Universitat Oberta de Catalunya

uoc.edu

Estudio de predictores de felicidad a nivel mundial

Área: Ciencia de datos aplicada a la salud
Autor: María Augusta Jimbo Granda
Director: José Luis Iglesias Allones
Codirector: Liliana Elvira Enciso Quispe

Marzo de 2019

Índice

Capítulo 1: Contexto y Justificación del Proyecto

OBJETIVOS DEL PROYECTO

Objetivo Principal:

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

Objetivos Específicos:

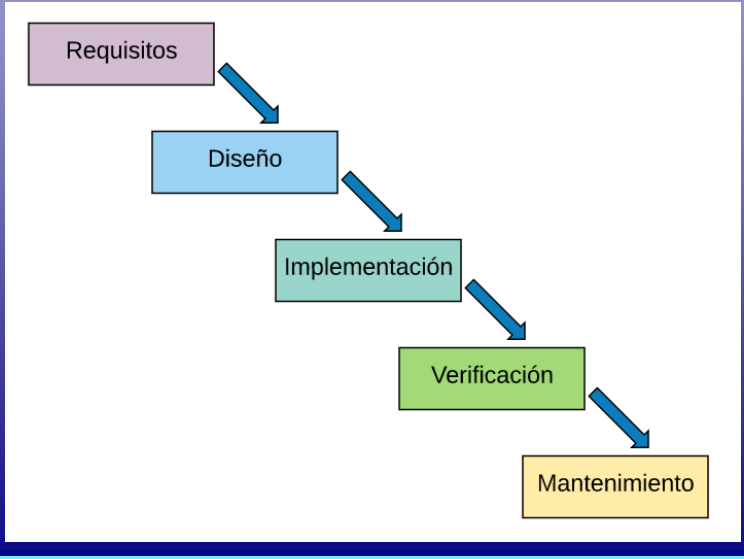
- Determinar los principales factores que contribuyen a la felicidad.
- Descubrir diferencias de factores entre países.
- Determinar si existe diferencia de felicidad en los tres años.
- Determinar en qué región se encuentran los países más felices y menos felices del mundo.
- Analizar la evolución que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

Capítulo 1: Contexto y Justificación del Proyecto

METODOLOGÍA Y ENFOQUE

Metodología

Para el desarrollo del proyecto, la metodología que se utilizó es la del modelo en cascada. A continuación, se presenta las etapas de esta metodología:



Capítulo 2: Estado del Arte del Proyecto

INVESTIGACIONES REALIZADAS SOBRE EL TEMA

METODOLOGÍA DE BÚSQUEDA DE INFORMACIÓN

- SMS (Systematic Mapping Study): Es un método definido para construir un esquema de clasificación y estructurar un campo de interés de estudio.

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 los pasos del proceso de mapeo sistemático.

Capítulo 2: Estado del Arte del Proyecto

SLR (Systematic Literatury Review): Lo que los distingue de los mapas sistemáticos es su análisis en profundidad en forma de un resumen narrativo detallado.

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. 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.

Capítulo 2: Estado del Arte del Proyecto

TRABAJOS RELACIONADOS

He consultado información de los predictores de felicidad en el mundo, en 3 bases de datos científicas importantes como son: sciencedirect, scopus e 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	

Capítulo 2: Estado del Arte del Proyecto

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”.

Capítulo 3: Metodología de Estudio y Herramientas

POBLACIÓN DE ESTUDIO

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:

- Dataset2015.csv (158 filas,12 columnas)
- Dataset2016.csv (157 filas,13 columnas)
- Dataset2017.csv (155 filas,12 columnas).

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

Capítulo 4: Resultados

PRESENTACIÓN DEL MODELO DE MINERÍA DE DATOS

El modelo de minería de datos construído, consistió en la integración de los siguientes elementos:

- La metodología en cascada en la cual la construcción del modelo se ejecutó vigilando el proceso y ajustando parámetros para continuar con la siguiente etapa.
- 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.
- Luego, 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.
- Aplicar ANOVA, es decir, aplicar un algoritmo de minería de datos (k-NN) para un problema de clasificación.

Capítulo 4: Resultados

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

Para contestar las interrogantes planteadas al inicio del proyecto, se ejecutan los pasos:

a) Obtención de los datos.

b) Realización de un análisis descriptivo de los datos.

c) Comprobación de los supuestos (normalidad y presencia de outliers).

d) Interpretación de los resultados.

El paso inicial para representar y analizar datos estadísticos, es la ejecución del **preprocesado**. En el presente proyecto, interesa dar respuesta a algunas interrogantes sobre la felicidad mundial, pero entre las más importantes están:

1. ¿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.

Capítulo 4: Resultados

2. ANOVA para muestras apareadas (Repeated Measures ANOVA).
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 paises medidos en 3 momentos diferentes de tiempo. Por esta razón, es más apropiado usar la prueba “repeated measures ANOVA”.
Para calcular ANOVA, se debe ir eliminando la variabilidad entre paises, para ello ejecutar los pasos:

a) Cálculo de ANOVA.

```
#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
```

Capítulo 4: Resultados

```
##   [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)
```

b) Identificación del término error.

```
#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
```

Capítulo 4: Resultados

c) 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
```

Capítulo 4: Resultados

d) Comprobación: Verificar si el nuevo *SSW* corresponde al *SSW* anterior menos el término de error entre paises.

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

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

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

e) 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
```

Capítulo 4: Resultados

2. ¿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) Preparación del data frame para aplicar ANOVA.

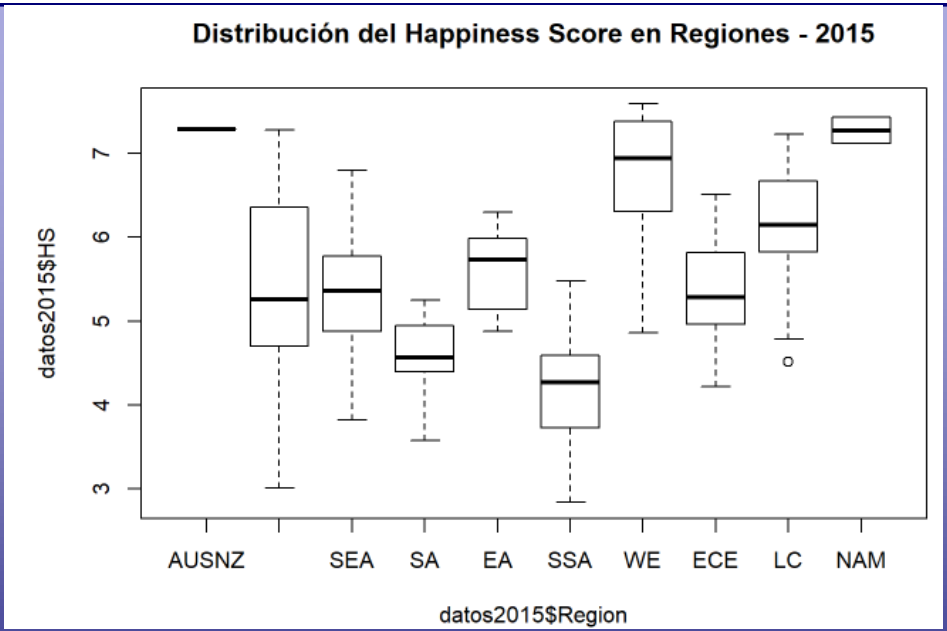
```
datset2015 <- data.frame(datos2015$Region, datos2015$HS)
str(datset2015)
```

```
## '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.

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

Capítulo 4: Resultados



c) ¿Qué método es el más apropiado para este caso de estudio?.
Se evidencia que existe diferencias en el número de muestras por *Region* en cuanto al análisis de su *HS*, por lo tanto; es necesario realizar un estudio ANOVA para muestras independientes.

Capítulo 4: Resultados

d) Cálculo de ANOVA entre todas las regiones.

```
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 en el boxplot, el resultado de *p-valor* con ANOVA y los datos obtenidos mediante modelado de tablas ANOVA, se ve que no hay relaciones entre las regiones. Se concluye la no relación de *HS* en las regiones analizadas.

Conclusiones

- 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).

Conclusiones

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:
Economy (GDP per Capita), family, health (life expectancy), freedom, trust (government corruption), 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.

Conclusiones

- ¿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.

- ¿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.

Trabajos Futuros

Los resultados obtenidos luego de los cálculos estadísticos sirven de base para analizar o comparar con data de años más actualizados 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.

GRACIAS POR SU ATENCIÓN