/o empresas [8].

2.1.4.1 LISTA DE LOS 10 PAÍSES MÁS FELICES DEL MUNDO SEGÚN LA ONU

Considerando los factores que influyen en la determinación de la felicidad, a continuación, se presenta la lista de los 10 países más felices del mundo del año 2017, ordenados según su puntuación.

1. **Noruega:** Es un país seguro y de paisajes pintorescos. Es también un país rico que invierte su ingreso petrolero en el desarrollo del país y en ahorrar a través de fondos para las generaciones futuras.
   La gente busca actividades de ocio menos caras, como hacer ejercicio y disfrutar la naturaleza.
2. **Dinamarca:** Un país bonito, acogedor y con un gran nivel de vida. El sistema de impuestos también sirve para sustentar un excelente sistema educativo desde la primera formación hasta la universidad. Los estudiantes pueden optar a becas y subvenciones mensuales durante siete años para completar su formación superior con calma.
3. **Islandia:** Sus paisajes son increíblemente asombrosos. Es un país muy seguro y cuenta con el desierto más grande en Europa**.**
   Islandia nunca ha tenido ejército ni ha entrado en conflicto. Los islandeses demuestran la importancia de una buena alimentación para una vida saludable**.**
4. **Suiza:** Es un país rico. Cuenta con bellezas naturales, sus residentes son cordiales y las políticas progresivas tienen que ver con la riqueza y cultura propia de un país.
5. **Finlandia:** No solo tiene una ciudad capital llena de creatividad gastronómica, si no también paisajes naturales, lagos y auroras boreales, y es un país que tiene la mejor educación del mundo.
   Es un país muy seguro e igualitario, ahí tanto los finlandeses como los migrantes a ese país son los más felices del mundo.

6. **Países bajos:** La gente realiza ejercicio físico moderado y respeta la ecología. En estos países existe poco desempleo, una desigualdad relativamente baja y una economía saludable.

   Los niños y jóvenes son felices porque tienen en general interacciones positivas en todos sus ambientes sociales, es decir; tienen un ambiente de apoyo en sus casas, entre sus amigos y también en su entorno educativo.
7. **Canadá:** Es un país que tiene un alto índice de desarrollo. Además, se constituye en uno de los países más seguros del mundo pues tiene una fuerte red de seguridad social y un compromiso compartido con valores como el respeto mutuo y un alegre multiculturalismo.
8. **Nueva Zelanda:** País que tiene una de las economías de más rápido crecimiento en el mundo, baja tasa de desempleo, baja inflación y excelentes oportunidades de empleo.

   Nueva Zelanda es conocida por sus paisajes espectaculares, estilo de vida relajado y gente amable.
9. **Suecia:** La gente de ese país respeta mucho la ecología. Suecia satisface, según sus residentes, todos los requisitos en seguridad, acogida a los extranjeros y belleza. En Suecia la gente pasa alegre por su estatus independiente.
10. **Australia:** Es un país querido por sus residentes por las sensaciones de seguridad, protección y paz que brinda. La violencia con armas de fuego es mínima.

    Los australianos se preocupan por el ambiente y bienestar animal. El país tiende a ser bastante económico con un servicio de salud universal de gran calidad y financiamiento gubernamental para la educación superior.

2.2 INVESTIGACIONES REALIZADAS SOBRE EL TEMA

2.2.1 METODOLOGÍA DE BÚSQUEDA DE INFORMACIÓN

Se ha utilizado principalmente 2 metodologías de búsqueda de información las cuales son:

a) **SMS (Systematic Mapping Study):** Mapeo Sistemático de Estudio, es un método definido para construir un esquema de clasificación y estructurar un campo de interés de estudio. El análisis de los resultados se centra en la frecuencia de las publicaciones por categorías dentro del esquema. De esta manera, se puede determinar la cobertura del campo de investigación. También

se pueden combinar diferentes facetas del esquema para responder a preguntas de investigación más específicas.

Un estudio de mapeo sistemático proporciona una estructura del tipo de informes de investigación y resultados que se han publicado, categorizándolos y a menudo ofrece un resumen visual, el mapa de sus resultados.

A continuación, se muestra en la Ilustración 3, los pasos del proceso de mapeo sistemático.

Ilustración 3. Pasos del proceso de mapeo sistemático [9].

Los pasos principales son: la definición de las preguntas de investigación, la búsqueda de artículos relevantes, la selección de artículos, la redacción de resúmenes, la extracción y mapeo de datos. Cada etapa del proceso tiene un resultado, siendo el resultado final del proceso el mapa sistemático. En un estudio de mapeo sistemático, se pueden considerar más artículos ya que no tienen que ser evaluados con tanto detalle.

b) SLR (Systematic Literatury Review): Es la Revisión Sistemática de Literatura, que sirve para diseñar un protocolo de búsqueda de información.

Este revisa los informes primarios existentes, los revisa en profundidad y describe su metodología y resultados. Una SLR tiene varios beneficios: una metodología bien definida reduce el sesgo, una gama más amplia de situaciones y contextos puede permitir conclusiones más generales, pero todo esto requiere de un esfuerzo considerable. Es decir, lo que los distingue de los mapas sistemáticos es su análisis en profundidad en forma de un resumen narrativo detallado.

En cuanto a la caracterización de revisiones sistemáticas por lo general, se las hace en función de sus objetivos de investigación, los criterios de inclusión y exclusión, el número de inclusiones y exclusiones, el esquema de clasificación y los medios de análisis.

Los mapas y revisiones sistemáticas son diferentes en términos de objetivos, amplitud, cuestiones de validez e implicaciones. Por lo tanto, deben utilizarse de manera complementaria y, por ejemplo, para el análisis, requieren métodos diferentes. Primero se puede realizar un mapa sistemático para obtener una visión general del área temática y luego entonces, el estado de la evidencia en temas específicos puede ser investigado usando una revisión sistemática [9].

2.2.2 TRABAJOS RELACIONADOS

Para el presente proyecto he consultado información de los predictores de felicidad en el mundo, en 3 bases de datos científicas importantes como son:

- Sciencedirect
- Scopus
- IEEE

Y he realizado 3 tipos de búsquedas que se describen en forma resumida a continuación:

Nro.	Tipo de búsqueda	Base de datos	Cadena de búsqueda	Filtros
1	Primera búsqueda	Todas	data mining and prediction	a. Años del 2015 al 2019 b. Tipos de artículos: Review articles, Research articles, Book chapters
2	Segunda búsqueda	Todas	data mining and prediction and factor	a. Años del 2015 al 2019 b. Tipos de artículos: Review articles, Research articles, Book chapters
3	Tercera búsqueda	Sciencedirect	data mining and prediction and factor and happiness	En todo este tipo de búsqueda, se usaron los filtros: a. Años del 2015 al 2019 b. Tipos de artículos
		Scopus	data happiness	
		IEEE	data mining and prediction and happiness	

De las búsquedas realizadas, he encontrado muchos artículos que usan la minería de datos para predecir en distintos ámbitos, pero más concretamente relacionados con la felicidad (tercer búsqueda), en donde se ha considerado a los 5 primeros resultados obtenidos, los mismos que se detallan a continuación:

Cadena: data mining and prediction and factor and happiness

Búsqueda por Base de Datos: Sciencedirect

Filtros:

a) Años del 2015 al 2019

b) Tipos de artículos: Review articles, Research articles, Book chapters

Cantidad de Resultados: 158

Nro.	Nombre del artículo	Referencia	Autores	Abstract	Resumen	Fecha de consulta
1	A comprehensive study on the effects of using data mining techniques to predict tie strength	[10]	Mohammad Karim Sohrabi, Soodeh Akbari	The use of social networks has grown significantly in recent years and this has led to the production of numerous volumes of data. The uncertainties that arise from the complexity of recognition decisions among people have led researchers to look for effective variables of intimacy among people, since there are several effective variables whose effectiveness rate is not precisely determined and their relationships are non-linear and complex, the use of data mining techniques can be considered as one of the practical solutions to this problem. Data mining could be considered as one of the applicable tools for researchers to explore relationships between users. In this paper, the prediction problem is modeled as a data mining problem on which different methods of supervised and unsupervised mining are applied. A comprehensive study is proposed on the effects of the use of different classification techniques such as decision trees, naive bayes, etc.; as well as some set classification methods such as Bagging and Boosting methods for prediction. LinkedIn social network is used as a real case study and experimental results are proposed on their extracted data. Several models are created, based on basic techniques and assembly methods, and their efficiencies are compared according to F measurement, accuracy and average execution time. The experimental results show that the model based on the behaviour of the profiles is much more accurate compared to model techniques based on profile data.	El uso de las redes sociales ha crecido notablemente en los últimos años y este hecho ha llevado a la producción de numerosos volúmenes de datos. Las incertidumbres que surgen de la complejidad de las decisiones de reconocimiento entre las personas han llevado a los investigadores a buscar variables efectivas de intimidad entre las persones, dado que existen varias variables efectivas cuya tasa de efectividad no está determinada con precisión y sus relaciones son no lineales y complejas, el uso de técnicas de minería de datos puede ser considerado como una de las soluciones prácticas para este problema. La minería de datos podría considerarse como una de las herramientas aplicables para que los investigadores exploren las relaciones entre los usuarios. En este trabajo, el problema de la predicción se modela como un problema de minería de datos sobre el que se aplican diferentes métodos de minería supervisada y no supervisada. Se propone un estudio exhaustivo sobre los efectos de la utilización de diferentes técnicas de clasificación como árboles de decisión, bayes ingenuos, etc.; además de algunos métodos de clasificación de conjuntos como los métodos de Bagging y Boosting para la predicción. LinkedIn social network se utiliza como un estudio de caso real y los resultados experimentales se proponen sobre sus datos extraídos. Se crean varios modelos, basados en técnicas básicas y métodos de ensamblaje, y se comparan sus eficiencias en función de la medida F, la precisión y el tiempo medio de ejecución. Los resultados experimentales muestran que el modelo basado en el comportamiento de los perfiles tiene una precisión mucho mayor en comparación con las técnicas de modelos basados en datos de perfiles.	8/5/2019

2	Who needs a reason to indulge? Happiness following reason-based indulgent consumption	[11]	Francine Espinoza Petersen, Heather Johnson Dretsch, Yuliya Komarova Loureiro	This research identifies a condition under which to indulge without a reason "feels good" and produces a very positive emotional reaction. The authors show that complacency with or without a reason and self-control of consumer traits interact to influence happiness after an indulgent purchase. While consumers with high self-control are happier when they have a reason to buy indulgent products, consumers with low self-control are happier when they have no reason to indulge. That is, indulging in a reason is less pleasurable for consumers with low self-control. This effect on happiness has an impact on subsequent judgments about the product and has important implications for the well-being of consumers as well as marketing managers. Through four studies, the effect on happiness in consumption is presented, the consequences of the effect are examined, and the evidence for the underlying process is reported.	Esta investigación identifica una condición bajo la cual darse el gusto sin una razón "se siente bien" y produce una reacción emocional muy positiva. Los autores muestran que la complacencia con o sin una razón y el autocontrol de los rasgos de los consumidores interactúan para influir en la felicidad después de una compra indulgente. Mientras que, los consumidores con alto autocontrol son más felices cuando tienen una razón para comprar productos indulgentes, los consumidores con bajo autocontrol son más felices cuando no tienen una razón para darse el gusto. Es decir, darse el gusto de una razón es menos placentero para los consumidores con un bajo autocontrol. Este efecto sobre la felicidad tiene un impacto en los juicios posteriores sobre el producto y tiene importantes implicaciones para el bienestar de los consumidores, así como para los directores de marketing. A través de cuatro estudios, se presenta el efecto sobre la felicidad en el consumo, se examina las consecuencias del efecto y se reporta la evidencia para el proceso subyacente.	8/5/2019
3	A survey on big data-driven digital phenotyping of mental health	[12]	Yunji Liang, Xiaolong Zheng, Daniel D. Zeng	The mental health landscape has changed enormously over the last two decades, but research into mental health is still in its infancy, with significant gaps in knowledge and a lack of accurate diagnosis. Today, large data and artificial intelligence offer new opportunities for the detection and prediction of mental problems. This review article outlines the vision of digital phenotyping of mental health (DPMH) by merging enriched data from ubiquitous sensors, social media, and health systems, and presents a broad picture of DPMH from the perspectives of detection and computation. First, a systematic review of the literature is conducted and then a research framework is proposed that highlights key mental health issues and discusses the challenges posed by enriched data for digital phenotyping. Five key lines of research are then developed, including affection recognition, cognitive analysis, detection of behavioral anomalies, social analysis, and biomarker analysis in the psychiatric context.	El panorama de la salud mental ha experimentado enormes cambios en las últimas dos décadas, pero la investigación sobre este tipo de salud se encuentra todavía en su fase inicial, con importantes lagunas de conocimiento y la falta de un diagnóstico preciso. Hoy en día, los grandes datos y la inteligencia artificial ofrecen nuevas oportunidades para la detección y predicción de problemas mentales. En este artículo de revisión, se esboza la visión del fenotipado digital de la salud mental (DPMH) mediante la fusión de los datos enriquecidos de los sensores ubicuos, los medios sociales y los sistemas de salud, y se presenta un amplio panorama de DPMH desde las perspectivas de la detección y la computación. En primer lugar, se realiza una revisión sistemática de la literatura y luego se propone el marco de investigación que destaca los aspectos clave relacionados con la salud mental, y se discuten los desafíos que plantean los datos enriquecidos para el fenotipado digital. A continuación, se desarrollan cinco líneas de investigación clave que incluyen reconocimiento de afecto, análisis cognitivo, detección de anomalías del comportamiento, análisis social y análisis de biomarcadores en el contexto psiquiátrico.	8/5/2019

4	The new income inequality and well-being paradigm: Inequality has no effect on happiness in rich nations and normal times, varied effects in extraordinary circumstances, increases happiness in poor nations, and interacts with individuals' perceptions, attitudes, politics, and expectations for the future	[13]	Jonathan Kelley, M. D. R. Evans	On the basis of more recent evidence, the views of academics are beginning to merge into the consensus that national income inequality is irrelevant to the subjective well-being of individuals in advanced nations and in normal times, as demonstrated by multi-level models with appropriate controls. For developing countries, the consensus is not as strong, but most of the evidence indicates a neutral or positive effect on inequality. On this basis, this paper provides exploratory analyses to stimulate future research, broadening understanding of the social, psychological and cultural forces that generate these results; analyzes changes over time and expectations for the future; and addresses the possibility that inequality may reduce well-being in extraordinary circumstances and for particular groups.	Sobre la base de pruebas más recientes, las opiniones de los académicos están empezando a fundirse en el consenso de que la desigualdad del ingreso nacional es irrelevante para el bienestar subjetivo de los individuos en las naciones avanzadas y en tiempos normales, como lo demuestran los modelos multiniveles con controles apropiados. Para los países en desarrollo, el consenso no es tan fuerte, pero la mayor parte de la evidencia indica un efecto neutro o positivo para la desigualdad. Sobre esta base, este documento proporciona análisis exploratorios para estimular la investigación futura, ampliando la comprensión de las fuerzas sociales, psicológicas y culturales que generan estos resultados; analiza los cambios a lo largo del tiempo y las expectativas para el futuro; y aborda la posibilidad de que la desigualdad pueda reducir el bienestar en circunstancias extraordinarias y para grupos particulares.	8/5/2019
5	We Are What We Generate - Understanding Ourselves Through Our Data	[14]	Muhammad Fahim Uddin, Jeongkyu Lee	In this work, the idea of understanding individuals through the data lens they produce is developed in the context of the main research work for the algorithm engine Predicting Educational Relevance For an Efficient Classification of Talent (PERFECT). It presents some of the research problems in terms of the relevance of such data and identifies the research problem as the basis for this work. A subset of the framework is presented that includes algorithms and mathematical constructs for the problem being identified. The conclusion is that such analytical and cognitive research can help improve education, health, labour economics, crime control, etc.	En este trabajo, se desarrolla la idea de entender a los individuos a través de la lente de datos que producen en el contexto del principal trabajo de investigación para el motor de algoritmos Predicting Educational Relevance For an Efficient Classification of Talent (PERFECT) (Predicción de la Relevancia Educativa para una Clasificación Eficiente de Talentos). Se presentan algunos de los problemas de la investigación en cuanto a la relevancia de tales datos e identifica el problema de la investigación como base para este trabajo. Se presenta un subconjunto del marco de trabajo que incluye algoritmos y construcciones matemáticas para el problema que se identifica. La conclusión que se obtiene es que dicha investigación analítica y cognitiva puede ayudar a mejorar la educación, la salud, la economía laboral, el control de la delincuencia, etc.	8/5/2019

Cadena: data happiness

Búsqueda por Base de Datos: Scopus

Filtros:

a) Años del 2015 al 2019

b) Tipos de artículos: Artículo electrónico, Material de conferencias, Libro electrónico y Tesis

Cantidad de Resultados: 2372

Nro.	Nombre del artículo	Enlace	Autores	ISSN	Abstract	Resumen	Fecha de consulta
1	A happiness degree predictor using the conceptual data structure for deep learning architectures	[15]	Pérez-Benito, Francisco Javier;Villacampa-Fernández, Patricia;Conejero	0169-2607	A deep learning architecture driven by the conceptual data structure for the prediction of Happiness.•The lower-level dimensions of psychological factors are separately ensembled for then being merged by higher-level dimensions until happiness is reached.•Two operators using the layers weights for the extraction of conclusions about the influence of the psychological factors in happiness are proposed.•The prediction of happiness is improved by not assuming linear relationships between factors.	Una arquitectura de aprendizaje profundo impulsada por la estructura conceptual de datos para la predicción de la Felicidad: las dimensiones inferiores de los factores psicológicos se ensamblan por separado para luego fusionarse con las dimensiones superiores hasta alcanzar la felicidad; se proponen dos operadores que utilizan los pesos de las capas para extraer conclusiones sobre la influencia de los factores psicológicos en la felicidad; la predicción de la felicidad se mejora al no asumir relaciones lineales entre los factores.	9/5/2019
2	Can happiness apps generate nationally representative datasets? - a case study collecting data on people's happiness using the german socio-economic panel.	[16]	Lucas, Richard.	1871-2584 (Print) 1871-2576 (Electronic)	In recent years, applications have become an important tool for collecting data, especially on people's happiness, two projects have received substantial attention from both the media and the scientific world: "Track your happiness" and "Mappiness". Both happiness applications used the experience sampling method to ask people a few times a day how they feel, what they do, with whom and where. Both studies have collected considerable data without giving participants any financial rewards. But quantity is not all that matters with respect to data collection, and so it is crucial to understand whether nationally representative datasets can be collected using such happiness applications.	En los últimos años, las aplicaciones se han convertido en una herramienta importante para recopilar datos, especialmente en los de la felicidad de las personas, dos proyectos han recibido una atención sustancial tanto de los medios de comunicación como del mundo científico: "Rastrea tu felicidad" y "Mappiness". Ambas aplicaciones de felicidad usaron el método de muestreo de experiencias para preguntar a las personas unas cuantas veces al día cómo se sienten, qué hacen, con quién y dónde. Ambos estudios han recopilado datos considerables sin dar a los participantes ninguna recompensa financiera. Pero la cantidad no es todo lo que importa con respecto a la recopilación de datos y, por lo tanto, es crucial entender si se pueden recopilar conjuntos de datos representativos a nivel nacional utilizando dichas aplicaciones de felicidad.	9/5/2019

3	The effect of pets on happiness: A data-driven approach via large-scale social media	[17]	Wu, Yuchen Yuan, Jianbo You, Quanzeng Luo, Jiebo	978-1-4673-9005-7	Psychologists have shown that pets have a positive impact on owners' happiness. In this paper, a novel and effective approach is proposed that exploits social media to study the effect of pets on owners' happiness. The proposed framework includes three main components: 1) collect Instagram user-level data consisting of about 300,000 images from 2,905 users; 2) build a convolutional neural network (CNN) for the classification of pets and combine it with timeline information to better identify pet owners and the control group; 3) measure the happiness confidence score by detecting and analyzing images of the pets themselves. The experimental results demonstrate the effectiveness of the proposed approach and it is believed that this approach can be applied to other related areas such as a large-scale and highly reliable methodology for the analysis of user activity through social media.	Los psicólogos han demostrado que las mascotas tienen un impacto positivo en la felicidad de los dueños. En este trabajo, se propone un enfoque novedoso y efectivo que explota los medios sociales para estudiar el efecto de las mascotas en la felicidad de los propietarios. El marco propuesto incluye tres componentes principales: 1) recopilar datos a nivel de usuario de Instagram que consisten en unas 300.000 imágenes de 2.905 usuarios; 2) construir una red neuronal convolucional (CNN) para la clasificación de las mascotas y combinarla con información de la línea de tiempo, para identificar mejor a los dueños de las mascotas y al grupo de control; 3) medir la puntuación de confianza de la felicidad detectando y analizando las imágenes de las propias mascotas. Los resultados experimentales demuestran la eficacia del enfoque propuesto y se crees que este enfoque puede aplicarse a otros ámbitos relacionados como una metodología a gran escala y de gran confianza para el análisis de la actividad de los usuarios a través de los medios sociales.	9/5/2019
4	Valuing Air Quality Using Happiness Data: The Case of China	[18]	Zhang, X.;Zhang, X.;Chen, X.	0921-8009	This document estimates the monetary value of the reduction of PM2.5, one of the main sources of air pollution in China. By comparing hedonic happiness in a nationally representative survey with daily air quality data according to the dates and counties of interviews in China, the relationship between local particle concentration and individual happiness is estimated. By maintaining constant happiness, compensation is calculated between particulate matter reduction and income, essentially a measure based on the happiness of willingness to pay to mitigate air pollution. People on average are willing to pay ¥258 per year per person for a 1% reduction in PM2.5.	Este documento estima el valor monetario de la reducción de las PM2,5, una de las principales fuentes de contaminación atmosférica en China. Al comparar la felicidad hedónica en una encuesta representativa a nivel nacional con los datos diarios de la calidad del aire de acuerdo con las fechas y condados de las entrevistas en China, se estima la relación entre la concentración local de partículas y la felicidad individual. Al mantener la felicidad constante, se calcula la compensación entre la reducción de la materia particulada y el ingreso, esencialmente una medida basada en la felicidad de la disposición a pagar para mitigar la contaminación del aire. La gente en promedio está dispuesta a pagar ¥258 por año por persona para una reducción del 1% en PM2.5.	9/5/2019

5	Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity	[19]	Nguyen, Quynh C.;Kath, Suraj;Meng, Hsien-Wen;Li, Dapeng;Smith, Ken R	0143-6228	Using publicly available twitter and geotagged data, neighborhood indicators are created for happiness, food, and physical activity in three large counties: Salt Lake, San Francisco, and New York. Manually tagged tweets and tweets tagged with algorithms had excellent levels of agreement: 73% for happiness; 83% for food; and 85% for physical activity. Social networks can be harnessed to provide a greater understanding of the well-being and health behaviors of communities, information that was previously difficult and costly to obtain consistently across geographic regions. More open access to neighborhood data can enable better design of programs and policies that address the social determinants of health.	Usando datos de twitter disponibles al público y geoetiquetados, se crean indicadores de vecindarios para la felicidad, la comida y la actividad física en tres condados grandes: Salt Lake, San Francisco y Nueva York. Los tweets etiquetados manualmente y los tweets etiquetados con algoritmos tuvieron excelentes niveles de acuerdo: 73% para la felicidad; 83% para la comida y 85% para la actividad física. Las redes sociales pueden ser aprovechadas para proporcionar una mayor comprensión del bienestar y los comportamientos de salud de las comunidades, información que antes era difícil y costosa de obtener de manera consistente en todas las regiones geográficas. Un acceso más abierto a los datos de los vecindarios puede permitir un mejor diseño de programas y políticas que aborden los determinantes sociales de la salud.	9/5/2019

Cadena: data mining and prediction and happiness

Búsqueda por Base de Datos: IEEE

Filtros:

a) Años del 2015 al 2019

Cantidad de Resultados: 5

Nro.	Nombre del artículo	Enlace	Autores	Año de publicación	ISSN/ISBN	Abstract	Resumen	Fecha de consulta
1	Feeling Analysis for Sadness and Happiness using Googlen-gram Database Googlen-gram Veritabanı ile Üzüntü ve Mutluluk Üzerine Duygu Analizi	[20]	Ilknur Dönmez ; Elena Battini Sönmez	2018	Electronic ISBN: 978-1-5386-7893-0 USB ISBN: 978-1-5386-7892-3 Print on Demand(PoD) ISBN: 978-1-5386-7894-7	The current era has been defined as "digital age" and "information age", since it is characterized by an exponential growth of data, generated both by the human being and by the machine, that is to say, by the Internet of things. The challenge is to turn "data" into "information", analysing the data and discovering the hidden patterns inside. In this work the two basic human feelings of Happiness and Sadness are extracted and analysed from a subset of the corpus of Google n-grams. The corpus n-grams of Google can be considered as an indicator of specific characteristics and behaviors of the human being. Under the hypothesis that the user's emotion can be extrapolated by the frequency of emotional words, this study applies regression to predict the importance of the emotional states of Happiness and Sadness in future years.	La era actual se ha definido como "era digital" y "era de la información", ya que se caracteriza por un crecimiento exponencial de los datos, generados tanto por el ser humano como por la máquina, es decir, por la Internet de las cosas. El reto es convertir los "datos" en "información", analizando los datos y descubriendo los patrones ocultos en su interior. En este trabajo se extraen y analizan los dos sentimientos humanos básicos de Felicidad y Tristeza de un subconjunto del corpus de Google n-grams. El corpus n-grams de Google puede ser considerado como un indicador de características y comportamientos específicos del ser humano. Bajo la hipótesis de que la emoción del usuario puede ser extrapolada por la frecuencia de las palabras emocionales, este estudio aplica regresión para predecir la importancia de los estados emocionales de Felicidad y Tristeza en años futuros.	7/5/2019
2	The Automatic Recognition of Sepedi Speech Emotions Based on Machine Learning Algorithms	[21]	Phuti J Manamela ; Madimetja J Manamela ; Thipe I Modipa ; Tshepisho J Sefara ; Tumisho B Mokgonyane	2018	Electronic ISBN: 978-1-5386-3060-0 Print on Demand(PoD) ISBN: 978-1-5386-3061-7	This paper analyzes a system of speech emotion recognition studies (SER) that classifies and recognizes six basic emotions (anger, sadness, disgust, fear, happiness and neutrality) of speech spoken in sepedi language (one of the official languages of South Africa). Recordings of the speech of speakers of the Sepedi language were collected and a television drama was broadcast to create an emotional corpus of speech. Thirty-	En este trabajo se analiza un sistema de estudios de reconocimiento de emociones del habla (SER) que clasifica y reconoce seis emociones básicas (ira, tristeza, disgusto, miedo, felicidad y neutralidad) del habla hablada en lengua sepedi (una de las lenguas oficiales de Sudáfrica). Se recogieron grabaciones del habla de los hablantes de la lengua Sepedi y se emitió un drama televisivo para crear corpus emocionales del habla. A continuación, se extrajeron 34	7/5/2019

						four voice functions were then extracted from the voice corpus, using the pyAudioAnalysis tool, to train and compare different algorithms using 10-fold cross validation. The experiments were conducted using WEKA's data mining software. The results showed that Auto-WEKA exceeds all standard algorithms (SVM. KNN and MLP). The recorded voice corpus produced good recognition accuracy compared to the voice corpus transmitted by television.	funciones de voz de los corpus de voz, utilizando la herramienta pyAudioAnalysis, para entrenar y comparar diferentes algoritmos utilizando la validación cruzada de 10 pliegues. Los experimentos se llevaron a cabo utilizando el software de minería de datos de WEKA. Los resultados mostraron que Auto-WEKA supera todos los algoritmos estándar (SVM. KNN y MLP). El corpus de voz grabado produjo una buena precisión de reconocimiento en comparación con el corpus de voz transmitido por televisión.	
3	Mining Social Emotions from Affective Text	[22]	Shenghua Bao ; Shengliang Xu ; Li Zhang ; Rong Yan ; Zhong Su ; Dingyi Han ; Yong Yu	2012	Print ISSN: 1041-4347 Electronic ISSN: 1558-2191 CD-ROM ISSN: 2326-3865	This article deals with the problem of extracting social emotions from the text. Recently, with the rapid development of Web 2.0, more and more documents are assigned by social users with tags of emotions such as joy, sadness and surprise. In this work, we intend to discover the connections between social emotions and affective terms and on the basis of which to predict social emotion from the content of the text automatically. More specifically, propose a joint emotion theme model by increasing the allocation of latent dirichlets with an additional layer for emotion modeling. It first generates a set of latent themes from the emotions, followed by the generation of affective terms from each theme. Experimental results from an online news collection show that the proposed model can effectively identify significant latent themes for each emotion. The evaluation of emotion prediction further verifies the effectiveness of the proposed model.	Este artículo se ocupa del problema de la extracción de emociones sociales del texto. Recientemente, con el rápido desarrollo de la web 2.0, cada vez más documentos son asignados por usuarios sociales con etiquetas de emociones como alegría, tristeza y sorpresa. En este trabajo, se pretende descubrir las conexiones entre las emociones sociales y los términos afectivos y en base a los cuales predecir la emoción social a partir del contenido del texto de forma automática. Más específicamente, proponer un modelo de tema de emoción conjunta mediante el aumento de la asignación de dirichlets latentes con una capa adicional para el modelado de emociones. Primero genera un conjunto de temas latentes a partir de las emociones, seguido de la generación de términos afectivos a partir de cada tema. Los resultados experimentales de una colección de noticias en línea muestran que el modelo propuesto puede identificar eficazmente temes latentes significativos para cada emoción. La evaluación de la predicción de emociones verifica aún más la eficacia del modelo propuesto.	7/5/2019

4	Unsupervised Learning Based On Artificial Neural Network: A Review	[23]	Happiness Ugochi Dike ; Yimin Zhou ; Kranthi Kumar Deveerasetty ; Qingtian Wu	2018	Electronic ISBN: 978-1-5386-7355-3 USB ISBN: 978-1-5386-7354-6 Print on Demand(PoD) ISBN: 978-1-5386-7356-0	Artificial neural networks (ANNs) have been effectively applied in numerous fields in order to predict, discover knowledge, classify, analyse time series, model, etc. There are some limitations to the use of supervised learning. These limitations can be overcome by using unsupervised learning techniques. A major problem associated with unsupervised learning is how to find hidden structures in unlabeled data. This article reviews the training/learning of unsupervised learning based on the artificial neural network. It provides a description of methods for selecting and repairing a series of hidden nodes in an ANN-based unsupervised learning environment. In addition, the situation, benefits, and challenges of unsupervised learning are also summarized.	Las redes neuronales artificiales (RNA) se han aplicado eficazmente en numerosos campos con el fin de predecir, descubrir conocimientos, clasificar, analizar series temporales, modelar, etc. Existen algunas limitaciones en el uso del aprendizaje supervisado. Estas limitaciones pueden superarse utilizando técnicas de aprendizaje no supervisadas. Un problema principal asociado con el aprendizaje no supervisado es cómo encontrar las estructuras ocultas en los datos no etiquetados. Este artículo revisa el entrenamiento/aprendizaje del aprendizaje no supervisado basado en la red neuronal artificial. Proporciona una descripción de los métodos para seleccionar y reparar una serie de nodos ocultos en un entorno de aprendizaje no supervisado basado en la RNA. Además, también se resumen la situación, los beneficios y los desafíos del aprendizaje no supervisado.	7/5/2019
5	Learning to recommend descriptive tags for health seekers using deep learning	[24]	Vidhi L. Chawda ; Vishwanath S. Mahalle	2017	Electronic ISBN: 978-1-5090-4715-4 Print on Demand(PoD) ISBN: 978-1-5090-4716-1	Health plays an important role for human happiness and well-being. Automatic disease prediction is important to overcome the problems of health seekers. Generally, people use Google to search their queries and that search engine responds with the answer, but that answer is in scattered format. The user does not get an exact answer to his or her queries. In addition, a novel scheme of deep learning is proposed to infer the disease according to the questions of the health search engines. This work first analyzes and categorizes the needs of health seekers and then asks for symptoms for the prediction of diseases. The user will then search for their query. The query is then processed to give disease prediction to the user or health seekers. The concept of hidden layers is used here. The first medical firms extract the mines from the raw traits. These characteristics	La salud juega un papel importante para la felicidad y el bienestar humano. La predicción automática de enfermedades es importante para superar los problemas de los buscadores de salud. Generalmente, la gente utiliza Google para buscar sus consultas y ese motor de búsqueda les responde con la respuesta, pero esa respuesta está en formato disperso. El usuario no obtiene respuesta exacta para sus consultas. Además, se propone un novedoso esquema de aprendizaje profundo para inferir la enfermedad de acuerdo a las preguntas de los buscadores de salud. En este trabajo primero se analiza y categoriza las necesidades de los buscadores de salud y luego se pide síntomas para la predicción de enfermedades. Entonces el usuario buscará su consulta. Luego se procesa la consulta para dar predicción de la enfermedad al usuario o a los buscadores de salud. Aquí se utiliza el concepto de capas ocultas. Las primeras firmas médicas extraen las minas	7/5/2019

						and signatures are considered entry nodes in one layer and hidden nodes in the next layer. This article presents an idea of the deep learning architecture that is used in healthcare for the diagnosis of disease.	de los rasgos en bruto. Estas características y firmas se consideran nodos de entrada en una capa y nodos ocultos en la capa siguiente. Este artículo presenta una idea de la arquitectura de aprendizaje profundo que se utiliza en el ámbito de la atención sanitaria para el diagnóstico de enfermedades.	

Se presenta a continuación un detalle de la cantidad de artículos obtenidos por cada una de las bases de datos científicas y por cada búsqueda con las que se trabajó.

Nro.	Base de datos	Primera búsqueda	Segunda búsqueda	Tercera búsqueda	Fecha
1	Sciencedirect	24120	19603	158	07-04-2019
2	Scopus	10716	353	2372*	01-05-2019
3	IEEE	3634	968	5**	07-05-2019
Total		**38470**	**20924**	**2535**	

***** Este dato resulta atípico porque si se colocaba la cadena de búsqueda: "data mining and prediction and factor and happiness" no se encontró ningún resultado, por lo que se modificó a: "data happiness".

****** Con la cadena "data mining and prediction and factor and happiness" no se encontró ningún resultado, por lo que se modificó a: "data mining and prediction and happiness".

La base de datos en donde se obtuvo mayor cantidad de artículos, relacionados con el tema del proyecto y donde resultó tener una interfaz amigable para consultar fue **Sciencedirect**, la cual desde ahora mismo la recomiendo utilizar.

Adicional, he encontrado muchos más **trabajos desarrollados y relacionados** con el presente tema de predictores de felicidad a nivel mundial, de los cuales he seleccionado 6 y se explican de manera general a continuación:

Artículos más destacados relacionados con predictores de felicidad a nivel mundial

Nro.	Nombre del artículo	Enlace	Resumen	Datos
1	Informe de la felicidad mundial	https://www.researchgate.net/publication/233401584_World_Happiness_Report	En el informe se hace hincapié que el dinero es un factor importante en la vida de un ser humano, pero solo con este no se consigue la felicidad. Que se debe proteger la tierra para conseguir una buena calidad de vida con lo cual se adoptará mejores estilos de vida y tecnologías que ayuden a mejorar la felicidad. La búsqueda de la felicidad está íntimamente vinculada a la búsqueda del desarrollo sostenible. Los factores externos importantes son: el ingreso, el trabajo, la comunidad y el gobierno, valores y religión. Entre los factores personales están: salud física y mental, la experiencia familiar, la educación, el género y la edad. Muchos de estos factores tienen una interacción bidireccional con la felicidad. Existen muchas otras variables que tienen un mayor efecto en la felicidad, que son: la confianza social, la calidad de trabajo y la libertad de elección y participación política. En muchos casos hay una interacción bidireccional entre el factor y la felicidad. La felicidad de una persona depende de sus propios valores, pero también de los valores de los que la rodean. Las personas que se preocupan más por los demás son típicamente más felices que las que se preocupan más por sí mismas. Amar y ser amado son condiciones clave para la felicidad humana. La educación está indirectamente relacionada con la felicidad a través de su efecto sobre los ingresos: la educación aumenta los ingresos y los ingresos aumentan la felicidad. Un buen trabajo es aquel que proporciona felicidad y satisfacción al trabajador. La felicidad se alcanza cuando las personas alcanzan la suficiencia en aproximadamente cuatro de los seis dominios o en la proporción equivalente de condiciones. Las investigaciones han demostrado que es posible recopilar datos significativos y fiables sobre el bienestar subjetivo. El bienestar subjetivo abarca tres aspectos diferentes: las evaluaciones cognitivas de la vida, las emociones positivas (alegría, orgullo) y las negativas (dolor, ira, preocupación). Todos estos aspectos del bienestar subjetivo deben medirse por separado para obtener una medida más completa de la calidad de vida de las personas y permitir una mejor comprensión de sus determinantes.	Recogidos y analizados del año 2010 y que luego fueron publicados en el 2012.

2	Informe de la felicidad mundial 2018	https://s3.amazonaws.com/happiness-report/2018/WHR_web.pdf	Hay grandes brechas de felicidad entre los países, y éstas continuarán creando grandes presiones para emigrar. Algunos de los que emigran entre países se beneficiarán y otros perderán. En general, los que se mudan a países más felices que los suyos ganarán en felicidad, mientras que los que se mudan a países más infelices tenderán a perder. Los que se quedan atrás no perderán en promedio. La inmigración seguirá planteando oportunidades y costos para los que se mudan, para los que se quedan atrás y para los nativos de los países de acogida de inmigrantes. Hay un nuevo país que encabeza la clasificación, Finlandia, pero los diez primeros puestos los ocupan los mismos países que en los dos últimos años, aunque con un cierto intercambio de puestos. Cuatro países diferentes han ocupado el primer lugar desde 2015: Suiza, Dinamarca, Noruega y ahora Finlandia. Seis variables clave que se ha comprobado que apoyan el bienestar: PIB, esperanza de vida, ayuda social, libertad en sociedad, confianza y generosidad. Las clasificaciones de felicidad de los inmigrantes se basan en toda la gama de datos de Gallup desde 2005 hasta 2017, lo que es suficiente para tener 117 países con más de 100 encuestados inmigrantes. La felicidad puede cambiar, y cambia, según la calidad de la sociedad en la que vive la gente. Los países con los inmigrantes más felices no son los países más ricos, sino los países con un conjunto más equilibrado de apoyos sociales e institucionales para una vida mejor. Una respuesta inmediata entre los lectores y comentaristas es sugerir que la gente debería mudarse a una comunidad más feliz para hacerse más feliz. Se consideraron tres tipos de resultados de felicidad: evaluaciones de vida, afecto positivo (experiencias de disfrute, felicidad y risa) y afecto negativo (experiencias de preocupación, tristeza y enojo). Una mayor felicidad a menudo acompaña a una mayor salud y seguridad. La migración internacional es, para muchas personas, un poderoso instrumento para mejorar sus vidas, dado que la mayoría de los migrantes y sus familias se benefician considerablemente de la migración. Sin embargo, no todos los migrantes y sus familias se alegran con la migración, y la felicidad de los migrantes no aumenta con el tiempo a medida que se aclimatan a su nuevo país. Es probable que los migrantes del campo a la ciudad carezcan de la información necesaria para poder juzgar la calidad de sus nuevas vidas en un mundo diferente.	Recogidos y analizados desde los años 2006 al 2017.

			La felicidad de los latinoamericanos se ve disminuida por sus muchos problemas sociales y económicos y que, de hecho, la felicidad podría aumentar si estos problemas se abordaran adecuadamente. Hay muchos factores positivos que contribuyen a la felicidad de los latinoamericanos; en particular, la abundancia y calidad de la cercanía, específica de los países de América Latina. Las relaciones interpersonales les permiten disfrutar de altos niveles de satisfacción en ámbitos de la vida que son particularmente importantes para los latinoamericanos: el ámbito social y, en especial, el familiar. El desafío del bienestar es una cuestión tanto de política y economía de alto nivel como de la suma de los esfuerzos individuales y comunitarios.	
3	Minería de Datos y Análisis de Datos	http://www.datamining.org.uk/MSC_THESIS_FINAL_VERSION.pdf	El objetivo del proyecto es explicar y predecir la felicidad global. Este fue un estudio transversal que incluyó 123 países, cada uno con un valor medio de felicidad. Esta verdad básica se estableció a partir de los datos de la encuesta, utilizando la respuesta a una pregunta sobre la satisfacción con la vida. El análisis inicial de los datos se realizó para descubrir patrones en los datos utilizando PCA, visualizaciones y correlaciones. Los datos se prepararon en primer lugar mediante la imputación de los valores que faltaban con el método k- del vecino más cercano. Se encontró que el conjunto de características era tan predictivo como las variables económicas. La selección de las características se realizó utilizando el lazo, los mínimos cuadrados y los árboles de decisión. La importancia de los resultados se determinó encontrando umbrales estadísticos de pruebas mediante pruebas de permutación y bootstrapping. Los aspectos más destacados de este proyecto son: • PCA descubrió una interesante relación entre la igualdad de género y la satisfacción en la vida. • Los árboles de decisión demostraron ser un método efectivo tanto en la selección de características como en la predicción de la satisfacción con la vida. • Nuestras características clave tuvieron un rendimiento significativamente mejor que las variables económicas. • Los modelos gráficos ayudaron a investigar la estructura de las relaciones de las variables. • Se utilizó un método de visualización de datos de efectivo para demostrar los resultados. La felicidad subjetiva también representa la felicidad de una persona. Las tres principales encuestas globales utilizadas en este tipo de investigación son la World Value	Recogidos y analizados desde los años 2008 al 2011.

			Survey (WVS), la Gallup World Poll (GWP) y la World Database of Happiness (WDH). La satisfacción con la vida se correlaciona con el "control de la vida", mientras que la felicidad se correlacionaba con la "relación estable". La felicidad es más emocional y la satisfacción de la vida es más cognitivo. Este trabajo ha proporcionado una valiosa perspectiva de la naturaleza dual de la correlación entre crecimiento y felicidad (a corto y largo plazo). Las implicaciones de una correlación entre la felicidad y el clima son enormes. La preparación de datos y la selección de características son dos partes críticas de un proyecto de minería de datos. KNN será útil para imputar valores perdidos en nuestro conjunto de datos. Se analizaron ocho encuestas, que se realizaron alrededor de 2008, y esto incluye 4 encuestas globales, 3 europeas y 1 latinoamericana. Antes del análisis, los datos se transformaron a una escala consistente de 0 a 10. Los datos de HPI (Gallup) y WVS han sido seleccionados como etiquetas de felicidad apropiadas. GAL se correlaciona mejor con algunas de las características que con WVS. Las características clave identificadas son: • Salud: esperanza de vida, tasa de mortalidad • Educación: enseñanza primaria, enseñanza secundaria. • Igualdad: proporción-mujeres-parlamento, distribución de ingresos • Libertad: índice de libertad.	
4	DATA1001 Project #1: Informe de Felicidad Global creado por la Red de Soluciones para el Desarrollo Sostenible de las Naciones Unidas	https://rpubs.com/koki25ando/DATA1001TeamBver1	La búsqueda de la felicidad ha sido parte de la humanidad durante más tiempo de lo que algunos piensan; algunos incluso argumentan que es la razón por la que continuamos haciendo algo más que sólo existir. Para el primer proyecto de Data1001, se decidió centrarse en el análisis y la presentación de datos en torno al Informe Global de Felicidad. Gallup es la empresa americana de análisis y datos. La SDSN los habría contratado para la tarea específica de realizar encuestas de felicidad en todo el mundo. Siendo tan grandes y acreditados como lo son ellos, les conviene asegurarse de que la información se recopile correctamente. La puntuación de felicidad se obtuvo a partir de los datos de Gallup, lo que se hizo mediante encuestas	Recogidos y analizados desde los años 2015 al 2017.

			telefónicas aleatorias en países donde ese era un método probado. En los países donde no lo era o donde las líneas telefónicas no estaban disponibles en todo el país, se llevaron a cabo encuestas cara a cara. A las personas encuestadas se les preguntó lo siguiente: "Imagínese una escalera, con escalones numerados del 0 en la parte inferior al 10 en la parte superior. La parte superior de la escalera representa la mejor vida posible para usted y la parte inferior de la escalera representa la peor vida posible para usted. ¿En qué escalón de la escalera dirías que te sientes personalmente parado en este momento?". Si te preguntáramos ahora, hoy, cuáles son tus metas en la vida, ¿qué quieres de la vida? Todos tendrían diferentes ideas/pensamientos. ¡Si!, todos tenemos diferentes deseos, prioridades y aspiraciones. Sin embargo, habría un tema subyacente de querer ser feliz. En 2012, la ONU declaró el 20 de marzo como Día Internacional de la Felicidad y produjo su primer Informe de Felicidad. Un estudio sobre el estado de la felicidad global y el reconocimiento de que "la búsqueda de la felicidad es un objetivo y un derecho humano fundamental". Cada vez más, la felicidad se considera la medida adecuada del progreso social y el objetivo de las políticas públicas. El informe es realizado por expertos de diferentes campos de la psicología, la salud, la economía, etc., y analiza seis condiciones clave de cada país: 1. prosperidad económica, incluido el trabajo decente para todos los que lo deseen; 2. la salud física y mental de los ciudadanos; 3. la libertad de los individuos para tomar decisiones clave en la vida; 4. redes de apoyo social fuertes y dinámicas (capital social); 5. valores públicos compartidos de generosidad; 6. confianza social, incluida la confianza en la honestidad de las empresas y del gobierno. La suma de las condiciones determina la nota global para cada país.	
5	¿Qué fórmula usar para obtenir la felicidad nacional?	https://www.elsevier.com/locate/seps	El Informe sobre la Felicidad Mundial contiene una clasificación internacional de la felicidad media nacional, medida por encuestas de evaluación de la vida personal. También contiene un análisis que intenta explicar las cifras de felicidad de más de 150 países utilizando datos sobre seis variables clave.	Recogidos y analizados

			Este análisis parte de la base de que los factores se combinan de manera aditiva y, por lo tanto, funcionan independientemente unos de otros. Por el contrario, exploramos un modelo multiplicativo, que permite la interactividad o sinergia entre los factores, así como la posibilidad de disminuir el beneficio marginal a niveles más altos de logro. Encontramos que este modelo proporciona un mejor ajuste a los datos y por lo tanto es superior en su poder explicativo. La implicación para los responsables de la formulación de políticas es que deben centrarse en mejorar los factores que son los más bajos para su nación, ya que esto proporcionará mayores beneficios relativos al bienestar subjetivo. A nivel individual, esto significa centrarse en mejorar las condiciones de aquellos que están experimentando los niveles más bajos de bienestar. Esto se conoce como la pregunta de la escalera de Cantril y el resultado se toma como una medida del bienestar subjetivo, o menos formalmente, la felicidad. A continuación, se intenta explicar la variación entre países por medio de estas seis variables explicativas: • PIB per cápita PIB per cápita, • apoyo social SS, • esperanza de vida saludable HLE, • libertad para tomar decisiones de vida FRE, • generosidad GEN, • percepción (ausencia de) corrupción.	desde los años 2014 al 2016.
6	La relación entre estatus y felicidad: Evidencia del sistema de castas en las zonas rurales	www.elsevier.com/locate/jbee	Un gran número de estudios empíricos han investigado la relación entre el estatus social y la felicidad; sin embargo, en los datos de observación, los problemas de identificación siguen siendo graves. Este estudio explota el hecho de que, en la India, a las personas se les asigna una casta desde que nacen. Dos encuestas similares de jefes de hogar (cada uno con N=1000) en las zonas rurales de Punjab y Andhra Pradesh muestran un patrón creciente de bienestar económico con jerarquía de castas. Esto ilustra que, en las regiones rurales estudiadas, la casta sigue siendo un determinante importante de las oportunidades en la vida. Adicional, las castas superiores están claramente más satisfechas que las castas inferiores y medias. Este resultado, que está en línea con las predicciones de las principales teorías de comparación social, es sólido en los dos estudios de caso. Sin embargo, el patrón a través de las castas baja y media es menos claro, reflejando la compleja relación teórica entre ser de rango medio, por un lado, y el comportamiento, las aspiraciones y el bienestar, por el otro. En la muestra de	Recogidos y analizados desde los años 2008 al 2010.

			Punjab, incluso encontramos un patrón significativo en forma de U, siendo las castas medias las menos felices. Las sociedades de castas pueden verse como sociedad de clase en la que se adquiere clase social con el nacimiento. Los sistemas de castas rechazan las relaciones estrechas con miembros de otras castas. Esa "pureza" de castas suele mantenerse mediante regla de endogamia, el matrimonio debe ser entre personas del mismo grupo social. La influencia del estatus social en la felicidad de las personas es un tema importante, que se refleja en la atención que ha estado recibiendo de investigadores de diferentes disciplinas. En primer lugar, este interés puede estar motivado por una auténtica preocupación política por la felicidad de las personas y la consiguiente necesidad de explorar sus determinantes. En segundo lugar, dado que las investigaciones muestran que las personas intentan en general maximizar su felicidad, comprender cómo la posición relativa se relaciona con la felicidad es un paso importante para predecir el comportamiento humano. Los estudios que han podido demostrar una relación de causa y efecto entre el estatus social y la felicidad generalmente concluyen que la felicidad, la satisfacción en el trabajo y otras variaciones de la satisfacción aumentan con el estatus social.	

Luego de lo detallado anteriormente he seleccionado trabajar en el proyecto basándome principalmente en los artículos 4, 5 y 2 en ese orden respectivamente según su relación existente, por las razones principales que son:

- Según los datasets de los años proporcionados para este proyecto, es decir; que estos artículos tienen los datos más actualizados que se han revisado y analizado.
- En todos ellos se relacionan seis variables clave que se ha comprobado que apoyan el bienestar, estas son: PIB o ingresos, esperanza de vida de salud, ayuda social, libertad en sociedad, confianza y generosidad. La suma de estas condiciones determina la nota global para cada país.
- Se consideraron tres tipos de resultados de felicidad: evaluaciones de vida, afecto positivo (experiencias de disfrute, felicidad y risa) y afecto negativo (experiencias de preocupación, tristeza y enojo).

2.3 MODELOS DE MINERÍA DE DATOS

2.3.1 DEFINICIÓN

Un modelo de minería de datos es un conjunto de datos, compuestos por estadísticas y patrones que se usan para descubrir patrones de grandes volúmenes de datos.

La minería de datos en la actualidad, tiene muchas aplicaciones comerciales, de marketing, industria, etc. Cuando se cuenta con gran cantidad de datos, hay que limpiarlos y organizarlos; para luego si decir que contamos con información.

A la información hay que tratarla con un modelo para así lograr obtener resultados o conclusiones a los que se les denomina "conocimiento". En otras palabras, el conocimiento es la información analizada.

Existen diferentes modelos de minería para realizar el análisis de los datos. Para crear un modelo de minería de datos, se recomienda seguir estos pasos:

- Cree la estructura de minería de datos subyacente e incluya las columnas de datos que sean necesarias.
- Seleccione el algoritmo más adecuado para la tarea analítica.
- Elija la estructura de columnas para usar en el modelo.
- Opcionalmente, puede establecer los parámetros para ajustar el procesamiento del algoritmo.

- Rellene el modelo con datos procesando la estructura correspondiente [25].

"*En la minería de datos se incluyen los siguientes tipos de algoritmos:*

- ✓ ***Algoritmos de clasificación**. - Predicen una o más variables discretas, basándose en los demás atributos del conjunto de datos.*
- ✓ ***Algoritmos de regresión**. - Predicen una o más variables numéricas continuas, como pérdidas o ganancias, basándose en otros atributos del conjunto de datos.*
- ✓ ***Algoritmos de segmentación**. - Dividen los datos en grupos, o clústeres, de elementos que tienen propiedades similares.*
- ✓ ***Algoritmos de asociación**. - Buscan correlaciones entre diferentes atributos de un conjunto de datos. La aplicación más común de esta clase de algoritmo es la creación de reglas de asociación que pueden usarse en un análisis de la cesta de compra.*
- ✓ ***Algoritmos de análisis de secuencias. -** Resumen las secuencias frecuentes o episodios en los datos, como una serie de clics en un sitio web o una serie de eventos de registro que preceden al mantenimiento del equipo*". [26]

2.3.2 PROCESAMIENTO DE MODELOS DE MINERÍA DE DATOS

Al procesar un modelo, los datos que lo estructuran se almacenan en memoria caché y el algoritmo los analiza. Luego, "*el algoritmo calcula un conjunto de estadísticas de resumen que describen los datos, identifica las reglas y los patrones en los datos, y después usa dichas reglas y patrones para completar el modelo.*

Una vez procesado el modelo de minería de datos, este contiene gran cantidad de información, y los patrones encontrados mediante el análisis, incluyen estadísticas, reglas y fórmulas de regresión. Se puede usar los visores personalizados para examinar esta información o se puede crear consultas de minería de datos para recuperarla y usarla para el análisis y la presentación" [25].

2.3.3 ¿CÓMO ESCOGER UN MODELO DE MINERÍA DE DATOS?

"*No hay un modelo óptimo de tratamiento de datos. Por lo tanto, el modelo a elegir depende de las circunstancias y necesidades. Factores a tener en cuenta son la efectividad del modelo para dar resultados de calidad, y el si resulta necesario o no que sea comprensible para el ser humano.*

En el caso de escoger una red neuronal, las operaciones que se aplican a los datos hay que determinarlas. ¿Cómo se hace esto? Digamos que "entrenando" a la red neuronal (a esto se le llama machine learning o aprendizaje automático) a través de algoritmos de optimización de forma que, dados unos datos de entrada, vamos informando al sistema de si el resultado es más o menos bueno. En sucesivas iteraciones, el sistema puede alcanzar un grado de perfeccionamiento adecuado para su explotación comercial" [27].

2.3.4 LAS BASES DE DATOS Y LA MINERÍA DE DATOS

"*Las bases de datos han sido sin duda una herramienta fundamental que ha permitido la evolución de la ciencia de la minería de datos. De hecho, a veces se usa el término "KDD (Knowledge Discovery in Databases o Descubrimiento de Conocimiento en Bases de Datos) como sinónimo de minería de datos.*

Una base de datos se puede decir que se constituye en una de las tres patas en que se apoya la minería de datos, y que son: 1. Bases de datos 2. Estadística y 3. Algoritmia" [27]. A continuación, en la Ilustración 4 se presenta los elementos que forman parte de la minería de datos.

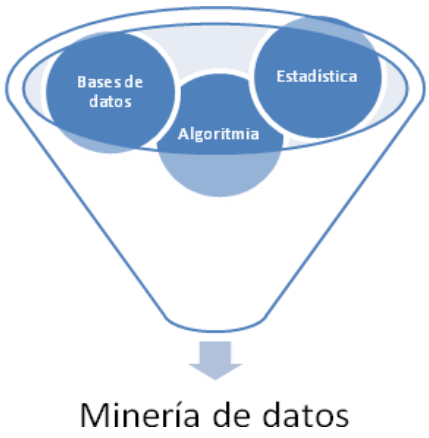


Ilustración 4. Minería de datos [27].

2.3.5 MINERÍA DE DATOS VISUAL

"*Una aplicación curiosa de la minería de datos es obtener imágenes representativas para realizar el análisis de datos. Esto permite mostrar lo que ocurre con miles de datos de forma gráfica*" [27]. En la Ilustración 5, se aprecia una figura de lo que es minería de datos visual.

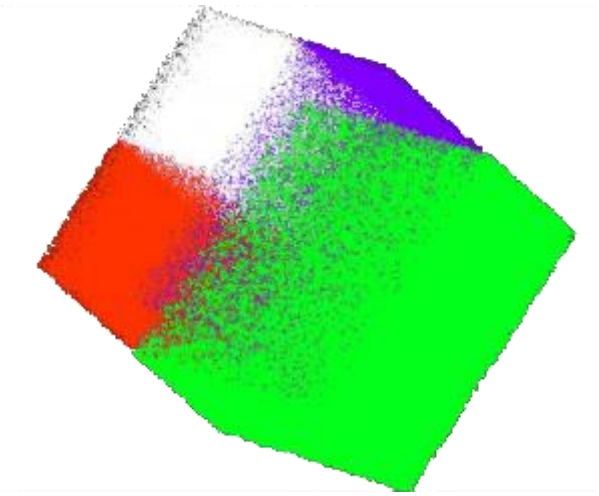


Ilustración 5. Minería de datos visual [27].

CAPÍTULO 3. METODOLOGÍA DE ESTUDIO Y HERRAMIENTAS

3.1 POBLACIÓN DE ESTUDIO

Para el proyecto he utilizado 3 datasets de los años 2015, 2016 y 2017, en donde a través de encuestas, se ha recogido la información que de detalla:

Dataset 2015.csv: Compuesto por 158 filas y 12 columnas. El detalle de las columnas es:

Nro.	Campo	Descripción
1	Country	Nombre del país
2	Region	Region a la que pertenece el país
3	Happiness Rank	Calificación del país basado en el puntaje de felicidad
4	Happiness Score	Métrica medida en 2015 mediante la formulación de la pregunta a las personas incluídas en la muestra: "¿Cómo calificarías tu felicidad en una escala del 0 al 10, donde 10 es el más feliz?"
5	Standard Error	El error estándar de la puntuación de felicidad
6	Economy (GDP per Capita)	La medida en que el PIB contribuye al cálculo de la puntuación de felicidad
7	Family	La medida en que la familia contribuye al cálculo de la puntuación de felicidad
8	Health (Life Expectancy)	La medida en que la esperanza de vida contribuyó al cálculo de la puntuación de felicidad
9	Freedom	La medida en que la libertad contribuyó al cálculo de la puntuación de felicidad
10	Trust (Government Corruption)	La medida en que la percepción de la corrupción contribuye a la puntuación de felicidad
11	Generosity	La medida en que la generosidad contribuyó al cálculo de la puntuación de felicidad
12	Dystopia Residual	La medida en que Dystopia Residual contribuyó al cálculo de la puntuación de felicidad

Dataset 2016.csv: Compuesto por 157 filas y 13 columnas. El detalle de las columnas es:

Nro.	Campo	Descripción
1	Country	Nombre del pais
2	Region	Region a la que pertenece el pais
3	Happiness Rank	Calificación del país basado en el puntaje de felicidad
4	Happiness Score	Métrica medida en 2016 mediante la formulación de la pregunta a las personas incluidas en la muestra: "¿Cómo calificarías tu felicidad en una escala del 0 al 10, donde 10 es el más feliz?"

5	Lower Confidence Interval	Menor intervalo de confianza de la puntuación de felicidad
6	Upper Confidence Interval	Intervalo de confianza superior de la puntuación de felicidad
7	Economy (GDP per Capita)	La medida en que el PIB contribuye al cálculo de la puntuación de felicidad
8	Family	La medida en que la familia contribuye al cálculo de la puntuación de felicidad
9	Health (Life Expectancy)	La medida en que la esperanza de vida contribuyó al cálculo de la puntuación de felicidad
10	Freedom	La medida en que la libertad contribuyó al cálculo de la puntuación de felicidad
11	Trust (Government Corruption)	La medida en que la percepción de la corrupción contribuye a la puntuación de felicidad
12	Generosity	La medida en que la generosidad contribuyó al cálculo de la puntuación de felicidad
13	Dystopia Residual	La medida en que Dystopia Residual contribuyó al cálculo de la puntuación de felicidad

Dataset 2017.csv: Compuesto por 155 filas y 12 columnas. El detalle de las columnas es:

Nro.	Campo	Descripción
1	Country	Nombre del pais
2	Happiness Rank	Calificación del país basado en el puntaje de felicidad
3	Happiness Score	Métrica medida en 2017 mediante la formulación de la pregunta a las personas incluidas en la muestra: "¿Cómo calificarías tu felicidad en una escala del 0 al 10, donde 10 es el más feliz?"
4	Whisker.high	Margen alto
5	Whisker.low	Margen bajo
6	Economy (GDP per Capita)	La medida en que el PIB contribuye al cálculo de la puntuación de felicidad
7	Family	La medida en que la familia contribuye al cálculo de la puntuación de felicidad
8	Health (Life Expectancy)	La medida en que la esperanza de vida contribuyó al cálculo de la puntuación de felicidad
9	Freedom	La medida en que la libertad contribuyó al cálculo de la puntuación de felicidad
10	Generosity	La medida en que la generosidad contribuyó al cálculo de la puntuación de felicidad
11	Trust (Government Corruption)	La medida en que la percepción de la corrupción contribuye a la puntuación de felicidad
12	Dystopia Residual	La medida en que Dystopia Residual contribuyó al cálculo de la puntuación de felicidad

A continuación, se presenta la cantidad de países por región y por año que se incluyen en cada uno de los datasets con los que se trabajó en el presente proyecto.

Nro.	Region	2015	2016	2017
1	Australia and New Zealand	2	2	2
2	Central and Eastern Europe	29	29	29
3	Eastern Asia	6	6	6
4	Latin America and Caribbean	22	24	22
5	Middle East and Northern Africa	20	19	19
6	North America	2	2	2
7	Southeastern Asia	9	9	8
8	Southern Asia	7	7	7
9	Sub-Saharan Africa	40	38	39
10	Western Europe	21	21	21
Total general		**158**	**157**	**155**

3.2 TÉCNICAS E INSTRUMENTOS DE RECOLECCIÓN DE DATOS

3.2.1 CRITERIOS PARA LA CONSTRUCCIÓN Y ELABORACIÓN DE LAS TÉCNICAS DE RECOLECCIÓN

Para construir y elaborar una técnica de recolección de datos se debe establecer varios criterios entre los cuales se encuentran:

- ✓ "*La naturaleza del objeto de estudio.*
- ✓ *Las posibilidades de acceso con los investigados.*
- ✓ *El tamaño de la población o muestra.*
- ✓ *Los recursos con los que se cuenta.*
- ✓ *La oportunidad de obtener datos.*
- ✓ *Tipo y naturaleza de la fuente de datos*" [28].

Entre las principales técnicas de recolección de datos están:

- ✓ Encuesta
- ✓ Entrevista
- ✓ Análisis documental
- ✓ Observación no experimental
- ✓ Observación experimental.

En el presente proyecto se utilizó 2 técnicas principales que son:

a) "Encuesta: *Con esta técnica de recolección de datos se dió lugar a establecer un contacto con las unidades de observación por medio de los cuestionarios previos*

establecidos. Entre las modalidades de encuesta existentes en la actualidad se puede destacar:

- *Encuestas por teléfono*
- *Encuestas por correo*
- *Encuesta personal*
- *Encuesta online.*

b) *Análisis documental:* *Mediante el análisis documental se recolectan datos de fuentes secundarias como son: Libros, boletines, revistas, folletos, y periódicos; es decir, se utilizan como fuentes para recolectar datos sobre las variables de interés. El instrumento que se acostumbra utilizar es la ficha de registro de datos*" [28].

Todas las variables que se usaron para la recolección de datos son cuantitativas a excepción de las variables Country y Region que son cualitativas.

En las siguientes gráficas (Ilustración 6 y 7) se muestra un diagrama en bloques de las entradas que se necesitan para establecer las conclusiones y los pasos que se deben considerar en cuanto al tratamiento de los datos.

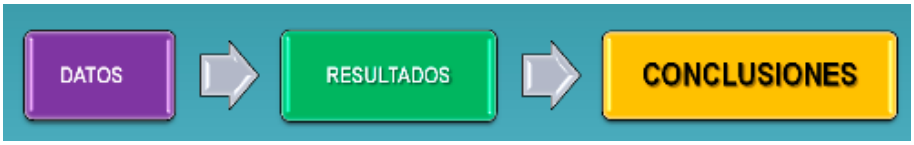


Ilustración 6. Diagrama de bloques[1].

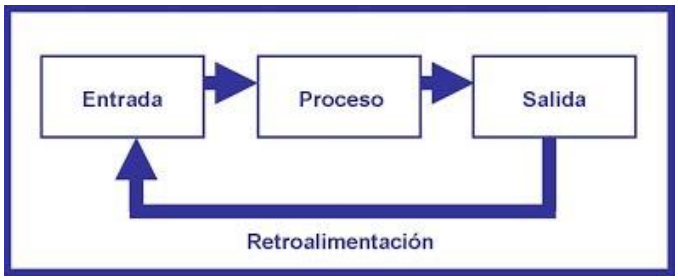


Ilustración 7. Pasos en el tratamiento de la información[2].

[1] Diagrama de bloques: http://cort.as/-JZYs

[2] Fases en el tratamiento de la información: http://cort.as/-Jb2x

3.3 HERRAMIENTAS Y LENGUAJES DE PROGRAMACIÓN

Para el desarrollo del análisis estadístico según los datasets, se utilizó un modelo de regresión múltiple y la implementación se la ejecutó en el lenguaje de programación R.

3.3.1 DEFINICIÓN DEL LENGUAJE R

"*R es un lenguaje con licencia GNU**,** es decir; es libre, gratuito y abierto.*
R funciona con paquetes gratuitos, como las librerías en otros lenguajes, y también se puede descargar y usar dichos paquetes" [29].

3.3.2 CARACTERÍSTICAS DEL LENGUAJE R

Algunas de las características principales del lenguaje R son:

- "*Es uno de los lenguajes más usados en Minería de Datos.*
- *Es gratuito.*
- *Cuenta con una interfaz amigable para el desarrollo.*
- *Posibilidad de crear gráficos, basado en Latex.*
- *Gran cantidad de herramientas estadísticas:* modelos lineales y no lineales*, t*ests estadísticos *y algoritmos de* clasificación *y* agrupamiento*.*
- *Posibilidad de crear tus propias funciones, además de objetos al ser su programación POO (orientada a objetos).*
- *Integración con distintas bases de datos.*
- *Puede tener un uso matemático, como sustitución a MATLAB*" [29].

3.3.3 CARACTERÍSTICAS DE R STUDIO

"*R Studio es un IDE de R.*

*Un Entorno de Desarrollo Integrado (IDE) proporciona todas las herramientas necesarias para poder programar en el lenguaje R**.***

Al instalar R Studio, se puede programar en R directamente en una consola, muy parecida al Símbolo del Sistema de Windows tal y como se aprecia en la Ilustración 8" [29].

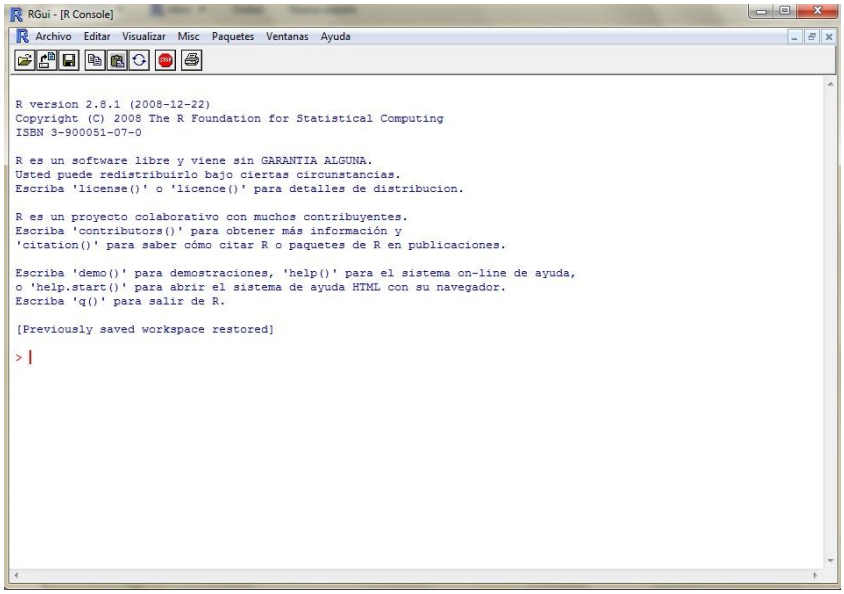


Ilustración 8. Entorno R [29].

"*Con R Studio existen muchas opciones, como la posibilidad de crear gráficos. R Studio también es gratuito, así que es la mejor opción para poder escribir en este lenguaje.*

A continuación, se muestra en la Ilustración 9 el entorno de R Studio" [29].

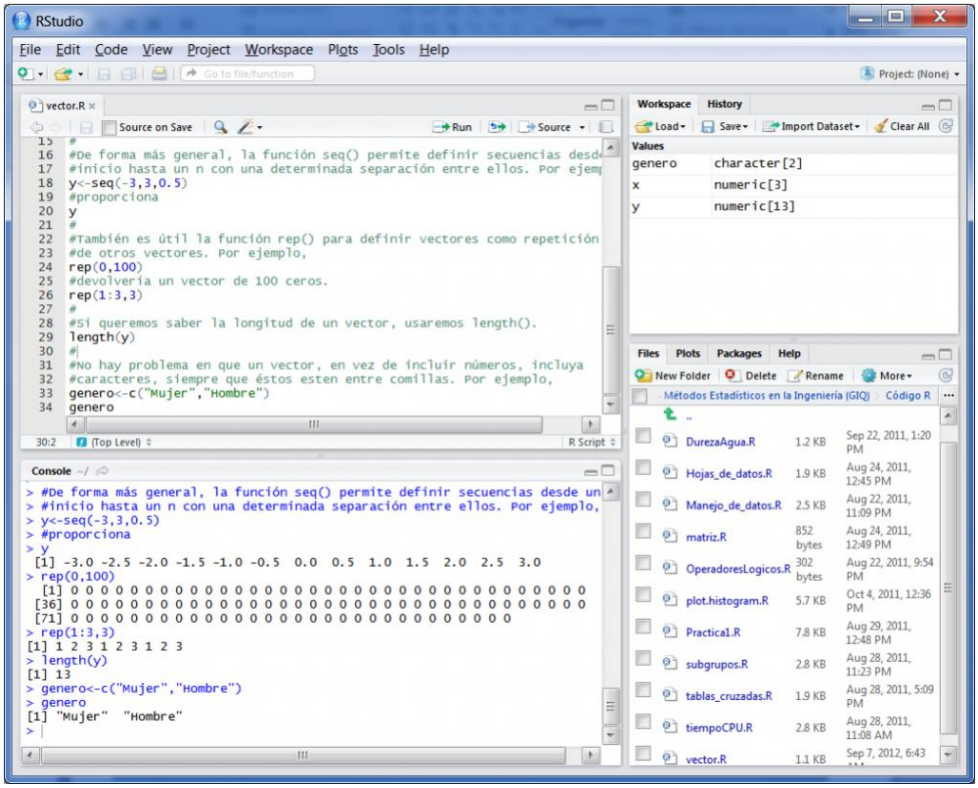


Ilustración 9. Entorno R Studio [29].

3.3.4 CARACTERÍSTICAS DE R COMMANDER

"*Con* R Commander *se puede usar la mayoría de análisis estadísticos más comunes.*

Se puede instalar R Commander desde el propio R Studio, como un paquete más y utilizar la gran cantidad de opciones de menú para poder programar *lo que más nos interesa.*

A continuación, se muestra en la Ilustración 10 el entorno de R Commander" [29].

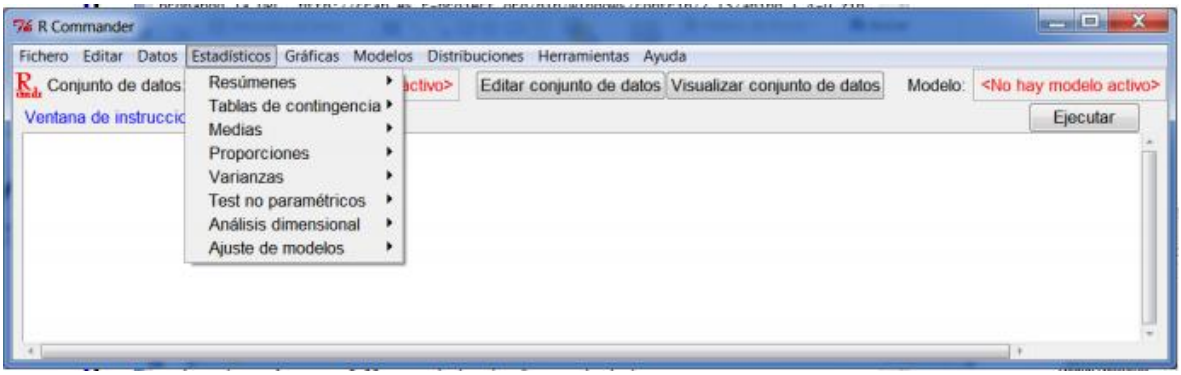


Ilustración 10. Entorno R Commander [29].

CAPÍTULO 4. RESULTADOS

4.1 PRESENTACIÓN DEL MODELO DE MINERÍA DE DATOS

El modelo de minería de datos construído para predecir la felicidad a nivel mundial en el presente trabajo de investigación, consistió en la integración de los siguientes elementos:

1) La metodología en cascada en la cual la construcción del modelo se ejecutó vigilando el proceso y rectificando o ajustando parámetros para continuar con la siguiente etapa.

2) Análisis del indicador sobre el nivel de felicidad y su desarrollo evolutivo a lo largo de los tres años, para lo cual se realizó una comparación detallada sobre si hay diferencias de este indicador entre las diferentes regiones del mundo. Se aplicó un análisis de varianza.

3) A continuación, se aplicó un modelo de regresión múltiple, para lo cual se consideró como variable dependiente el nivel de felicidad (HS) y como variables explicativas el resto de variables que intervinieron en el conjunto de datos.

4) Aplicar ANOVA, es decir, aplicar un algoritmo de minería de datos (k-NN, que es el algortimo de vecinos más cercanos) para un problema de clasificación. A partir del ANOVA, se investigó si existen diferencias significativas entre las diferentes configuraciones del algoritmo; y en caso afirmativo, con cuál de las versiones es la que obtiene los mejores resultados.

4.2 PRESENTACIÓN, ANÁLISIS DE DATOS Y DISCUSIÓN DE LOS RESULTADOS.

PROCEDIMIENTO

En forma general, para contestar las interrogantes planteadas al inicio del proyecto, se ejecutan los siguientes pasos:

1. Obtención de los datos.
2. Realización de un análisis descriptivo de los datos.
3. Comprobación de los supuestos (normalidad y presencia de outliers).
4. Interpretación de los resultados.

El paso inicial para representar y analizar datos estadísticos, es la ejecución del preprocesado.

PREPROCESADO

Dentro de esta etapa se ha considerado las siguientes 5 subetapas que son:

- Carga de ficheros
- Actualización de nombres de columnas
- Validación de conjunto de datos
- Tipos de datos de variables y conversiones
- Abreviación de nombres de las regiones.

A continuación, se describen cada una de ellas.

- **Carga de ficheros**

  Carga de tres ficheros, en variables de tipo data frame, las cuales son: *datos2015*, *datos2016 y datos2017*. Por cada uno se valida la carga.

```
datos2015<-read.csv("2015.csv", sep=",",na.strings = "NA", dec = ".")
datos2016<-read.csv("2016.csv", sep=",",na.strings = "NA", dec = ".")
datos2017<-read.csv("2017.csv", sep=",",na.strings = "NA", dec = ".")
#Revisión de datos de los primeros registros de los archivos
head(datos2015)
```

```
##       Country         Region Happiness.Rank Happiness.Score Standard.Error
## 1 Switzerland Western Europe              1           7.587        0.03411
## 2     Iceland Western Europe              2           7.561        0.04884
## 3     Denmark Western Europe              3           7.527        0.03328
## 4      Norway Western Europe              4           7.522        0.03880
## 5      Canada  North America              5           7.427        0.03553
## 6     Finland Western Europe              6           7.406        0.03140
##   Economy..GDP.per.Capita.  Family Health..Life.Expectancy. Freedom
## 1                  1.39651 1.34951                  0.94143 0.66557
## 2                  1.30232 1.40223                  0.94784 0.62877
## 3                  1.32548 1.36058                  0.87464 0.64938
## 4                  1.45900 1.33095                  0.88521 0.66973
## 5                  1.32629 1.32261                  0.90563 0.63297
## 6                  1.29025 1.31826                  0.88911 0.64169
##   Trust..Government.Corruption. Generosity Dystopia.Residual
## 1                       0.41978    0.29678           2.51738
## 2                       0.14145    0.43630           2.70201
## 3                       0.48357    0.34139           2.49204
## 4                       0.36503    0.34699           2.46531
## 5                       0.32957    0.45811           2.45176
## 6                       0.41372    0.23351           2.61955
```

```
head(datos2016)
```

```
##       Country         Region Happiness.Rank Happiness.Score
## 1     Denmark Western Europe              1           7.526
## 2 Switzerland Western Europe              2           7.509
## 3     Iceland Western Europe              3           7.501
## 4      Norway Western Europe              4           7.498
## 5     Finland Western Europe              5           7.413
## 6      Canada  North America              6           7.404
##   Lower.Confidence.Interval Upper.Confidence.Interval
## 1                     7.460                     7.592
## 2                     7.428                     7.590
## 3                     7.333                     7.669
## 4                     7.421                     7.575
## 5                     7.351                     7.475
## 6                     7.335                     7.473
##   Economy..GDP.per.Capita.  Family Health..Life.Expectancy. Freedom
## 1                  1.44178 1.16374                  0.79504 0.57941
## 2                  1.52733 1.14524                  0.86303 0.58557
## 3                  1.42666 1.18326                  0.86733 0.56624
## 4                  1.57744 1.12690                  0.79579 0.59609
## 5                  1.40598 1.13464                  0.81091 0.57104
## 6                  1.44015 1.09610                  0.82760 0.57370
##   Trust..Government.Corruption. Generosity Dystopia.Residual
## 1                       0.44453    0.36171           2.73939
## 2                       0.41203    0.28083           2.69463
## 3                       0.14975    0.47678           2.83137
## 4                       0.35776    0.37895           2.66465
## 5                       0.41004    0.25492           2.82596
## 6                       0.31329    0.44834           2.70485
```

```
head(datos2017)
```

```
##       Country Happiness.Rank Happiness.Score Whisker.high Whisker.low
## 1      Norway              1           7.537     7.594445    7.479556
## 2     Denmark              2           7.522     7.581728    7.462272
## 3     Iceland              3           7.504     7.622030    7.385970
## 4 Switzerland              4           7.494     7.561772    7.426227
## 5     Finland              5           7.469     7.527542    7.410458
## 6 Netherlands              6           7.377     7.427426    7.326574
##   Economy..GDP.per.Capita.   Family Health..Life.Expectancy.   Freedom
## 1                 1.616463 1.533524                0.7966665 0.6354226
## 2                 1.482383 1.551122                0.7925655 0.6260067
## 3                 1.480633 1.610574                0.8335521 0.6271626
## 4                 1.564980 1.516912                0.8581313 0.6200706
## 5                 1.443572 1.540247                0.8091577 0.6179509
## 6                 1.503945 1.428939                0.8106961 0.5853845
##   Generosity Trust..Government.Corruption. Dystopia.Residual
## 1  0.3620122                     0.3159638          2.277027
## 2  0.3552805                     0.4007701          2.313707
## 3  0.4755402                     0.1535266          2.322715
## 4  0.2905493                     0.3670073          2.276716
## 5  0.2454828                     0.3826115          2.430182
## 6  0.4704898                     0.2826618          2.294804
```

- **Actualización de nombres de las columnas**

  Unificar los nombres de las columnas de cada data frame con palabras más cortas para que sea fácilmente manejable. Por ejemplo, en lugar de "*Happiness Score*" usar "*HS*". Realizar el mismo procedimiento para todas las columnas de los 3 archivos. Así:

```
#Modificación o actualización de columnas de datasets
colnames(datos2015)<-c("Country","Region","HR","HS","SE","EPC","Family","LE","Freedom","GC","Generosity","DS")
colnames(datos2016)<-c("Country","Region","HR","HS","LCI","UCI","EPC","Family","LE","Freedom","GC","Generosity","DS")
colnames(datos2017)<-c("Country","HR","HS","WI","WL","EPC","Family","LE","Freedom","Generosity","GC","DS")
```

- **Validación del conjunto de datos**

  Revisar si los tres conjuntos de datos tienen el mismo número de países o filas.

```
str(datos2015$Country)
```

```
##  Factor w/ 158 levels "Afghanistan",..: 136 59 38 106 25 46 100 135 101 7 ...
```

```
str(datos2016$Country)
```

```
##  Factor w/ 157 levels "Afghanistan",..: 38 135 58 104 45 26 98 99 7 134 ...
```

```
str(datos2017$Country)
```

```
##  Factor w/ 155 levels "Afghanistan",..: 105 38 58 133 45 99 26 100 132 7 ...
```

Observación: En el año 2015 se registran 158 países, en el año 2016 se registran 157 países y en el año 2017 se registran 155 países; es decir; que no se está analizando el mismo número de países para los tres años, aunque es mínima la diferencia; por lo tanto, no existe ninguna afectación.

- **Tipos de datos de variables y conversiones**

  Aquí se realitzarà una consulta sobre los diferentes tipos de datos de las variables que intervienen; si es necesario, se debe aplicar las conversiones necesarias.

  Año 2015

```
variables2015 <- sapply(datos2015,class)
kable(data.frame(var=names(variables2015),clase=as.vector(variables2015)))
```

var	clase
Country	factor
Region	factor
HR	integer
HS	numeric
SE	numeric
EPC	numeric
Family	numeric
LE	numeric
Freedom	numeric
GC	numeric
Generosity	numeric
DS	numeric

Año 2016

```
variables2016 <- sapply(datos2016,class)
kable(data.frame(var=names(variables2016),clase=as.vector(variables2016)))
```

var	clase
Country	factor
Region	factor
HR	integer
HS	numeric
LCI	numeric
UCI	numeric
EPC	numeric
Family	numeric
LE	numeric
Freedom	numeric
GC	numeric
Generosity	numeric
DS	numeric

Año 2017

```
variables2017 <- sapply(datos2017,class)
kable(data.frame(var=names(variables2017),clase=as.vector(variables2017)))
```

var	clase
Country	factor
HR	integer
HS	numeric
WI	numeric
WL	numeric
EPC	numeric
Family	numeric
LE	numeric
Freedom	numeric
Generosity	numeric
GC	numeric
DS	numeric

Observación: Como los tipos de variables se encuentran acordes a la información almacenada en cada una de las columnas, entonces, no es necesario realizar ningún tipo de conversión.

- **Abreviación de nombres de regiones**

  Abreviar los nombres de las regiones para trabajar más fácilmente, para ello se utiliza estas abreviaciones:

  - AUSNZ (Australia and New Zealand)
  - MENA (Middle East and Northern Africa)
  - SEA (Southeastern Asia)
  - SA (Southern Asia)
  - EA (Eastern Asia)
  - SSA (Sub-Saharan Africa)
  - WE (Western Europe)
  - ECE (Central and Eastern Europe)
  - LC (Latin America and Caribbean)
  - NAM (North America)

Luego de realizar la respectiva **depuración** y configuración de los registros y columnas de los datasets, se presenta la cantidad de registros existentes por cada una de las regiones de los años 2015 y 2016. En el año 2017, no existe data por regiones.

Los datos del año 2015 son:

```
#Consultar los tipos de regiones y actualizar según las abreviaturas.
#Año 2015
table(datos2015$Region)
```

```
##
##       Australia and New Zealand      Central and Eastern Europe
##                               2                              29
##                    Eastern Asia     Latin America and Caribbean
##                               6                              22
## Middle East and Northern Africa                   North America
##                              20                               2
##               Southeastern Asia                   Southern Asia
##                               9                               7
##              Sub-Saharan Africa                  Western Europe
##                              40                              21
```

```
#Actualización año 2015
datos2015$Region <- factor(datos2015$Region, levels=c("Australia and New Zealand","Middle East and Northern Africa","Southea
stern Asia","Southern Asia","Eastern Asia","Sub-Saharan Africa","Western Europe","Central and Eastern Europe","Latin Americ
a and Caribbean","North America"), labels=c("AUSNZ","MENA","SEA","SA","EA","SSA","WE","ECE","LC","NAM"))
table(datos2015$Region)
```

```
##
## AUSNZ  MENA   SEA    SA    EA   SSA    WE   ECE    LC   NAM
##     2    20     9     7     6    40    21    29    22     2
```

Los datos del año 2016 son:

```
#Año 2016
table(datos2016$Region)
```

```
##
##       Australia and New Zealand      Central and Eastern Europe
##                               2                              29
##                    Eastern Asia     Latin America and Caribbean
##                               6                              24
## Middle East and Northern Africa                   North America
##                              19                               2
##               Southeastern Asia                   Southern Asia
##                               9                               7
##              Sub-Saharan Africa                  Western Europe
##                              38                              21
```

```
#Actualización año 2016
datos2016$Region <- factor(datos2016$Region, levels=c("Australia and New Zealand","Middle East and Northern Africa","Southea
stern Asia","Southern Asia","Eastern Asia","Sub-Saharan Africa","Western Europe","Central and Eastern Europe","Latin Americ
a and Caribbean","North America"), labels=c("AUSNZ","MENA","SEA","SA","EA","SSA","WE","ECE","LC","NAM"))
table(datos2016$Region)
```

```
##
## AUSNZ  MENA   SEA    SA    EA   SSA    WE   ECE    LC   NAM
##     2    19     9     7     6    38    21    29    24     2
```

En el presente proyecto, interesa dar respuesta a **algunas interrogantes sobre la felicidad a nivel mundial**, pero entre las más importantes están:

- **¿EXISTEN DIFERENCIAS DE FELICIDAD EN LOS TRES AÑOS?**
  Para dar respuesta a esta interrogante, se usaron 2 métodos:
  - ANOVA de un factor
  - ANOVA para muestras apareadas.

1. ANOVA DE UN FACTOR (ONE WAY ANOVA)

Para poder determinar si existen diferencias que sean significativas en el nivel de felicidad en los tres años, se sigue los pasos que se describen a continuación:

a) Preparación de los datos

Para poder comparar los datos de cada país a lo largo de los tres años, se debe agrupar la información, para ello considerar:

Paso 1: Generar un data frame que contenga los datos de un país por cada fila o registro. Las columnas que se envían son: *country* y los valores de *HS* de los años 2015, 2016 y 2017 respectivamente para lo cual se utiliza la función merge.

```
pruebas2015 <- data.frame(Country = datos2015$Country,HS2015 = datos2015$HS)
pruebas2016 <- data.frame(Country = datos2016$Country,HS2016 = datos2016$HS)
pruebas2017 <- data.frame(Country = datos2017$Country,HS2017 = datos2017$HS)
merge_1 <- merge (x = pruebas2015, y = pruebas2016, by = "Country", all = TRUE)
Paises_1 <- merge (x = merge_1, y = pruebas2017, by = "Country", all = TRUE)
str(Paises_1)
```

```
## 'data.frame':    166 obs. of  4 variables:
##  $ Country: Factor w/ 166 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ HS2015 : num  3.58 4.96 5.61 4.03 6.57 ...
##  $ HS2016 : num  3.36 4.66 6.36 3.87 6.65 ...
##  $ HS2017 : num  3.79 4.64 5.87 3.8 6.6 ...
```

```
head(Paises_1,20)
```

```
##                   Country HS2015 HS2016 HS2017
## 1             Afghanistan  3.575  3.360  3.794
## 2                 Albania  4.959  4.655  4.644
## 3                 Algeria  5.605  6.355  5.872
## 4                  Angola  4.033  3.866  3.795
## 5               Argentina  6.574  6.650  6.599
## 6                 Armenia  4.350  4.360  4.376
## 7               Australia  7.284  7.313  7.284
## 8                 Austria  7.200  7.119  7.006
## 9              Azerbaijan  5.212  5.291  5.234
## 10                Bahrain  5.960  6.218  6.087
## 11             Bangladesh  4.694  4.643  4.608
## 12                Belarus  5.813  5.802  5.569
## 13                Belgium  6.937  6.929  6.891
## 14                  Benin  3.340  3.484  3.657
## 15                 Bhutan  5.253  5.196  5.011
## 16                Bolivia  5.890  5.822  5.823
## 17 Bosnia and Herzegovina  4.949  5.163  5.182
## 18               Botswana  4.332  3.974  3.766
## 19                 Brazil  6.983  6.952  6.635
## 20               Bulgaria  4.218  4.217  4.714
```

Paso 2: Transformar el data frame anterior en otro data frame que contenga en cada fila la siguiente información: *country, group, HS*. En donde en la variable "*country*" se almacena el país de donde proviene la muestra. En la variable "*group*", hay que almacenar el grupo al que pertenece el dato (2015, 2016, 2017). La variable *HS* es el nivel de felicidad del país en aquel año correspondiente.

```
pruebas2_2015 <- data.frame(Country = datos2015$Country,HS = datos2015$HS, Group = '2015')
pruebas2_2016 <- data.frame(Country = datos2016$Country,HS = datos2016$HS, Group = '2016')
pruebas2_2017 <- data.frame(Country = datos2017$Country,HS = datos2017$HS, Group = '2017')
merge_2 <- merge (x = pruebas2_2015, y = pruebas2_2016, all = TRUE)
Paises_2 <- merge (x = merge_2, y = pruebas2_2017, all = TRUE)
str (Paises_2)
```

```
## 'data.frame':    470 obs. of  3 variables:
##  $ Country: Factor w/ 166 levels "Afghanistan",..: 1 1 1 2 2 2 3 3 3 4 ...
##  $ HS     : num  3.36 3.58 3.79 4.64 4.66 ...
##  $ Group  : Factor w/ 3 levels "2015","2016",..: 2 1 3 3 2 1 1 3 2 3 ...
```

b) Presentación de datos con diagrames de caja: Con las distribuciones del nivel de felicidad en los tres años. A continuación, se presenta en la Ilustración 11, los datos en un boxplot sobre las distribuciones del nivel de felicidad (happiness score) en los tres años.

```
grupos2015 <- subset (Paises_2$HS, Paises_2$Group == "2015")
grupos2016 <- subset (Paises_2$HS, Paises_2$Group == "2016")
grupos2017 <- subset (Paises_2$HS, Paises_2$Group == "2017")
boxplot(Paises_2$HS ~ Paises_2$Group, main="Distribución del nivel de felicidad en tres años", names = c("2015","2016","201
7"))
```

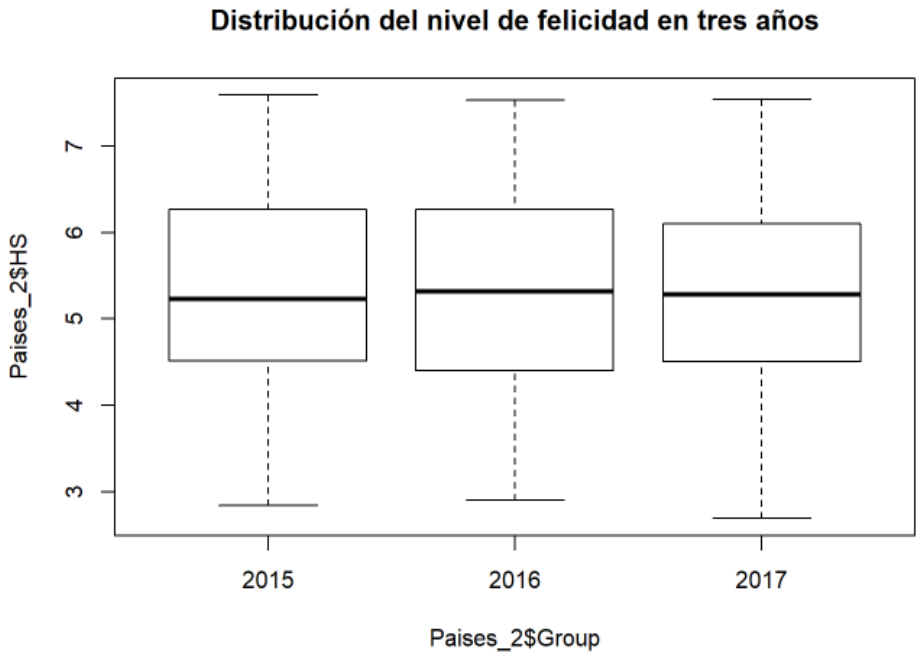


Ilustración 11. Distribución de la felicidad en los tres años.

c) Cálculo ANOVA de un factor: Sirve para investigar si existe diferencias en el nivel de felicidad entre los tres años. Así:

```
HSaov <- aov( Paises_2$HS~Paises_2$Group, data=Paises_2 )
HSaov
```

```
## Call:
##    aov(formula = Paises_2$HS ~ Paises_2$Group, data = Paises_2)
##
## Terms:
##                 Paises_2$Group Residuals
## Sum of Squares          0.0678  606.2388
## Deg. of Freedom              2       467
##
## Residual standard error: 1.139366
## Estimated effects may be unbalanced
```

```
summary (aov(Paises_2$HS~Paises_2$Group))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Paises_2$Group   2    0.1  0.0339   0.026  0.974
## Residuals      467  606.2  1.2982
```

```
numSummary(Paises_2$HS, groups=Paises_2$Group, statistics=c("mean","sd"))
```

```
##          mean       sd data:n
## 2015 5.375734 1.145010    158
## 2016 5.382185 1.141674    157
## 2017 5.354019 1.131230    155
```

Según los datos obtenidos, si existe diferencias en cuanto a los niveles de felicidad en los últimos tres años. Pese a que los valores son casi similares en cuanto al cálculo de la media, se puede decir que en el año 2016 hay mayor nivel de felicidad que en el resto de años.

d) Identificación de las variables: *SST* (suma total de cuadrados), *SSW* (con la suma de cuadrados), *SSB* (entre la suma de cuadrados) y los grados de libertad. Con estos valores, se procede al cálculo manual del valor *F*, el valor crítico (con un nivel de confianza = 95%), y el valor *p*. Interpretar los resultados.

```
variablegrupos2015 <- sum ((grupos2015 - mean(grupos2015))^2)
variablegrupos2016 <- sum ((grupos2016 - mean(grupos2016))^2)
variablegrupos2017 <- sum ((grupos2017 - mean(grupos2017))^2)
#SSW (Within Sum of Squares)
SSW <- variablegrupos2015 + variablegrupos2016 + variablegrupos2017
SSW
```

```
## [1] 606.2388
```

```
valor <- c(grupos2015,grupos2016,grupos2017)
SST <- sum ((valor - mean(valor))^2)
#SST (Total Sum of Squares)
SST
```

```
## [1] 606.3066
```

```
SSB <- SST - SSW
#SSB (Between Sum of Squares)
SSB
```

```
## [1] 0.06783971
```

```
k <-3  #número de grupos
N <-length(valor)
N
```

```
## [1] 470
```

```
#Grados de Libertad
(k-1)*(N-1)
```

```
## [1] 938
```

```
#valor F
F <- (SSB / (k-1)) / (SSW/ (N-k))
F
```

```
## [1] 0.02612926
```

```
#Valor crítico
qf(0.05, k-1, (k-1)*(N-1),lower.tail=FALSE)
```

```
## [1] 3.00532
```

```
#Valor de P
pf(F,k-1, (k-1)*(N-1),lower.tail=FALSE)
```

```
## [1] 0.9742099
```

Observación: Como el valor *p* es superior a 0,05 entonces no se descarta una hipótesis de que los niveles de felicidad sean similares en los últimos tres años con un 95% de confianza. El valor crítico es mayor al valor de *F*, por lo tanto, se concluye que no existen diferencias significativas.

2. ANOVA para muestras apareadas (Repeated Measures ANOVA)

Se han realizado y comprobado cálculos de ANOVA asumiendo muestras independientes, pero de lo que se ha observado las muestras de los tres grupos correspondientes a los tres años están relacionadas.

Esta relación se da ya que se trata de los mismos países medidos en 3 momentos diferentes de tiempo. Por esta razón, es más apropiado usar la prueba "repeated measures ANOVA (ANOVA con medidas repetidas)". En esta prueba la variabilidad entre sujetos (países) se sustrae de la variabilidad dentro de los grupos.

Para calcular ANOVA, se debe ir eliminando la variabilidad entre sujetos, para ello hay que ejecutar los pasos que se indican a continuación:

- **Cálculo de ANOVA**: Se realiza la corrección apropiada y correspondiente a la prueba de ANOVA apareada (Repeated Measures).
  Con la corrección realizada, se obtiene 146 observaciones de 4 variables: el nombre del país (*country*) y los valores de nivel de felicidad (*HS*) en los años 2015, 2016 y 2017 respectivamente. Con los datos obtenidos se define el índice, el mismo que es necesario en los pasos siguientes, es decir:

```
#Corrección de Datos de Paises_1
Paises_3 <- subset (na.omit(Paises_1))
str(Paises_3)
```

```
## 'data.frame':    146 obs. of  4 variables:
##  $ Country: Factor w/ 166 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ HS2015 : num  3.58 4.96 5.61 4.03 6.57 ...
##  $ HS2016 : num  3.36 4.66 6.36 3.87 6.65 ...
##  $ HS2017 : num  3.79 4.64 5.87 3.8 6.6 ...
```

```
length(Paises_3)
```

```
## [1] 4
```

```
valor3<-c(Paises_3$HS2015,Paises_3$HS2016,Paises_3$HS2017) #438
indice<-c(1:length(Paises_3$Country))
indice<-as.factor(indice)
indice
```

```
##   [1] 1   2   3   4   5   6   7   8   9   10  11  12  13  14  15  16  17
##  [18] 18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
##  [35] 35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51
##  [52] 52  53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68
##  [69] 69  70  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85
##  [86] 86  87  88  89  90  91  92  93  94  95  96  97  98  99  100 101 102
## [103] 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119
## [120] 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136
## [137] 137 138 139 140 141 142 143 144 145 146
## 146 Levels: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 ... 146
```

```
#El grupo es 1,2,3, que pertenece a los años 2015, 2016, 2017 respectivamente.
n3<-rep(146,3)
group3 = rep(1:3, n3)
group3<-as.character(group3) #438
paired.data3 <- data.frame( indice=rep(indice,3), group3, valor3 )
#Al calcular ANOVA, se indica el error o variabilidad entre años
Countries3aov <- aov( valor3~group3 + Error(indice/group3), data=paired.data3 )
summary(Countries3aov)
```

```
##
## Error: indice
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 145  560.4   3.864
##
## Error: indice:group3
##           Df Sum Sq  Mean Sq F value Pr(>F)
## group3     2  0.011 0.005599   0.178  0.837
## Residuals 290  9.122 0.031454
```

- **Identificación del término error:** Se obtiene por varios métodos.

```
#SSW (Within Sum of Squares)
SSW
```

```
## [1] 606.2388
```

```
#SST (Total Sum of Squares)
SST
```

```
## [1] 606.3066
```

```
#SSB (Between Sum of Squares)
SSB
```

```
## [1] 0.06783971
```

- **Obtención de la variabilidad:** La cual se obtiene a través de la media.

```
limite <- length(Paises_3$HS2015)
limite
```

```
## [1] 146
```

```
N3 <- length (valor3)
N3
```

```
## [1] 438
```

```
media<-mean(valor3)
media
```

```
## [1] 5.391833
```

```
medidass<-0
for (i in 1:limite){
  mean.indice <- mean( paired.data3[paired.data3$indice==i,]$valor3 )
  medidass <- medidass + (mean.indice - media)^2
}
medidass <- 3*medidass #se multiplica la variabilidad por el número de grupos
medidass
```

```
## [1] 560.3508
```

```
#Una vez se obtiene medidass, se calcula SSW - medidass y el resultado
errorss <- SSW - medidass
#Error Apareado
errorss
```

```
## [1] 45.88805
```

- **Comprobación:** Verificar si el nuevo *SSW* corresponde al *SSW* anterior menos el término de error entre sujetos o países.

```
errorss1 <- medidass + errorss
#Error No Apareado
errorss1
```

```
## [1] 606.2388
```

```
#Si son los mismos, esto se comprueba mediante el error no apareado
```

- **Comparación del valor de *F* de ANOVA** con muestras apareadas según el valor obtenido anteriormente con ANOVA de muestas independientes. Se aplica la fórmula que se aprecia a continuación:

```
F <- (SSB/(k-1) )/ ( errorss/((limite-1)*(k-1)) )
F
```

```
## [1] 0.2143643
```

Mediante la corrección apropiada para la ejecución de Anova apareado, se analiza los valores de *F*, se evidencia la variación con el valor *F* de la primera prueba de Anova, para lo cual se denota que, **en pruebas de Anova existen diferencias significativas con la aplicación de ambos modelos de ANOVA**.

De lo que se observa, las muestras de los tres grupos están relacionadas, dado que se trata de los mismos países medidos en 3 momentos diferentes de tiempo. Por esta razón, es más apropiado usar la prueba "repeated measures ANOVA". En esta prueba la variabilidad entre sujetos se sustrae de la variabilidad dentro de los grupos.

- **¿EXISTEN DIFERENCIAS EN EL NIVEL DE FELICIDAD ENTRE LAS DISTINTAS REGIONES?**

  Cálculo ANOVA

  Para dar respuesta a esta pregunta, se realiza un cálculo sobre los datos del año 2015. A continuación, se describen los pasos a seguir:

  a) **Preparación del data frame** para aplicar ANOVA. En este caso, el data frame debe contener la variable "*region*" como un factor, esta es la variable independiente, y la variable "*HS*" que es de tipo numérico, es la variable dependiente. Cada fila representa una región y un valor concreto de *HS* para la región específica.

```
datset2015 <- data.frame(datos2015$Region, datos2015$HS)
str(datset2015)
```

```
## 'data.frame':    158 obs. of  2 variables:
##  $ datos2015.Region: Factor w/ 10 levels "AUSNZ","MENA",..: 7 7 7 7 10 7 7 7 1 1 ...
##  $ datos2015.HS    : num  7.59 7.56 7.53 7.52 7.43 ...
```

  b) **Presentación de un diagrama de cajas** en la Ilustración 12 con la distribución del *HS* en las distintas regiones del conjunto de datos.

```
boxplot(datos2015$HS ~ datos2015$Region, plot=TRUE,main="Distribución del Happiness Score en Regiones - 2015")
```

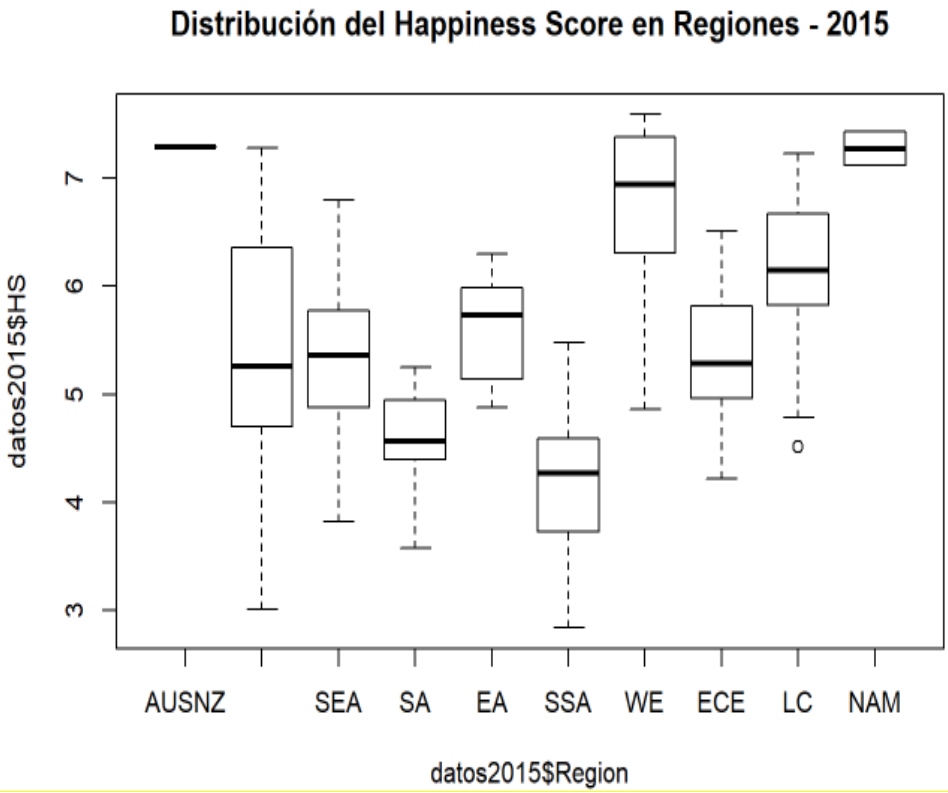


Ilustración 12. Distribución de felicidad en regiones, año 2015.

c) ¿Qué método es el más apropiado para este caso de estudio: ANOVA para muestras independientes o ANOVA con muestas apareadas?

Según revisiones realizadas de los datos con split en R, se evidencia que existe diferencias en el número de muestras por Región en cuanto al análisis de su *HS*, por lo tanto; es necesario realizar un estudio ANOVA para muestras independientes.

d) Cálculo de ANOVA entre todas las regiones, según el método seleccionado.

```
HSaov <- aov( datos2015$HS ~ datos2015$Region )
HSaov
```

```
## Call:
##    aov(formula = datos2015$HS ~ datos2015$Region)
##
## Terms:
##                 datos2015$Region Residuals
## Sum of Squares         123.68339  82.15118
## Deg. of Freedom                9       148
##
## Residual standard error: 0.7450339
## Estimated effects may be unbalanced
```

```
summary (HSaov)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## datos2015$Region   9 123.68  13.743   24.76 <2e-16 ***
## Residuals        148  82.15   0.555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
numSummary(datos2015$HS, groups=datos2015$Region, statistics=c("mean","sd"))
```

```
##           mean          sd data:n
## AUSNZ 7.285000 0.001414214      2
## MENA  5.406900 1.101381902     20
## SEA   5.317444 0.950020146      9
## SA    4.580857 0.570526490      7
## EA    5.626167 0.554052855      6
## SSA   4.202800 0.609557099     40
## WE    6.689619 0.824581802     21
## ECE   5.332931 0.570445811     29
## LC    6.144682 0.728560053     22
## NAM   7.273000 0.217788889      2
```

```
#Tabla de instrucción ANOVA
model.tables(HSaov)
```

```
## Tables of effects
##
##  datos2015$Region
##     AUSNZ     MENA      SEA      SA     EA    SSA     WE     ECE      LC
##     1.909  0.03117 -0.05829 -0.7949 0.2504 -1.173  1.314 -0.0428  0.7689
## rep 2.000 20.00000  9.00000  7.0000 6.0000 40.000 21.000 29.0000 22.0000
##       NAM
##     1.897
## rep 2.000
```

Según los datos graficados en el boxplot, el resultado de *P-Valor* con ANOVA (2e-16) y los datos obtenidos mediante el modelado de tablas ANOVA, se evidencia que no existen relaciones entre cada una de las regiones. Se puede evidenciar que entre Australia and New Zealand (AUSNZ) y North América (NAM) existe una ligera relación en sus varianzas, pero el valor P es muy pequeño por lo que se

puede concluir la no relación de *HS* en las regiones analizadas (Rechazo a hipótesis nula).

PRUEBAS A POSTERIORI (POST-HOC TESTS)

Estas pruebas se usan en razón de que se han comprobado diferencias significativas de felicidad entre las distintas regiones, las cuales se han identificado por el análisis ANOVA en el data frame *datos2015*. Para realizar estas pruebas, se trabajó con el fichero 2015.

Se ha realizado ya en las secciones anteriores, el cálculo del análisis de varianza en 2 grupos de muestras que son: la evolución del *HS* a lo largo de los años y la comparativa entre las diferentes regiones. Se especifica que si se puede rechazar la hipótesis nula, ya que hay evidencias para decir que al menos dos grupos tienen medias diferentes, pero el ANOVA no determina qué grupos tienen medias diferentes. A continuación, hay que realizar el cálculo para identificar qué grupos son diferentes, este cálculo sólo se hace en el caso de que se haya rechazado la hipótesis nula.

En primer lugar, independientemente del resultado obtenido en los puntos anteriores, en el caso de que haya diferencias significativas identificadas por el análisis ANOVA en el data frame *datos2015* en comparación entre las diferentes regiones. Cuando se observan diferencias significativas, se procede a aplicar pruebas a posteriori.

Luego, se realiza una comparación múltiple por cada par de regiones, para determinar cuáles son las regiones que presentan diferencias en el nivel de felicidad. Esta prueba se llama “pairwise comparison” "y se basa en aplicar el *test t* para cada par de grupos (regiones). La aplicación de este *test t* tiene un estadístico similar al caso del test para dos muestras independientes con varianza igual. Una de las suposiciones del método ANOVA es que todos los grupos tienen varianza igual. Para estimar la varianza de la muestra, se usa la varianza global estimada a partir de todos los datos de la muestra.

Los pasos a seguir para ejecutar esta prueba son:

Cálculo ANOVA

a) Identificación de *SSW* (Within Sum of Squares) usando el cálculo ANOVA sobre las regiones del data frame 2015. Calcular la varianza de toda la muestra según la fórmula: *var=SSW/(n-k)*.

```
#SSW (Within Sum of Squares)
AUSNZ <- subset(datos2015$HS, datos2015$Region == "AUSNZ")
MENA <- subset(datos2015$HS, datos2015$Region == "MENA")
SEA <- subset(datos2015$HS, datos2015$Region == "SEA")
SA <- subset(datos2015$HS, datos2015$Region == "SA")
EA <- subset(datos2015$HS, datos2015$Region == "EA")
SSA <- subset(datos2015$HS, datos2015$Region == "SSA")
WE <- subset(datos2015$HS, datos2015$Region == "WE")
ECE <- subset(datos2015$HS, datos2015$Region == "ECE")
LC <- subset(datos2015$HS, datos2015$Region == "LC")
NAM <- subset(datos2015$HS, datos2015$Region == "NAM")
#-----------------------------------------------
valor1 <- c(AUSNZ, MENA, SEA, SA, EA, SSA, WE, ECE, LC, NAM)
numgrup <-10  #número de grupos
N1 <-length(valor1)
#-----------------------------------------------
AUSNZ15 <- sum ((AUSNZ - mean(AUSNZ))^2)
MENA15 <- sum ((MENA - mean(MENA))^2)
SEA15 <- sum ((SEA - mean(SEA))^2)
SA15 <- sum ((SA - mean(SA))^2)
EA15 <- sum ((EA - mean(EA))^2)
SSA15 <- sum ((SSA - mean(SSA))^2)
WE15 <- sum ((WE - mean(WE))^2)
ECE15 <- sum ((ECE - mean(ECE))^2)
LC15 <- sum ((LC - mean(LC))^2)
NAM15 <- sum ((NAM - mean(NAM))^2)
#SSW (Within Sum of Squares)
SSW_1 <- AUSNZ15 + MENA15 + SEA15 + SA15 + EA15 + SSA15 + WE15 + ECE15 + LC15 + NAM15
SSW_1
```

```
## [1] 82.15118
```

```
varssw1=SSW_1/(N1-numgrup)
varssw1
```

```
## [1] 0.5550756
```

b) Cálculo del *estadístico t* para un par de muestras de manera ilustrativa. Se trabaja con "ECE" y "SA". El *estadístico t* es equivalente al que corresponde al contraste de dos muestras con varianzas poblacionales conocidas. Para el cálculo de *t*, entonces se tiene:

- En el numerador: la diferencia de las medias.
- En el denominador: la raíz cuadrada *(var * (1/n1 + 1/n2))*, donde *var* es la varianza.

```
mECE <- mean(ECE)
lECE <- length(ECE)
mSA <- mean(SA)
lSA <- length(SA)
SSW_2 <- SA15 + ECE15
SSW_2
```

```
## [1] 11.06444
```

```
varssw2=SSW_2/((lECE+lSA)-2)
t <- ((mECE-mSA)/sqrt(varssw2*(1/lECE + 1/lSA)))
t
```

```
## [1] 3.130632
```

```
t1 <- ((2*varssw2) - t)
t1
```

```
## [1] -2.479783
```

Grados de libertad para este estadístico son: (n-k)

```
glEA <- ((lECE+lSA)-2)
glEA
```

```
## [1] 34
```

c) Cálculo del valor *p* asociado al *estadístico t* también calculado. ¿Se puede rechazar la hipótesis nula de que **las dos regiones** tienen igual nivel de felicidad con un nivel de confianza del 95%?. Se debe realizar una prueba bilateral para comprobar. Así:

```r
#pvalor = p (t <= -2.48) + p (t >= 32.13)
#pvalor = 2p (t >= 3.13 /H0) = 0.0058
#Valor de la tabla de Grados Libertad
pvalor <- 0.0058
pvalor
```

```
## [1] 0.0058
```

```r
#Validación:
t.test(ECE, SA, alternative="greater", conf.level=0.95, paired=FALSE, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  ECE and SA
## t = 3.1304, df = 9.1312, p-value = 0.005952
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.312384      Inf
## sample estimates:
## mean of x mean of y
##  5.332931  4.580857
```

Observación: Según los valores obtenidos, *pvalor* es menor que 3.13 y mediante la validación con *t.test*, permite comprobar que uno de los valores de *HS* es superior al otro (es decir, **ECE > SA**), de esta forma se puede rechazar entonces la hipótesis nula. Por tanto, estas no tienen el mismo nivel de felicidad.

d) Cálculo de todas las comparaciones entre pares de regiones de la muestra usando la función **pairwise.t.test**. Esta función usa el mismo procedimiento que se ha ilustrado en el punto anterior, pero ahora es para todos los pares de muestras. Hay que aplicar la prueba bilateral. También identificar los pares que son significativamente diferentes con un valor **alfa = 0.05** (95% de nivel de confianza).

```r
pairwise.t.test(datos2015$HS, datos2015$Region, alternative = "two.sided",paired = FALSE, conf.level = 0.95)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  datos2015$HS and datos2015$Region
##
##      AUSNZ   MENA    SEA     SA      EA      SSA     WE      ECE
## MENA 0.02346 -       -       -       -       -       -       -
## SEA  0.02424 1.00000 -       -       -       -       -       -
## SA   0.00042 0.22741 0.61992 -       -       -       -       -
## EA   0.14332 1.00000 1.00000 0.22741 -       -       -       -
## SSA  2.3e-06 9.4e-07 0.00250 1.00000 0.00076 -       -       -
## WE   1.00000 5.7e-06 0.00029 5.4e-08 0.05371 < 2e-16 -       -
## ECE  0.01331 1.00000 1.00000 0.28407 1.00000 2.0e-07 1.0e-07 -
## LC   0.55950 0.03803 0.11947 0.00012 1.00000 3.4e-16 0.28407 0.00519
## NAM  1.00000 0.02424 0.02424 0.00044 0.14409 2.5e-06 1.00000 0.01388
##      LC
## MENA -
## SEA  -
## SA   -
## EA   -
## SSA  -
## WE   -
## ECE  -
## LC   -
## NAM  0.55950
##
## P value adjustment method: holm
```

Observación: De acuerdo a los datos obtenidos, existen determinados pares de muestras que son significativamente diferentes de 95% entre ellos están: **SSA - AUSNZ, SSA - MENA, WE - MENA, WE - SA, WE - SSA, ECE - SSA, ECE - WE, LC - SSA y NAM - SSA**; obteniedo un total de 9 pares de muestras. Adicional, también hay pares de muestras que no se vinculan para nada entre sí.

ANOVA NO PARAMÉTRICO

Se usa este tipo de pruebas para establecer **supuestos poco exigentes** (como simetría o continuidad).

Además, se utiliza como base el fichero 2015, es decir; se continua con el análisis ANOVA entre regiones del año 2015.

Condiciones de aplicación de ANOVA

El test ANOVA requiere que los datos de la muestra cumplan dos suposiciones básicas: normalidad e igualdad de varianzas (homocedasticidad, la cual se da si el error cometido por el modelo tiene siempre la misma varianza). A continuación, se comprueba si se cumplen estas condiciones. De hecho, las suposiciones se tienen que verificar antes de aplicar ANOVA. En este caso, por razón de facilidad, se lo realiza después de aplicar ANOVA.

a) Verificación de la suposición de normalidad. Se puede representar gráficamente con la función *qqnorm* y también con el test "shapiro.test" enviando como parámetro el valor *HS* (happiness score). Esta verificación se visualiza en la Ilustración 13.

```
qqnorm(datos2015$HS, xlab="Quantiles Normales", ylab="Quantiles Muestrales", main="Q-Q plot Datos Normales")
```

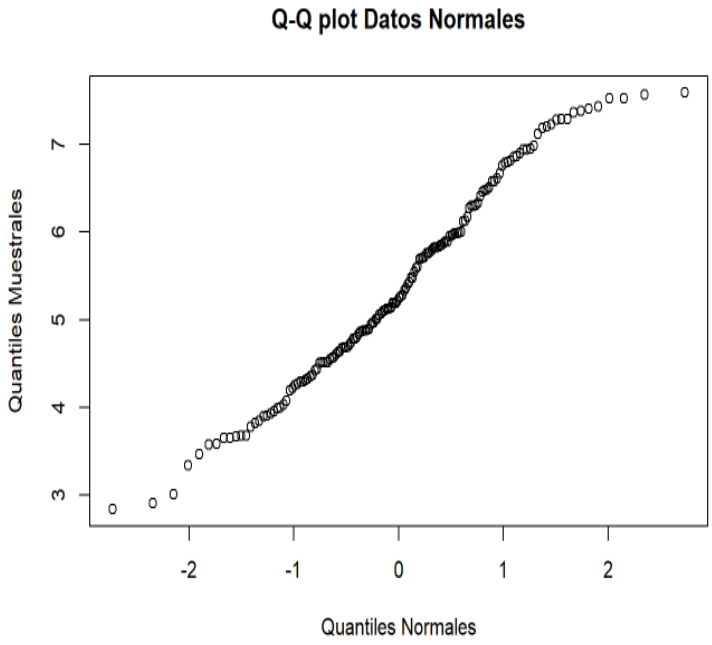


Ilustración 13. Quantiles normales.

```
shapiro.test(datos2015$HS)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  datos2015$HS
## W = 0.97948, p-value = 0.01878
```

Interpretación: Siendo la hipótesis nula que la población esté distribuida normalmente, el *p-valor* es 0.01878 el cual es menor que **alfa** (95% de confianza), entonces la hipótesis nula se rechaza concluyendo que **los datos no vienen de una distribución normal**.

b) A continuación, se realiza la verificación sobre si se cumple la suposición de homogeneidad de varianzas. Para esto, se aplica la prueba "bartlett.test".

```
bartlett.test(datos2015$HS, datos2015$Region)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  datos2015$HS and datos2015$Region
## Bartlett's K-squared = 27.188, df = 9, p-value = 0.001302
```

Interpretación: Al verificar la homogeneidad de varianzas, se identifica que al menos dos de ellas son diferentes, este dato se comprueba en la gráfica del conjunto de muestras y mediante el test de Bartlett que especifica el valor de *p-value* menor que 0.05, de esta manera se rechaza la hipótesis nula, de forma similar a la prueba realizada con "shapiro.test"

REGRESIÓN

La **regresión lineal múltiple** consiste en ajustar modelos lineales entre una variable dependiente y más de una variable independiente.

A continuación, se realiza un análisis de regresión con dos datos del 2015, para lo cual se ejecuta los siguientes pasos:

Aplicación de un modelo de regresión lineal múltiple.

Para aplicar un modelo de regresión lineal múltiple, se considera como variable dependiente el nivel de felicidad (*HS*) y como variables explicativas: *GDP, family, life expectancy, freedom, trust* y *generosity*.

```
ModeloRegMul <- lm(HS~EPC+Family+LE+Freedom+GC+Generosity, data=datos2015)
summary(ModeloRegMul)
```

```
##
## Call:
## lm(formula = HS ~ EPC + Family + LE + Freedom + GC + Generosity,
##     data = datos2015)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40484 -0.31734 -0.02814  0.37189  1.50130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8602     0.1905   9.766  < 2e-16 ***
## EPC           0.8607     0.2203   3.907 0.000141 ***
## Family        1.4089     0.2227   6.327 2.69e-09 ***
## LE            0.9753     0.3163   3.084 0.002433 **
## Freedom       1.3334     0.3850   3.463 0.000694 ***
## GC            0.7845     0.4365   1.797 0.074302 .
## Generosity    0.3889     0.3910   0.995 0.321471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.551 on 151 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.7684
## F-statistic: 87.81 on 6 and 151 DF,  p-value: < 2.2e-16
```

Interpretación de los resultados de los valores *p* y explicación de los coeficientes significativos.

De acuerdo al modelo de regresión múltiple para datos cuantitativos, se evidencia un coeficiente de bondad de 0.7772; mientras que, en el análisis de las variables que interfieren en el nivel de felicidad de los países, tiene un mayor impacto el PIB per cápita, Familia y Libertad, la Expectativa de vida tiene un efecto de un 95% con respecto al resto de predictores.

A continuación en la Ilustración 14 se presentan los valores residuales obtenidos del modelo.

Análisis de los residuos del modelo e interpretación del resultado.

```
par(mfrow=c(2,2))
ModeloRes <- lm (HS~EPC+Family+LE+Freedom+GC+Generosity, data=datos2015)
plot (ModeloRes)
```

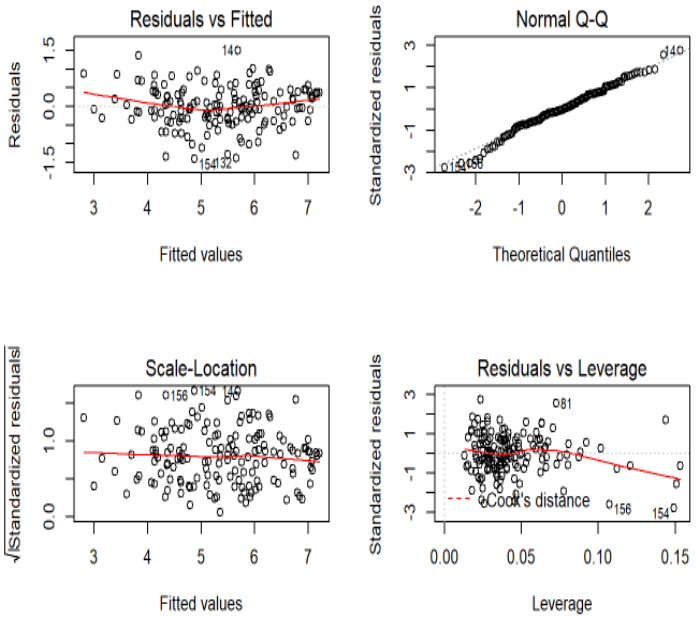


Ilustración 14. Valores residuales del modelo.

Interpretación: En el análisis de residuos, se evidencia que no existen patrones distintos, aunque en la parte inferior de la gráfica se evidencia la presencia de algunos valores elevados. En el análisis de la normal, en la gráfica *Q-Q* se verifica que la variable dependiente está distribuida normalmente, mientras que en la revisión de Escala-Ubicación, los datos se distribuyen por igual a lo largo de los rangos de los predictores. En la relación Residuos vs apalancamiento, es el aspecto típico cuando no hay casos influyentes. En conclusión, **las variables independientes influyen en el resultado de *HS***.

CONCLUSIONES

Como conclusiones se presentan las siguientes:

- La mayoría de autores, por no decir todos, coinciden que con la felicidad se obtiene una buena salud.
- Existen otros hábitos que también aumentan el grado de felicidad, los cuales son: hacer deporte, rodearse de buenas personas, tener una visión optimista en la vida, etc.
- La hipótesis nula establece que todas las medias de la población (medias de los niveles de los factores) son iguales mientras que la hipótesis alternativa establece que al menos una es diferente.
- Mediante el método de ordenamiento de Kruskal-Wallis se obtiene una mejor visibilidad de la relación de grupos y correlación de los resultados de manera eficiente.
- Las variables independientes: GDP, family, life expectancy, freedom, trust y generosity influyen en el resultado de HS (nivel de felicidad).

Según los resultados obtenidos, se responde a las interrogantes planteadas al inicio del proyecto que son:

¿Cuáles son los principales factores que contribuyen a la felicidad?

Los principales factores contribuyentes a la felicidad son:

a. Economy (GDP per Capita)
b. Family
c. Health (Life Expectancy)
d. Freedom
e. Trust (Government Corruption)
f. Generosity

¿Existen diferencias importantes en dichos factores entre países?

Si.

La puntuación de la family tiende a tener el mayor impacto en la puntuación de la felicidad y Economy (GDP per Capita) tiene el segundo mayor impacto.

La confianza (trust) tiene la puntuación más baja de todas las condiciones observadas. Los países que tienen poca o ninguna confianza en los gobiernos, lo

hacen para que los ciudadanos se sientan privados del derecho al voto y no puedan tomar las decisiones de vida que desean, lo que se demuestra en la correlación entre la baja confianza y los bajos puntajes de libertad.

¿Existen diferencias de felicidad en los tres años?

Según los datos obtenidos, si, existe diferencias en cuanto a los niveles de felicidad en los últimos tres años. Pese a que los valores son casi similares en cuanto al cálculo de la media, se puede decir que en el año 2016 hay mayor nivel de felicidad que en el resto de años.

¿Existen relaciones entre las distintas regiones según el nivel de felicidad?

No.

Según se visualiza en los boxplots, no existen relaciones entre cada una de las regiones. Se puede evidenciar que entre Australia and New Zealand (AUSNZ) y North América (NAM) existe una ligera relación en sus varianzas, pero el valor P es muy pequeño por lo que se puede concluir la no relación de HS (nivel de felicidad) en las regiones analizadas.

¿En qué región se encuentran los países más felices y menos felices del mundo?

Los países "más felices" están situados en Europa, especialmente en Dinamarca y Suiza. Mientras tanto, los países "menos felices" están situados en África.

RECOMENDACIONES

Como recomendaciones se presentan las siguientes:

- Si alguien más va a trabajar sobre este mismo tema de predictores de felicidad, realizar búsquedas de información en inglés, ya que en español no hay mucha información; esto en bases de datos científicas como Scupus, Sciencedirect, etc, en donde se debe colocar la o las palabras claves a buscar.
- Extraer de fuentes fidedignas las fórmulas de cómo los países en el mundo hacen el cálculo de la felicidad, con la finalidad de comparar la información subida a los datasets que se publican en repositorios digitales.

TRABAJOS FUTUROS

Los resultados obtenidos luego de los cálculos estadísticos sirven de base para analizar o comparar con data de años más actualizados (cuando se cuente con la data) sobre el mismo tema de predictores de felicidad.

Con el mismo modelo construido, en donde se realizó el análisis de los datos con ANOVA se puede predecir factores de salud a partir del 2018 que no fue analizado en el presente proyecto, así como también en deportivos, climatológicos, etc; solo habría que adaptarle la data correspondiente. Es decir, usar el modelo cuando se cuente con datos de poblaciones que siguen una distribución normal con varianzas iguales entre los niveles de factores.

GLOSARIO DE TÉRMINOS

- PIB (Producto Interno Bruto). Es el valor monetario de los bienes y servicios finales producidos por una economía en un período determinado.
- KDD (Knowledge Discovery in Databases). Es un proceso metodológico y además secuencial que se sigue para encontrar conocimiento en un conjunto de datos en bruto.
- Minería de datos. Es el proceso de búsqueda en grandes bases de datos para encontrar información útil que sirva para la toma de decisiones. También se utiliza el término en inglés «data mining».
- Método. Modo ordenado y sistemático de proceder para llegar a un resultado o fin determinado.
- Factor. Elemento, circunstancia, influencia, que contribuye a producir un resultado.
- Varianza. Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media. Formalmente se calcula como la suma de los residuos al cuadrado divididos entre el total de observaciones. También se puede calcular como la desviación típica al cuadrado.
- Desviación media. Es la medida aritmética de los valores absolutos de las desviaciones respecto a la medida.
- Desviación estándar. Es un índice numérico de la dispersión de un conjunto de datos. Mientras mayor es la desviación estándar, mayor es la dispersión de la población. La desviación estándar es un promedio de las desviaciones individuales de cada observación con respecto a la media de lo que es una distribución.
- SMS (Systematic Mapping Study). Mapeo Sistemático de Estudio, es un método definido para construir un esquema de clasificación y estructurar un campo de interés de estudio.
- SLR (Systematic Literatury Review). Es la Revisión Sistemática de Literatura, que sirve para diseñar un protocolo de búsqueda de información.
- Modelo de minería de datos. Conjunto de datos, estadísticas y patrones que se aplican a los nuevos datos para generar predicciones.
- Algoritmo. Se denomina algoritmo a un grupo finito de operaciones organizadas de manera lógica y ordenada que permite solucionar un determinado problema. Se trata de una serie de instrucciones o reglas establecidas que, por medio de una sucesión de pasos, permiten arribar a un resultado o solución. Los algoritmos se pueden expresar de diversas formas: lenguaje natural, lenguaje de programación, pseudocódigo y diagramas de flujo.

- Redes neuronales. Son un modelo para encontrar esa combinación de parámetros y aplicarla al mismo tiempo. En el lenguaje propio, encontrar la combinación que mejor se ajusta es "entrenar" la red neuronal. Una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones, es decir, para "aplicar" la combinación. Técnica de inteligencia artificial más representativa de la Minería de datos
- Iteración. Significa repetir varias veces un proceso con la intención de alcanzar una meta deseada, objetivo o resultado. Cada repetición del proceso también se le denomina una "iteración", y los resultados de una iteración se utilizan como punto de partida para la siguiente iteración.
- Base de datos. Es una colección de información organizada de forma que un programa de ordenador pueda seleccionar rápidamente los fragmentos de datos que necesite. Una base de datos es un sistema de archivos electrónico. Las bases de datos tradicionales se organizan por campos, registros y archivos.
- Lenguaje de programación. Es un lenguaje formal que proporciona una serie de instrucciones que permiten a un programador escribir secuencias de órdenes y algoritmos a modo de controlar el comportamiento físico y lógico de una computadora con el objetivo de que produzca diversas clases de datos. A todo este conjunto de órdenes y datos escritos mediante un lenguaje de programación se le conoce como programa.
- El análisis de datos. Se encarga de examinar un conjunto de datos con el propósito de sacar conclusiones sobre la información para poder tomar decisiones, o simplemente ampliar los conocimientos sobre diversos temas.
- Depuración. Es el proceso de identificar y corregir errores de programación. En inglés se conoce como debugging, porque se asemeja a la eliminación de bichos (bugs), manera en que se conoce informalmente a los errores de programación.
- ANOVA o AVAR. Es una de las técnicas más utilizadas en los análisis de los datos de los diseños experimentales (técnica del análisis de la varianza). Se utiliza cuando queremos contrastar más de dos medias, por lo que puede verse como una extensión de la prueba t para diferencias de dos medias.

ANEXOS

Anexo 1: Fórmula a usar para determinar la felicidad

El Informe 2017 promedia los resultados de las encuestas realizadas en el trienio 2014-2016. A continuación, se intenta explicar la variación entre países por medio de estas seis variables explicativas:

• PIB per cápita PIB per cápita
• Apoyo social SS
• Esperanza de vida saludable HLE
• Libertad para tomar decisiones de vida FRE
• Generosidad GEN
• Percepción (ausencia de) corrupción PER.

Las variables de la encuesta (WHR 2017, Helliwell et al., p.17) se midieron así:

- El apoyo social es el promedio nacional de las respuestas binarias (0 o 1) a la pregunta: "Si estabas en problemas, ¿tienes parientes o amigos con los que puedas contar para que te ayuden cuando los necesites, o no?".
- Los datos sobre la esperanza de vida se obtienen de los Indicadores del Desarrollo Mundial. Luego se ajusta a la esperanza de vida saludable utilizando datos de la Organización Mundial de la Salud.
- La libertad se define como el promedio nacional de respuestas binarias a la pregunta: "¿Estás satisfecho o insatisfecho con tu libertad de elegir lo que haces con tu vida?".
- La percepción de la corrupción es el promedio de respuestas binarias a dos preguntas:
  "¿Está la corrupción generalizada en el gobierno o no?", y
  "¿Está la corrupción generalizada en las empresas o no?".
- La generosidad se define como el residuo de la regresión del promedio nacional de respuestas a la pregunta: "¿Ha donado dinero a una organización benéfica en el último mes?.

BIBLIOGRAFÍA

[1] Kaggle, W. H. (2018). Kaggle. Obtenido de https://www.kaggle.com/unsdsn/world-happiness/home

[2] Factores de felicidad: http://cort.as/-JbQw

[3] Modelo de mineria de datos: http://cort.as/-JblD

[4] EcuRed. (s.f.). *EcuRed*. Obtenido de https://www.ecured.cu/Miner%C3%ADa_de_Datos

[5] Openclassrooms. (30 de 10 de 2017). Openclassrooms. Obtenido de https://openclassrooms.com/en/courses/4309151-gestiona-tu-proyecto-de-desarrollo/4538221-en-que-consiste-el-modelo-en-cascada

[6] Paises más felices del mundo: https://cnnespanol.cnn.com/2018/03/15/paises-mas-felices-2018-informe-latinoamerica/

[7] Unidas, N. (s.f.). UN.ORG. Obtenido de https://www.un.org/es/events/happinessday/

[8] Britos, P. (febrero de 2014). *Univeridad Tecnológica Nacional de Buenos Aires.* Corbin, J. A. (s.f.). *Psicología y Mente*. Obtenido de https://psicologiaymente.com/miscelanea/paises-mas-felices-mundo-onu

[9] Petersen, K., & Feldt, R. (s.f.). Robertfeldt.net. Obtenido de http://www.robertfeldt.net/publications/petersen_ease08_sysmap_studies_in_se.pdf

[10] Mohammad Karim Sohrabi, Soodeh Akbari, A comprehensive study on the effects of using data mining techniques to predict tie strength, Computers in Human Behavior, Volume 60, 2016, Pages 534-541, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2016.02.092. (http://www.sciencedirect.com/science/article/pii/S0747563216301558)

[11] Francine Espinoza Petersen, Heather Johnson Dretsch, Yuliya Komarova Loureiro, Who needs a reason to indulge? Happiness following reason-based indulgent

consumption, International Journal of Research in Marketing, Volume 35, Issue 1, 2018, Pages 170-184, ISSN 0167-8116, https://doi.org/10.1016/j.ijresmar.2017.09.003.

[12] Yunji Liang, Xiaolong Zheng, Daniel D. Zeng, A survey on big data-driven digital phenotyping of mental health, Information Fusion, Volume 52, 2019, Pages 290-307, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2019.04.001.

[13] Jonathan Kelley, M.D.R. Evans, The new income inequality and well-being paradigm: Inequality has no effect on happiness in rich nations and normal times, varied effects in extraordinary circumstances, increases happiness in poor nations, and interacts with individuals' perceptions, attitudes, politics, and expectations for the future, Social Science Research, Volume 62, 2017, Pages 39-74, ISSN 0049-089X, https://doi.org/10.1016/j.ssresearch.2016.12.007.

[14] Muhammad Fahim Uddin, Jeongkyu Lee, We Are What We Generate - Understanding Ourselves Through Our Data, Procedia Computer Science, Volume 95, 2016, Pages 335-344, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.09.343.

[15] Pérez-Benito, F.J., Villacampa-Fernández, P., Conejero, J.A., García-Gómez, J.M., Navarro-Pardo, E. A happiness degree predictor using the conceptual data structure for deep learning architectures (2019) Computer Methods and Programs in Biomedicine, 168, pp. 59-68. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85034833368&doi=10.1016%2fj.cmpb.2017.11.004&partnerID=40&md5=b6c1e84d751b167a1486ef5505db97f2 DOI: 10.1016/j.cmpb.2017.11.004

[16] Ludwigs, K., Lucas, R., Veenhoven, R. et al. Applied Research Quality Life (2019). https://doi.org/10.1007/s11482-019-09723-2

[17] Wu, Y., Yuan, J., You, Q., Luo, J. The effect of pets on happiness: A data-driven approach via large-scale social media (2016) Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, art. no. 7840808, pp. 1889-1894. Cited 3 times. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85015258712&doi=10.1109%2fBigData.2016.7840808&partnerID=40&md5=be161f175173f601e892c28275de6f28 DOI: 10.1109/BigData.2016.7840808

[18] Zhang, X., Zhang, X., & Chen, X. (2017). Valuing air quality using happiness data: The case of china. *Ecological Economics, 137*, 29-36. doi:10.1016/j.ecolecon.2017.02.020

[19] Nguyen, Q. C., Kath, S., Meng, H. -., Li, D., Smith, K. R., VanDerslice, J. A., . . . Li, F. (2016). Leveraging geotagged twitter data to examine neighborhood happiness, diet, and physical activity. *Applied Geography, 73*, 77-88. doi:10.1016/j.apgeog.2016.06.003

[20] I. Dönmez and E. B. Sönmez, "Feeling Analysis for Sadness and Happiness using Googlen-gram Database Googlen-gram Veritabanı ile Üzüntü ve Mutluluk Üzerine Duygu Analizi," *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, Sarajevo, 2018, pp. 56-60. doi: 10.1109/UBMK.2018.8566245

[21] P. J. Manamela, M. J. Manamela, T. I. Modipa, T. J. Sefara and T. B. Mokgonyane, "The Automatic Recognition of Sepedi Speech Emotions Based on Machine Learning Algorithms," *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD)*, Durban, 2018, pp. 1-7. doi: 10.1109/ICABCD.2018.8465403

[22] S. Bao *et al*., "Mining Social Emotions from Affective Text," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 9, pp. 1658-1670, Sept. 2012. doi: 10.1109/TKDE.2011.188

[23] H. U. Dike, Y. Zhou, K. K. Deveerasetty and Q. Wu, "Unsupervised Learning Based On Artificial Neural Network: A Review," *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, Shenzhen, 2018, pp. 322-327. doi: 10.1109/CBS.2018.8612259

[24] V. L. Chawda and V. S. Mahalle, "Learning to recommend descriptive tags for health seekers using deep learning," *2017 International Conference on Inventive Systems and Control (ICISC)*, Coimbatore, 2017, pp. 1-7. doi: 10.1109/ICISC.2017.8068589

[25]Duncan, O. (7 de mayo de 2018). *MICROSFT*. Obtenido de https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/mining-models-analysis-services-data-mining?view=sql-server-2017#bkmk_mdlArch

[26]Duncan, O. (30 de abril de 2018). MICROSFT. Obtenido de https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017

[27]Krall, C. (s.f.). *Aprenderaprogramar.com*. Obtenido de https://www.aprenderaprogramar.com/index.php?option=com_content&view=article&id=258:mineria-de-datos-2o-parte-modelos-tecnicas-herramientas-dv00106a&catid=45&Itemid=164

[28]Tamayo, C., & Siva, I. (s.f.). Universidad Católica Los Angeles de Chimbote. Obtenido de http://www.postgradoune.edu.pe/pdf/documentos-academicos/ciencias-de-la-educacion/23.pdf

[29]Lenguajesdeprogramacion.net. (s.f.). *lenguajesdeprogramacion.net.* Obtenido de https://lenguajesdeprogramacion.net/r/