Investigación clínica

Modelos para la gestión de pacientes con problemas derivados de la diabetes



Autores

Mayo 2018

Contenido

[1. Entendimiento del negocio 3](#_Toc514602458)

[- Objetivos de negocio 4](#_Toc514602459)

[2. Development Plan 5](#_Toc514602460)

[- Análisis inicial y pre-procesado de los datos 5](#_Toc514602461)

[2..1. Tabla análisis de los datos 5](#_Toc514602462)

[2..2. Gráficos 8](#_Toc514602463)

[2..3. Limpieza de dato 26](#_Toc514602464)

[- Selección de los campos significativos 27](#_Toc514602465)

[- Entrenamiento de los modelos 29](#_Toc514602466)

[- Comparación de los resultados 30](#_Toc514602467)

[- Conclusiones 31](#_Toc514602468)

# Entendimiento del negocio

[[1]](#footnote-1)La diabetes es una enfermedad crónica que aparece cuando el páncreas no produce insulina suficiente o cuando el organismo no utiliza eficazmente la insulina que produce. La insulina es una hormona que regula el azúcar en la sangre. El efecto de la diabetes no controlada es la hiperglucemia (aumento del azúcar en la sangre), que con el tiempo daña gravemente muchos órganos y sistemas, especialmente los nervios y los vasos sanguíneos.

Existen tres tipos de diabetes: diabetes gestacional, diabetes de tipo1 y diabetes de tipo 2. Esta última representa la mayoría de los casos mundiales. Hasta hace poco, este tipo de diabetes solo se observaba en adultos, pero en la actualidad también se está manifestando en niños.

El coste médico para una persona con diabetes alcanza un promedio aproximado de $12,000 al año. Esta cantidad es más del doble de los gastos médicos para las personas que no tienen diabetes. En particular, en los Estados Unidos, el coste médico asociado a la diabetes supera los $200 billones.

En los últimos años, la diabetes ha aumentado con mayor rapidez en los países de ingresos medianos y bajos.

De acuerdo con la Organización Mundial de la Salud, el número de personas con diabetes ha aumentado de 108 millones en 1980 a 422 millones en 2014 y para 2035 se prevén 592 millones de enfermos. Será la séptima causa de mortalidad en 2030.

En 2014, el 8,5% de los adultos (18 años o mayores) tenía diabetes. En 2015 fallecieron 1,6 millones de personas como consecuencia directa de la diabetes y los niveles altos de glucemia fueron la causa de otros 2,2 millones de muertes en 2012.

Millones de personas en todo el mundo podrían estar en riesgo de muerte prematura debido a un mal diagnóstico de diabetes o porque no están recibiendo tratamientos efectivos para la enfermedad. Esta es la conclusión de una investigación llevada a cabo en siete países por el Instituto de Métrica y Evaluación de la Salud (IMHS) de la Universidad de Washington, Estados Unidos.

## Objetivos de negocio

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Goal | Attributes | Indicator of success | Dataset | Bussines goal it helps to achive (%) |
| Evitar recaída de pacientes enfermos de diabetes | Datos aportados por parte del cliente | 85% | Dataset Group9.csv |  |
| Evitar pruebas de diagnóstico innecesarias | Datos apostados por parte del cliente | 80% | Dataset Group9.csv |  |

# Development Plan

## Análisis inicial y pre-procesado de los datos

Se han aportado dos ficheros para la comprensión de los datos:

* **DataSet Group 9.csv**

Fichero en formato CSV con todos los datos de los pacientes.  
Contiene un total de cinco mil y una entradas y treinta y seis features.

* **DataAnalytics - Práctica final.pdf**

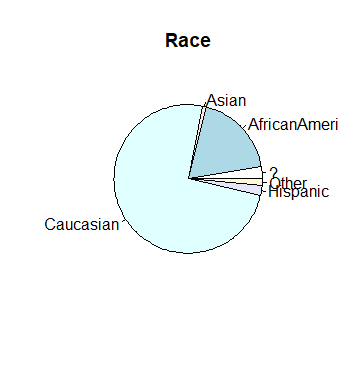
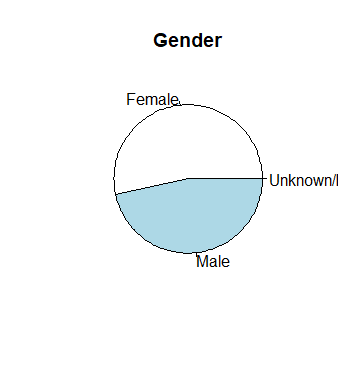
Páginas 2-4.  
Fichero formato pdf con descripción del problema, descripción de los campos del fichero DataSet Group 9.csv y tabla de valores para algunas de las features de dicho fichero.

### Tabla análisis de los datos

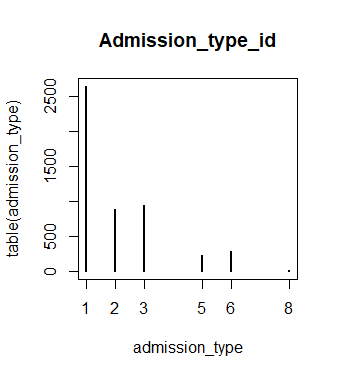
Ha continuación podemos ver una tabla que contiene todas las variables de estudio con una pequeña descripción.

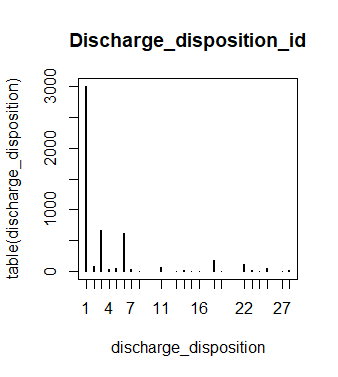
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Nombre de la variable** | **Descripción** | **Tipo de variable** | **Valor mínimo** | **Valor máximo** | **N.º de nulos** |
| **X1** | Número de la entrada del fichero | Auto numérico discreto | 0 | 4999 | 0 |
| **patient\_nbr** | Identificador del paciente en el hospital | Numérico discreto | 135 | 189332087 | 0 |
| **race** | Raza del paciente | Categórica nominal | - | - | 134 |
| **gender** | Género del paciente | Categórica nominal | - | - | 1(Unkown) |
| **age** | Intervalo de edad del paciente | Intervalo numérico entero | [0-10) | [90-100) | 0 |
| **admission\_type\_id** | Tipo de ingreso del paciente en el hospital | Categórico numeral | - | - | 283 |
| **Nombre de la variable** | **Descripción** | **Tipo de variable** | **Valor mínimo** | **Valor máximo** | **N.º de nulos** |
| **discharge\_disposition\_id** | Razón del alta médica del paciente | Categórico numeral | - | - | 176 |
| **admission\_source\_id** | Área de procedencia del hospital | Categórica numérica | - | - | 335 |
| **time\_in\_hospital** | Tiempo que ha pasado ingresa el paciente en el hospital. Medido en días | Numérica discreta | 1 | 14 | 0 |
| **num\_lab\_procedures** | Número de test en laboratorio ligados al paciente durante su ingreso | Numérica discreta | 1 | 114 | 0 |
| **num\_procedures** | Número de procedimientos ligados al paciente (a parte de los test en laboratorio) | Numérica discreta | 0 | 6 | 0 |
| **num\_medications** | Número distinto de medicamentos suministrados durante el ingreso | Numérica discreta | 1 | 67 | 0 |
| **number\_outpatient** | Número de visitas como paciente externo en el año anterior al ingreso | Numérica discreta | 0 | 27 | 0 |
| **number\_emergency** | Número de visitas como urgencia en el año anterior al ingreso | Numérica discreta | 0 | 63 | 0 |
| **number\_inpatient** | Número de ingresos en el hospital el año anterior | Numérica discreta/Categórica númerica | 0 | 16 | 0 |
| **diag\_1** | Diagnostico | Categórica numeral | 5 | 410.0 | 0 |
| **diag\_2** | Diagnostico | Categórica numeral | 8 | 410.0 | 0 |
| **diag\_3** | Diagnostico | Categórica numeral | 8 | 410.0 | 0 |
| **diag\_4** | Diagnostico | Categórica numeral | 2.500.805.279.287,00 | 62.008.680.532.454.900,00 | 0 |
| **number\_diagnoses** | Número de diagnósticos realizados al paciente | Numérica discreta | 1 | 16 | 0 |
| **nateglinide** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **Nombre de la variable** | **Descripción** | **Tipo de variable** | **Valor mínimo** | **Valor máximo** | **N.º de nulos** |
| **chlorpropamide** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **acetohexamide** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **glipizide** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **glyburide** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **tolbutamide** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 4012 |
| **pioglitazone** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **rosiglitazone** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **troglitazone** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **examide** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **insulin** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **glimepiride-pioglitazone** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **metformin-rosiglitazone** | Cambio en la dosis del medicamento o | Categórica nominal | - | - | 0 |
| **metformin-pioglitazone** | Cambio en la dosis del medicamento | Categórica nominal | - | - | 0 |
| **change** | Cambio en el tipo de tratamiento del paciente | Booleana | - | - | 0 |
| **diabetesMed** | Tratamiento prescrito | Booleana | - | - | 0 |
| **readmitted** | paciente ha vuelto a visitar el hospital después de este día | Booleana | - | - | 0 |

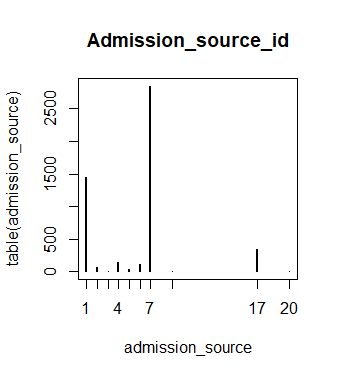
### Gráficos

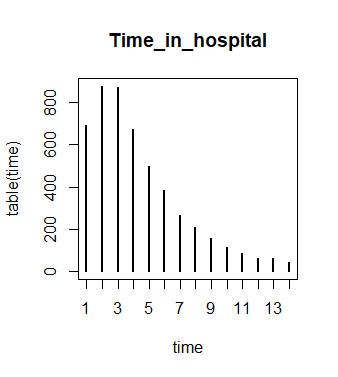
Ha continuación podemos ver gráficos sobre las distribuciones de las variables que se consideran más significativas para la mejor compresión.  


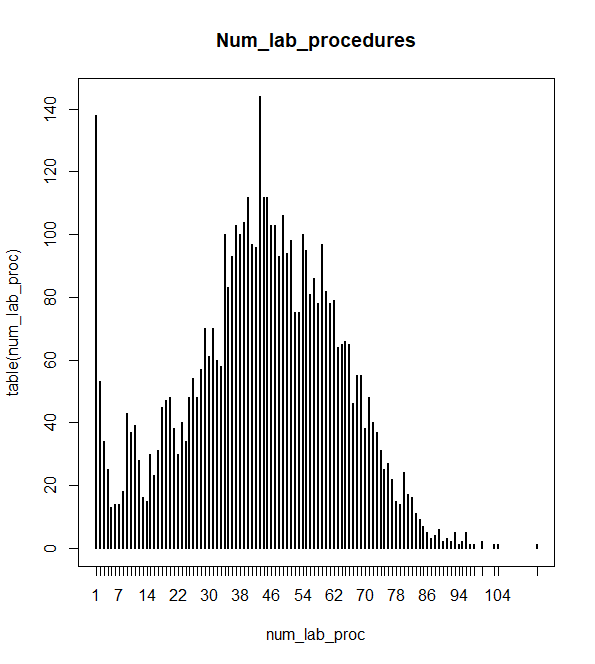
#### 

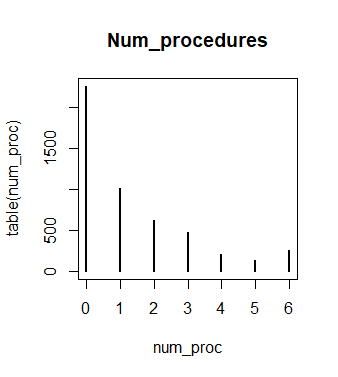


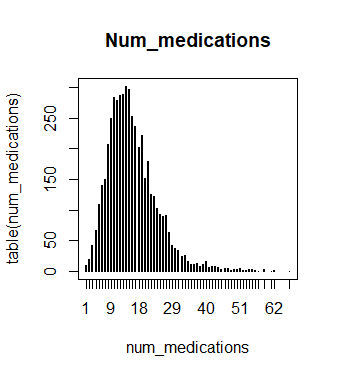


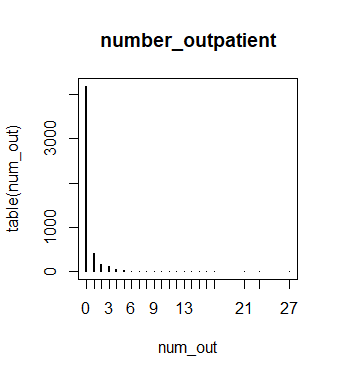


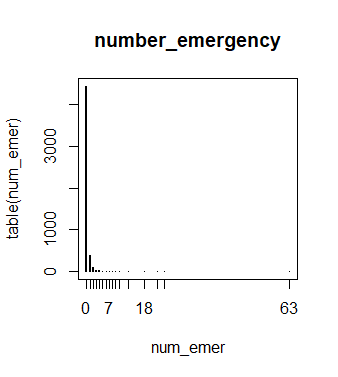


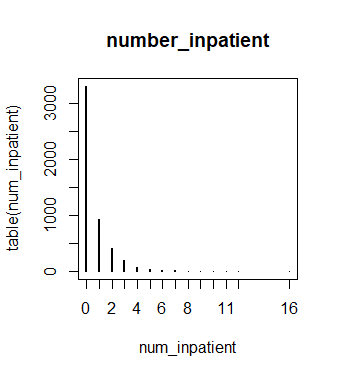


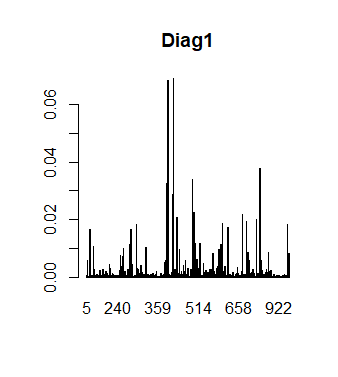
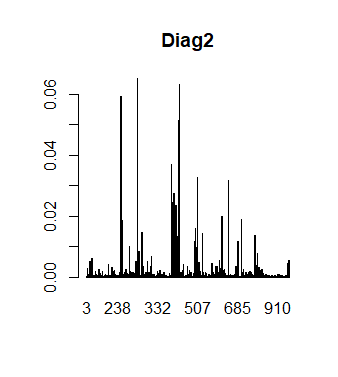


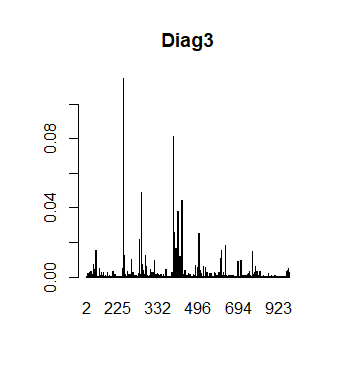


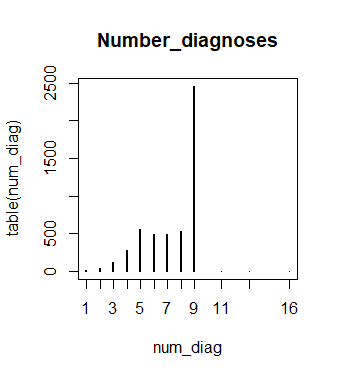


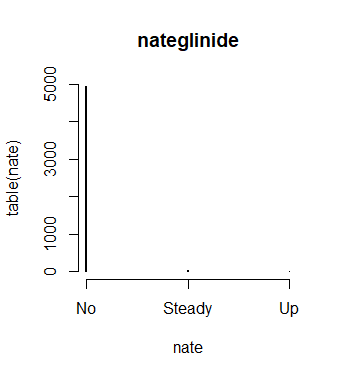


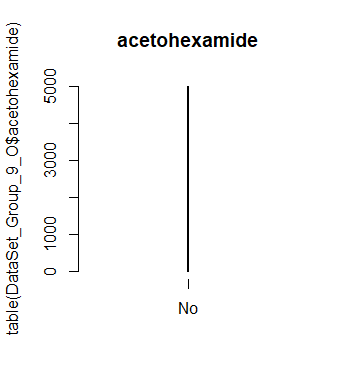
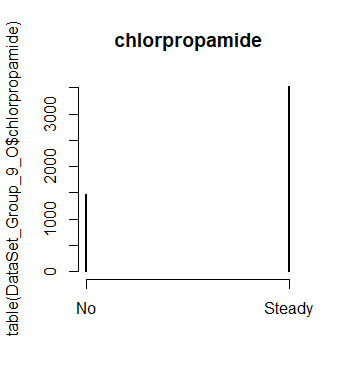


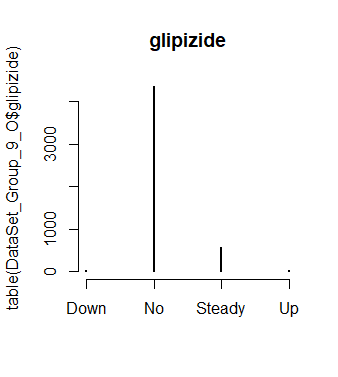


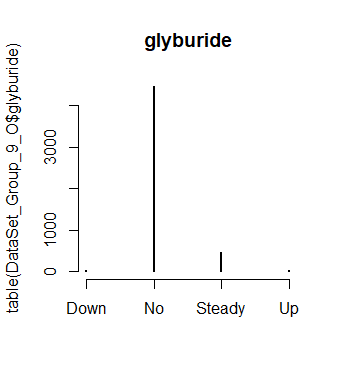


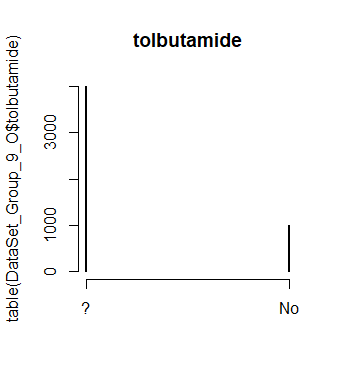


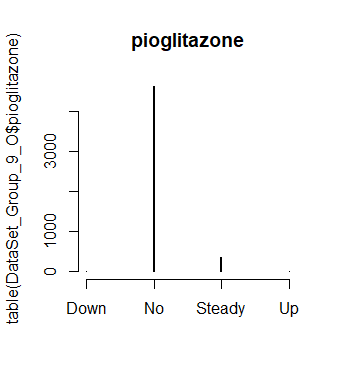
A continuación, podemos ver los diferentes niveles de azúcar de los pacientes (Steady, Up, No, Down) en los distintos tratamientos:  
  


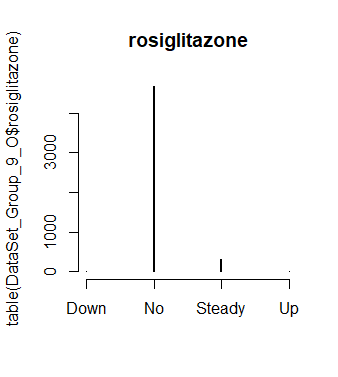


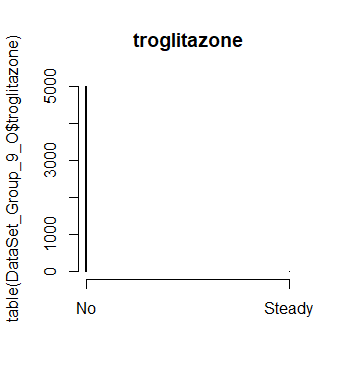


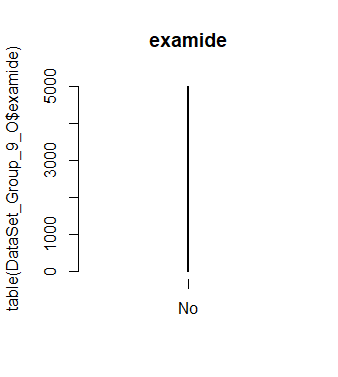


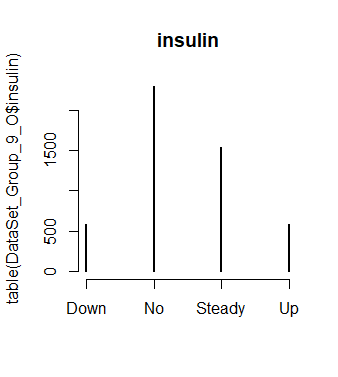


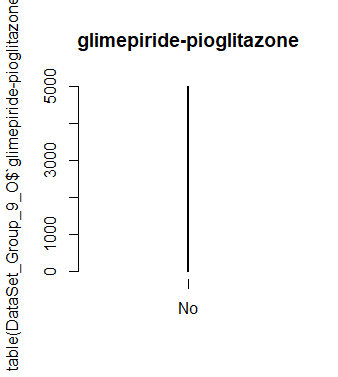


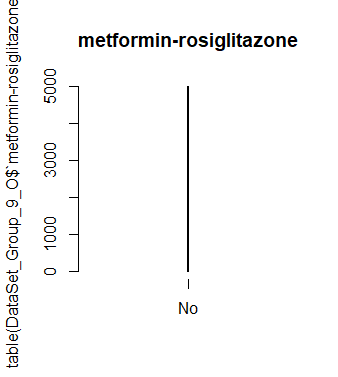


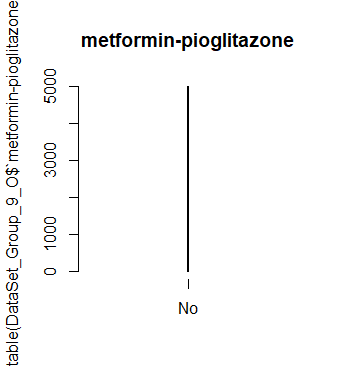


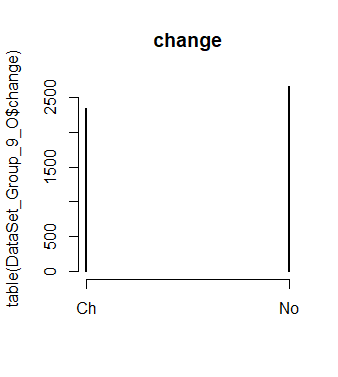


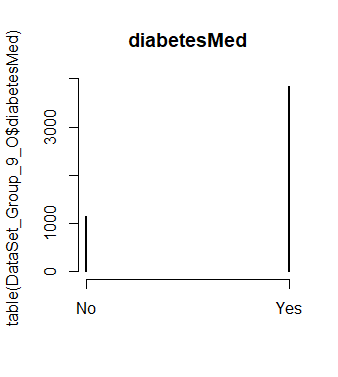




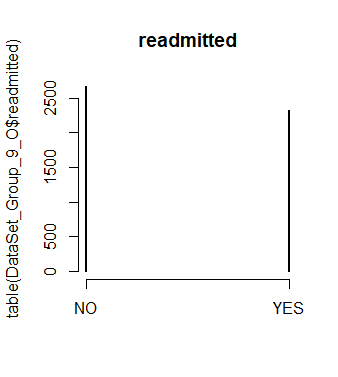








**Resultados del dataset: ¿Volverá al hospital?**



### Limpieza de dato

El proceso de limpieza de datos es vital para un buen análisis de los datos, ya que una mala limpieza de los datos nos puede llevar a resultados erróneos.

La limpieza de estos datos se ha llevado a cabo en dos fases:

* **Búsqueda de duplicados**: esta búsqueda se ha llevado a cabo a través de la herramienta Microsoft Excel que nos permite abrir el dataset de datos y aplicar directamente una función de búsqueda de duplicados.

Como resultado hemos obtenido que no existe ningún duplicado dentro de nuestro conjunto de datos.

* **Análisis de los datos:** Existen algunos valores que no se ajustan al rango de valores usual del dataset. En general, se ha podido observar que todos estos datos pertenecen a valores con valor desconocido. En concreto las variables que cuentan con este tipo de datos son:
  + **Race:**  Contiene un total de 135 entradas con valor ‘?’.
  + **Gender:** Contiene únicamente un nulo. Este valor nulo conincide con una categoría a la que se le ha llamado ‘Unknown’.
  + **Admission\_type\_id:** Contiene un total de 284 nulos. Estos valores nulos corresponden al valor ‘6’.
  + **Discharge\_disposition\_id:** Contiene un total de 176 nulos que corresponden al valor ‘18’.
  + **Admission\_source\_id:** Contiene un total de 335 nulos que pertenecen al valor ‘17’.
  + **Tolbutamide:** Contiene un total de 4012 nulos con valor ‘?’.

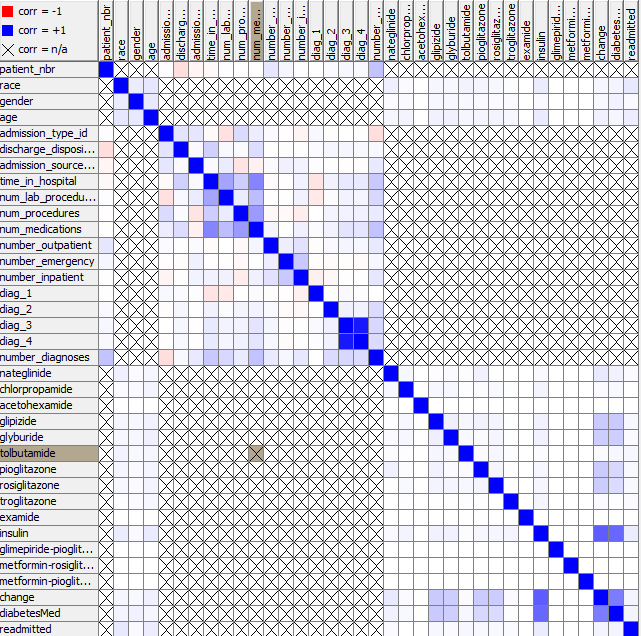
El tratamiento de todos estos valores nulos se ha llevado a cabo a través de la herramienta KNIME, que es una plataforma de minería de datos que no permite el tratamiento de estos valores.  
KNIME nos da la opción de tratar los siguientes valores: cadenas de caracteres, números enteros, números decimales.  
Para nuestro análisis hemos optado por asignar valores de mediana para que la distribución de los datos no se vea afectada y para el tratamiento de las cadenas de caracteres hemos optado por asignar el valor más frecuente. Esta corrección únicamente se ha llevado a cabo para los campos race (lo que supone un 0,027% del total de los datos) y gender (0,0002% de los datos).

## Selección de los campos significativos

En nuestro conjunto de datos observamos variables que no son relevantes para el estudio y por este motivo las hemos descartarlas del modelo, como son:

* **X1**: identificadores del registro.
* **Patient\_nbr**: identificador del paciente en el hospital.
* **Tobultamide:** En esta variable, el número de entradas nulas es 4012 de un total de 5000 entradas, lo que supone que desconocemos un total del 80,24% de los datos.
* **Acetoexamide:** Variable que puede tomar cuatro valores diferentes (Steady, No,Up,Down). Únicamente toma el valor ‘No’, por lo que a la hora de hacer el análisis es una variable que no aporta información al estudio.
* **Examide:** Variable que puede tomar cuatro valores diferentes (Steady, No,Up,Down). Únicamente toma el valor ‘No’, por lo que a la hora de hacer el análisis es una variable que no aporta información al estudio.
* **Glimepiride-pioglitazone:** Variable que puede tomar cuatro valores diferentes (Steady, No,Up,Down). Únicamente toma el valor ‘No’, por lo que a la hora de hacer el análisis es una variable que no aporta información al estudio.
* **Metformin-rosiglitazone:** Variable que puede tomar cuatro valores diferentes (Steady, No,Up,Down). Únicamente toma el valor ‘No’, por lo que a la hora de hacer el análisis es una variable que no aporta información al estudio.
* **Metformin-pioglitazone:** Variable que puede tomar cuatro valores diferentes (Steady, No,Up,Down). Únicamente toma el valor ‘No’, por lo que a la hora de hacer el análisis es una variable que no aporta información al estudio.

Para descartar la relación entre algunas de las variables, hemos optado por hacer un test de correlación con la herramienta KNIME, donde podemos observar la siguiente matriz de correlación:



Podemos observar que existe correlación entre las variables del dataset diag\_3 y diag\_4 con una corrección de 1.  
Existen otras variables como son insuline y change, con un valor de correlación 0,6402. Con este valor, no podemos descartar el uso de una de estas dos variables ya que no es lo suficientemente significativo para ello.

Hemos decidido descartar la variable diag\_3, ya que, al realizar diferentes entrenamientos de los modelos, los resultados obtenidos con la variable diag\_3 son peores (peor porcentaje de exactitud y mayor número de falsos positivos) que con los variable diag\_4.

## Entrenamiento de los modelos

En entrenamiento de los diferentes modelos, al igual que en los apartados anteriores, se ha llevado a cabo a través del programa KNIME.  
Los ficheros de los diferentes modelos entrenados los podemos encontrar en la carpeta ‘Modelos’, que se encuentra en el mismo archivo zip que este documento.

Los modelos entrenados cuentan con validación cruzada, que [[2]](#footnote-2)es una [técnica](https://es.wikipedia.org/wiki/T%C3%A9cnica) utilizada para [evaluar](https://es.wikipedia.org/wiki/Evaluar) los resultados de un [análisis estadístico](https://es.wikipedia.org/wiki/An%C3%A1lisis_estad%C3%ADstico) y garantizar que son [independientes](https://es.wikipedia.org/wiki/An%C3%A1lisis_discriminante) de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la [media aritmética](https://es.wikipedia.org/wiki/Media_aritm%C3%A9tica) obtenida de las medidas de [evaluación](https://es.wikipedia.org/wiki/Evaluaci%C3%B3n) sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la [predicción](https://es.wikipedia.org/wiki/Predicci%C3%B3n) y se quiere [estimar](https://es.wikipedia.org/wiki/Estimaci%C3%B3n_estad%C3%ADstica) la precisión de un modelo que se llevará a cabo a la práctica.  
En el entrenamiento de los modelos se ha contado con el nodo ‘X-partitioner’ y ‘X-Agregator’ para llevarla a cabo.

Los modelos entrenados son los siguientes:

* **Clasificador Bayesiano**:  [clasificador](https://es.wikipedia.org/wiki/Clasificaci%C3%B3n) [probabilístico](https://es.wikipedia.org/wiki/Probabilidad) fundamentado en el [teorema de Bayes](https://es.wikipedia.org/wiki/Teorema_de_Bayes) y algunas hipótesis simplificadoras adicionales.
* **Árbol de decisión**: modelo de predicción que dado un conjunto de datos los analiza a través de construcciones lógicas y así poder distinguir las diferentes opciones de un problema.
* **K-nearest neighbors(Vecinos más próximos):** clasificador que nos da un conjunto de prototipos que obtiene obteniendo la función de densidad de los datos.
* **Red neuronal**: algoritmo que recrea el método de aprendizaje del cerebro humano.
* **Random forest:** es una combinación de árboles predictores.

## Comparación de los resultados

* **Clasificador Bayesiano**

Tras el entrenamiento obtenemos un nivel de exactitud del 61,22%.

Sobre si el paciente volverá o no en menos de treinta días al hospital, obtenemos los siguientes resultados:

|  |  |  |
| --- | --- | --- |
| **Readmitted** | **NO** | **YES** |
| **NO** | 1831 | 842 |
| **YES** | 1097 | 1230 |

* **Árbol de decisión**:

Tras el entrenamiento obtenemos un nivel de exactitud del 56,76%.

Sobre si el paciente volverá o no en menos de treinta días al hospital, obtenemos los siguientes resultados:

|  |  |  |
| --- | --- | --- |
| **Readmitted** | **NO** | **YES** |
| **NO** | 1587 | 1060 |
| **YES** | 1079 | 1221 |

* **K-nearest neighbors (Vecinos más próximos):**

Tras el entrenamiento obtenemos un nivel de exactitud del 54%.

Sobre si el paciente volverá o no en menos de treinta días al hospital, obtenemos los siguientes resultados:

|  |  |  |
| --- | --- | --- |
| **Readmitted** | **NO** | **YES** |
| **NO** | 414 | 126 |
| **YES** | 340 | 126 |

* **Red neuronal**

Tras el entrenamiento obtenemos un nivel de exactitud del 51,8%.

Sobre si el paciente volverá o no en menos de treinta días al hospital, obtenemos los siguientes resultados:

|  |  |  |
| --- | --- | --- |
| **Readmitted** | **NO** | **YES** |
| **NO** | 1576 | 1097 |
| **YES** | 1313 | 1014 |

* **Random Forest**

Tras el entrenamiento obtenemos un nivel de exactitud del 62,56 %.

Sobre si el paciente volverá o no en menos de treinta días al hospital, obtenemos los siguientes resultados:

|  |  |  |
| --- | --- | --- |
| **Readmitted** | **NO** | **YES** |
| **NO** | 2008 | 665 |
| **YES** | 1207 | 1120 |

## Conclusiones

Observamos que mejor porcentaje de exactitud lo obtenemos con el modelo **Random Forest**, con un valor del 62,56%.  
Además, es el modelo que mejor número de falsos positivos tiene, con un total de 665.

1. Informe mundial sobre la diabetes, OMS. [↑](#footnote-ref-1)
2. Wikipedia: https://es.wikipedia.org/wiki/Validaci%C3%B3n\_cruzada [↑](#footnote-ref-2)