Reporte

Proyecto Final

—

Predicción de Reingreso de Pacientes con Diabetes

—

Michael Fallas León

INTRODUCCIÓN

Para éste proyecto se utiliza el lenguaje de programación Python y algunas bibliotecas relacionadas al análisis de datos y aprendizaje automático con el fin de explorar datos y crear modelos, y para profundizar en la aplicación de algoritmos de aprendizaje automático. Además, se busca seleccionar el modelo que mejor se adapte a los datos comparando sus características de rendimiento y ajustando parámetros con el fin de optimizar los resultados.

Los datos seleccionados para el proyecto utilizan factores relacionados con el reingreso y otros datos de pacientes que padecen de diabetes. El conjunto de datos representa 10 años (1999 - 2008) de atención clínica en 130 hospitales y redes de entrega integradas de Estados Unidos. Incluye más de 50 características que representan los resultados del paciente y del hospital. El conjunto contiene información de la base de datos para encuentros que satisfacían los siguientes criterios:

1. Es un encuentro hospitalario (ingreso hospitalario).
2. Es un encuentro diabético; es decir, se ingresó al sistema cualquier tipo de diabetes como diagnóstico.
3. La duración de la estadía fue de al menos 1 día y como máximo 14 días.
4. Se efectuaron pruebas de laboratorio durante el encuentro.
5. Se administraron medicamentos durante el encuentro.

Tomando en cuenta lo anterior, los datos contienen atributos como el número de paciente, raza, sexo, edad, tipo de admisión, tiempo en el hospital, especialidad médica del médico de admisión, número de pruebas de laboratorio realizadas, resultados de la prueba HbA1c, diagnóstico, número de medicamentos, medicamentos para la diabetes, visitas de emergencia en el año anterior a la hospitalización, número de visitas ambulatorias, número de hospitalizaciones, entre otros.

Según la OMS, para el año 2014 existían alrededor de 422 millones de personas con algún tipo de diabetes. Con el conjunto de datos actual se puede predecir si un paciente será o no readmitido en el hospital, con el fin de modifical el tratamiento y así evitar el

reingreso de este. Por lo tanto, el conjunto de datos nos muestra 3 tipos de salida distintos:

- Sin reingreso.
- Reingreso en menos de 30 días. Ésta situación no es buena, porque puede indicar que el tratamiento no era el indicado.
- Reingreso en más de 30 días. Ésta situación puede deberse al estado del paciente.

Los datos se presentan en nombre del Centro de Investigación Clínica y Traslacional, Virginia Commonwealth University, un receptor de NIH CTSA Grant UL1 TR00058 y un receptor de los datos CERNER. John Clore (jclore '@' vcu.edu), Krzysztof J. Cios (kcios '@' vcu.edu), Jon DeShazo (jpdeshazo '@' vcu.edu) y Beata Strack (strackb '@' vcu.edu) . Estos datos son un resumen desidentificado de la base de datos de Datos de salud (Cerner Corporation, Kansas City, MO).

Fuente: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

Objetivos

- Determinar si es posible predecir la readmisión o no del paciente.
- Determinar la influencia, de distintos atributos, en si el paciente será o no reingresado en el hospital

Problema

Muchos pacientes reingresan al hospital después de salir del internado, por lo que se desea conocer si es posible predecir el reingreso con el fin de determinar si es necesario algún cambio en la medicación.

Hipótesis

A continuación, se mencionan las hipótesis planteadas:

- Entre mayor edad tenga el paciente es más probable que reingrese nuevamente al hospital.
- Entre más tiempo permaneció internado el paciente en el hospital la última vez, es menos probable que reingrese por un problema similar.
- La probabilidad de que el paciente no sea ingresado al hospital es mayor si éste presenta una prescripción de medicamentos para la diabetes.
- El cambio de medicación disminuye la probabilidad de que el paciente no reingrese al hospital.

Análisis y Visualización de Datos

En la presente sección se muestra un resumen del análisis obtenido de la exploración de los datos (EDA), donde se detallan aspectos que destacan en cada uno de los atributos, así como la visualización de estos. Con el fin de entender influencias de los atributos, al momento de definir si el paciente será o no reingresado.

A continuación, se muestran visualizaciones para determinados atributos, así como el análisis de estos:

Edad del paciente
El atributo de edad fue modificado de la siguiente manera:

- Child: [0,10[
- Young Adults: [10,30[
- Adults: [30,40[
- Middle-Aged Adults: [40,70[
- Older Adults: [70,100[

Se desea conocer si la edad de los pacientes tiene influencia en su reingreso en el hospital después de estar hospitalizado. Esto por medio de un gráfico de barras que muestre la distribución de los pacientes dentro de cada rango de edad. (ver figura 1).

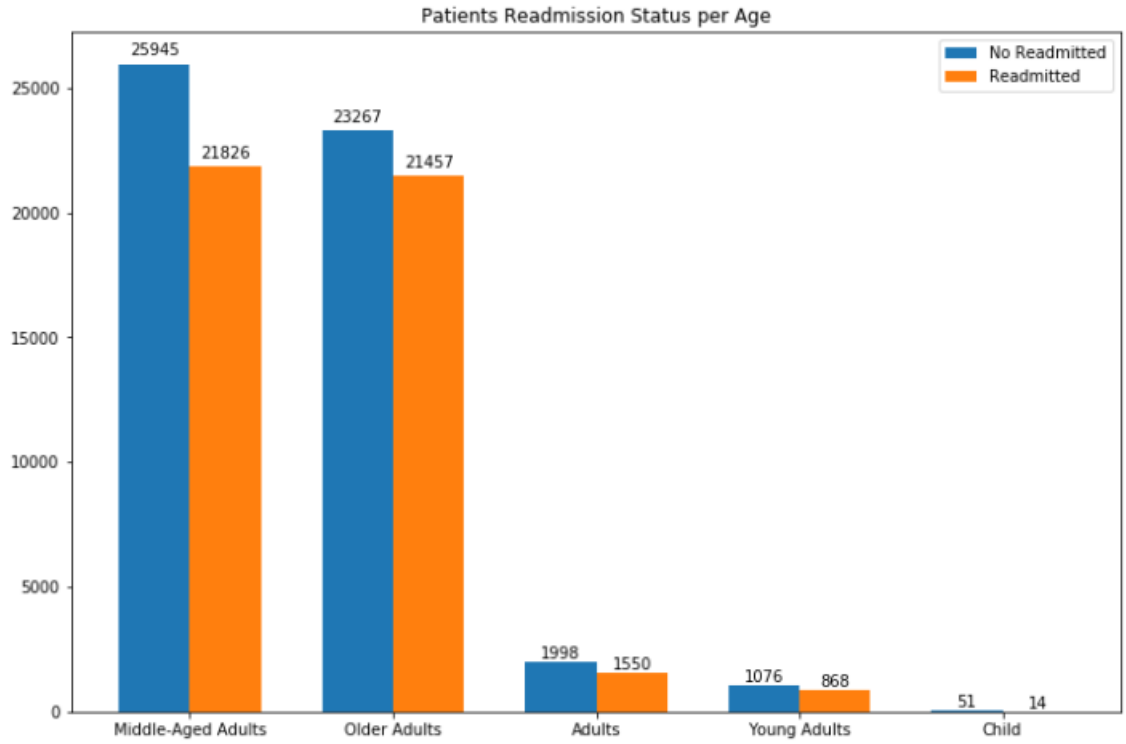

Figura 1. Gráfico de barras: comportamiento de reingreso vs la edad del paciente

No se logra observar una clara relación con la edad ya que la diferencia en cada uno de los rangos es muy similar, a excepción de la categoría "Child"; sin embargo, en ésta última solamente se dispone de un total de 65 observaciones, lo cual es muy poco en comparación a las demás categorías. Por lo tanto, la hipótesis que menciona que entre mayor edad tenga el paciente, éste es más probable que reingrese nuevamente al hospital se descarta.

Tiempo en el hospital

Se procede a observar si el tiempo en el hospital que pasa el paciente en la última vez que fue internado tiene relación con su futuro reingreso; por lo tanto, el siguiente gráfico de cajas nos muestra dicha relación.

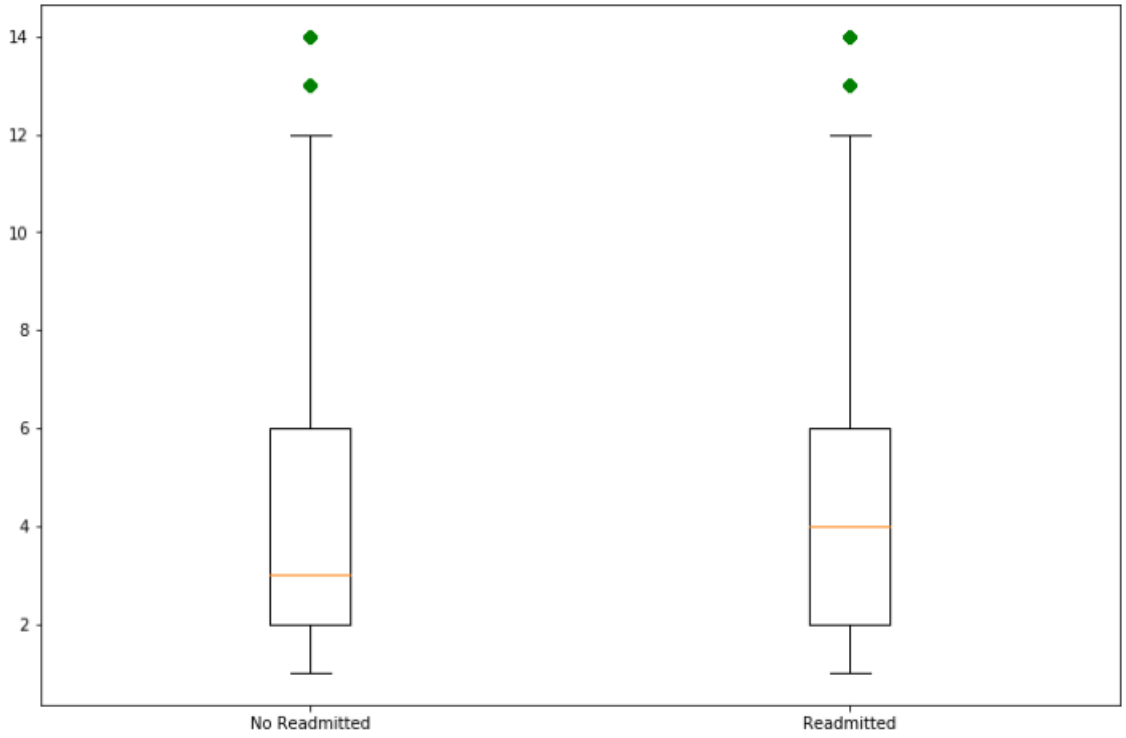

Figura 2. Gráfico de cajas: Tiempo en el hospital de los pacientes que serán o no reingresados

En éste caso, no parece que la cantidad de tiempo que el paciente pasó internado la última vez tenga influencia en si será o no reingresado por un problema similar, ya que no se observa una clara tendencia. La única y muy pequeña diferencia es que, el 50% de los pacientes que no fueron reingresados permanecieron en el hospital por al menos 3 días la última vez que fueron internados, mientras que el 50% de los pacientes que fueron reingresados permanecieron como máximo 4 días. Por lo tanto, se descarta la hipótesis que menciona que entre más tiempo permaneció internado el paciente en el hospital la última vez, es menos probable que reingrese por un problema similar.

Prescripción de medicamentos para la diabetes

En la figura 3, se observa si existe alguna relación entre el estado de readmisión del paciente y si éste posee una prescripción de medicamentos para la diabetes.

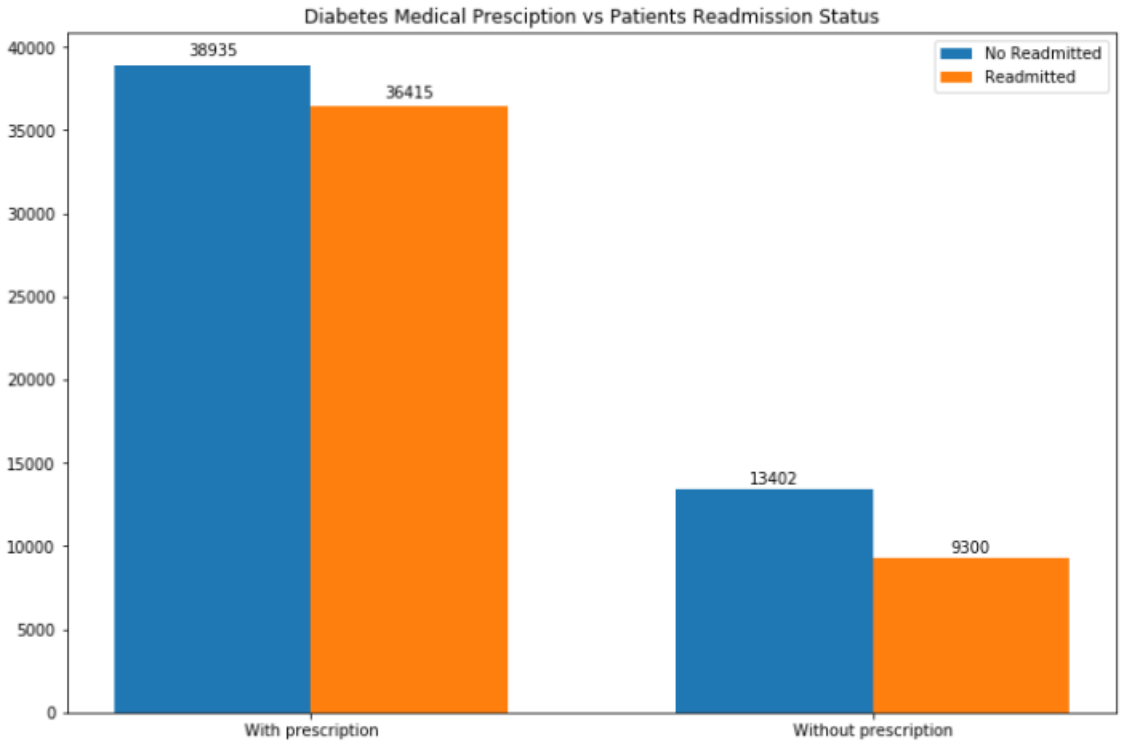

Figura 3. Gráfico de barras: Cantidad de pacientes que son o no reingresados con/sin prescripción de medicamentos para la diabetes

Con el gráfico de barras anterior se observa que, la gran mayoría de los pacientes poseen prescripción de medicamentos relacionados a la diabetes. Además, el porcentaje de personas con prescripción de medicamentos para la diabetes que no reingresan al hospital es de un 51.6%, mientras que el porcentaje de personas sin prescripción de medicamentos para la diabetes que no reingresan es de 59%. Por lo tanto, para un paciente sin prescripción de medicamentos para la diabetes es un 7% más probable que no sea reingresado, lo que descarta la hipótesis planteada sobre esta relación, ya que se observa lo contrario. Cabe destacar que, la diferencia de observaciones es muy grande, ya que hay 52 648 pacientes con prescripción de medicamentos de más con respecto a la cantidad de pacientes sin dicha prescripción.

Cambio en la medicación

En la figura 4, se visualiza si el cambio en la medicación del paciente tiene alguna influencia en si éste reingresa o no.

Se puede observar que un 55.5% de los pacientes que no se le cambió la medicación no reingresaron al hospital, mientras que el porcentaje de pacientes que se le cambió la medicación y no reingresaron al hospital es de un 50.9%. Por lo que, se observa un ligero favorecimiento si no se cambia la medicación, **fortaleciendo la hipótesis** que indica que *el cambio de medicación disminuye la probabilidad de que el paciente no reingrese al hospital*. Cabe destacar que, la cantidad de observaciones para cada uno de los dos casos es muy similiar, 52 774 observaciones de pacientes que no cambian la medicación y, 45 278 observaciones de pacientes que cambian de medicación.

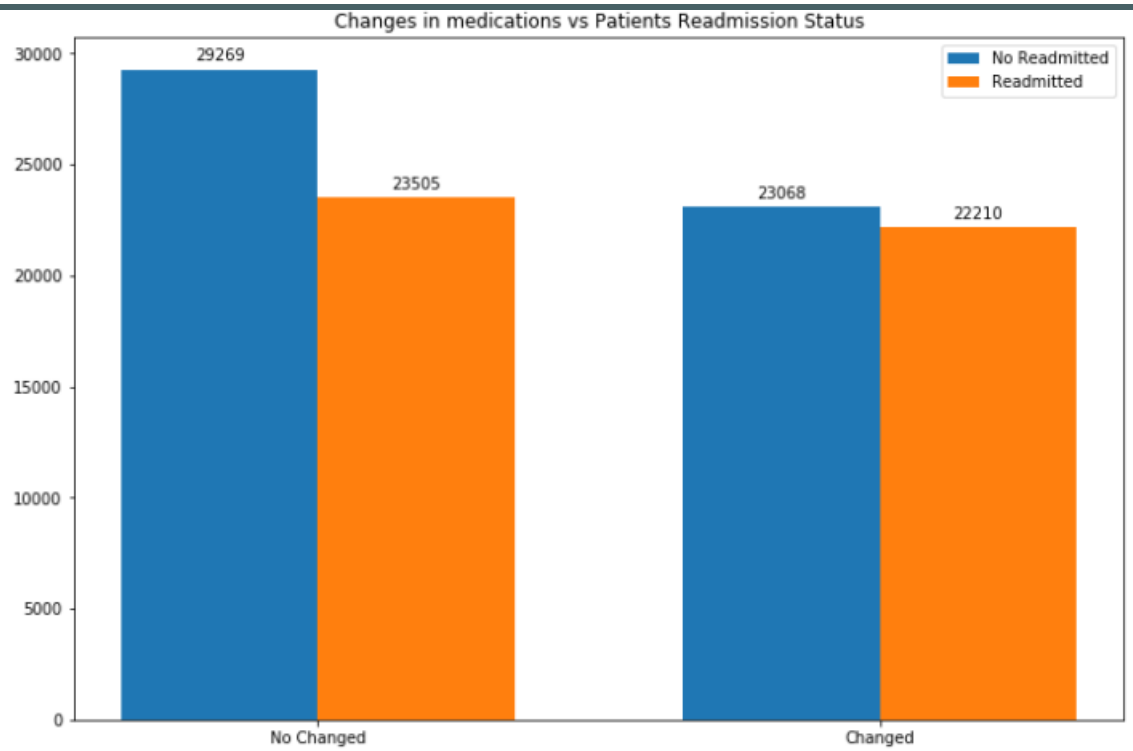

Figura 4. Gráfico de barras: cambios en la medicación de los pacientes vs su estado de readmisión

Modelos Construidos

Se construyeron modelos basados en los datos disponibes, utilizando tres tipos de algoritmos con el fin de determinar la solución que prediga con un mejor rendimiento si el cliente efectuará o no el pago de la cuota el siguiente mes.

Cabe destacar que, los atributos de *Diagnosis* están categorizados en códigos, con el fin de disminuir la cantidad de categorías posibles en estos atributos se hace uso de la tabla 2 del paper *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records*, donde se muestra el rango de códigos para distintos grupos.
(fuente: https://www.hindawi.com/journals/bmri/2014/781670/)

Además, se decide establecer solamente 5 valores para la edad de los pacientes, por lo tanto, las nuevas categorías abarcar las siguientes edades:

- Child: [0,10[
- Young Adults: [10,30[
- Adults: [30,40[
- Middle-Aged Adults: [40,70[
- Older Adults: [70,100[

Los algoritmos utilizados para la construcción de los modelos son los siguientes: *Random Forest*, *Support Vector Machine* y *K-Nearest Neighbors*. La siguiente tabla muestra los parámetros de rendimiento en las predicciones para los modelos obtenidos.

Tabla 1. Parámetros de rendimiento de los modelos

Model	Accuracy	Kappa
modelRF	61.2%	0.22
modelSVM	56.1%	0.09
ModelKNN	59.7%	0.16

Observando los resultados anteriores, el modelo con mejor rendimiento corresponde al modelo que utiliza el algoritmo de *Random Forest Classifier*, ya que posee un 61.2% de exactitud y el valor de kappa más alto.

A continuación, se muestra la importancia que le da el modelo a los 10 atributos más importantes:

1. feature 4: num_lab_procedures (0.140556)
2. feature 6: num_medications (0.119614)
3. feature 3: time_in_hospital (0.083281)

4. feature 10: diag_1 (0.069359)
5. feature 12: diag_3 (0.067728)
6. feature 11: diag_2 (0.066355)
7. feature 9: number_inpatient (0.055864)
8. feature 13: number_diagnoses (0.053160)
9. feature 5: num_procedures (0.052350)
10. feature 33: insulin (0.030321)

Conclusiones

Por medio del modelo desarrollado es posible predecir con una precisión de aproximadamente un 61.2% si el paciente reingresará o no al hospital. Por lo tanto, con los datos disponibles y el modelo con mejor rendimiento no es posible determinar de forma precisa si el paciente regresará o no a ser internado, ya que la confiabilidad es muy baja.

Por lo tanto, si se desea mejorar el rendimiento del modelo, se recomienda disponer de más atributos que brinden más información de los padecimientos cada vez que se ingresa al hospital y el estado del paciente al salir, ya que pueden mostrar más información de un posible reingreso.

Es importante mencionar que los atributos más importantes para el modelo fueron el número de procedimientos, número de medicamentos, el tiempo en el hospital y los diagnósticos. Por lo tanto, con el fin de determinar si el paciente reingresará o no, estos atributos son los que ofrecen más información.