Predicción de readmisiones clínicas en pacientes con diabetes

Carlos Alberto Arbeláez Giraldo

Santiago Velásquez Hernández

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Tutor
Ing. John Freddy Duitama Muñoz, Doctor (PhD)

Universidad de Antioquia
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2021

Cita	(Arbeláez Giraldo & Velásquez Hernández, 2021)
Referencia **Estilo APA 7 (2020)**	Arbeláez Giraldo, C. A., & Velásquez Hernández, S. (2021). *Predicción de readmisiones clínicas en pacientes con diabetes* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte II.

Tabla de contenido

Lista de Figuras

Lista de tablas

1. **Resumen ejecutivo**

La monografía aborda el problema de readmisiones clínicas en pacientes con diabetes mellitus. Este tipo de problema, en el contexto de EE. UU y que se encuentra mayormente en la literatura, comprende tres diferentes categorías (readmisión mayor a 30 días, readmisión menor a 30 días, sin readmisión). Actualmente, las predicciones de readmisiones clínicas u hospitalarias tienen como finalidad mejorar los tratamientos sobre los pacientes y reducir los costos asociados a la readmisión de estos.

La mayoría de los registros y características de los conjuntos de datos de readmisiones médicas coinciden en la información demográfica de las personas como lo son raza, sexo, edad, etc. También se presentan datos asociados a los diagnósticos hechos por los médicos tratantes y la medicación suministrada durante su estancia en el centro clínico o unidad de cuidados intensivos. El caso particular de la información usada en la monografía no dista de esto. Es válido mencionar que la información usada es multicentro, lo cual implica que no se posee el sesgo de una región u hospital especifico que es uno de los retos actuales de este dominio.

Las iteraciones para la obtención del modelo se abordaron en dos instancias. La primera de ellas tenía como propósito generar un modelo que sirviera de línea base para comprender que tan acertada puede ser la clasificación vista desde una perspectiva binaria, es decir paciente readmitido o no. En un segundo momento se extendió el problema a una clasificación multiclase en la cual pretendió darle especificidad al problema según como está en la literatura encontrada, además se usaron nuevos algoritmos para explorar que tan buenas son las predicciones (para este tipo de problemas) cuando se usan algoritmos más potentes a los normalmente conocidos.

Los resultados obtenidos en las dos fases de modelado fueron alentadores, todos los modelos obtenidos entregaron una precisión superior al 70%, siendo el XGboost el mejor modelo obtenido y el seleccionado con un 72% de precisión en el problema multiclase.

2. **Descripción del problema**

La readmisión hospitalaria se define como el regreso a un centro de salud de un paciente dado de alta en periodos superiores o inferiores a 30 días posteriores a su remisión o salida del hospital. La readmisión de un paciente indica que el tratamiento recibido previamente estuvo incompleto o no fue satisfactorio [1]. Esto se ve reflejado en sobre costos por volver a atender a un paciente dado de alta, solo en EE.UU las readmisiones hospitalarias evitables han costado cerca de 2.5 billones de dólares desde el 2012 [2].

En la plataforma Kaggle se plantea como problema predecir si un paciente con Diabetes mellitus tendrá una readmisión en los intervalos de 30 a 90 días.

2.1. Origen de los datos

La información pertenece al programa de recolección de datos de centros médicos Health Facts (Cerner Corporation, Kansas City, MO), este programa cuenta con una bodega de datos que guarda registros clínicos de hospitales en EE.UU; la información recolectada contiene registros de todas las instituciones pertenecientes a dicha organización. Contiene datos demográficos, diagnósticos, procedimientos realizados, códigos de laboratorio, medicamentos, etc.

En La información corresponde a 10 años (1999-2008) de registros de cuidados clínicos, que están distribuidos entre 130 hospitales de EE.UU: medio oeste (18 hospitales), noreste (58), sur(28), oeste (16). En total hay 74,036,643 millones de registros clínicos en Health Facts, de los cuales solo 17,880,231 millones son registros únicos. Para la extracción final del dataset se aplicaron los siguientes criterios para escoger los registros médicos de pacientes que aplican para el caso de estudio.

(1) Es un registro hospitalario (Admisión hospitalaria)
(2) Es un registro “diabético”, es decir, durante el cual se ingresó al sistema cualquier tipo de diabetes como diagnóstico.
(3) La estancia hospitalaria fue de al menos 1 día y como máximo 14 días.

(4) Durante el encuentro se realizaron pruebas de laboratorio.

(5) Se administraron medicamentos durante el encuentro.

Se identificaron 101,766 registros que cumplen con todos los criterios mencionados anteriormente y se utilizaron en análisis posteriores [3].

3. **Estado del arte**

En 2013 el Programa de Reducción de Readmisiones Hospitalarias (HRRP) de los Centros de Servicios de Medicare y Medicaid (CMS) comenzó a penalizar financieramente a los hospitales de EE. UU, con tasas excesivas de readmisión de 30 días [4]. Debido a esto y entre otras muchas acciones que tomaron las instituciones de salud, se incrementó la investigación para la creación de modelos que permitieran calcular la readmisión de pacientes.

Los modelos de machine learning son usados para la toma de decisiones médicas, desafortunadamente estos pueden estar sujetos a fallas en su desarrollo y validación, así como a limitaciones en su utilidad clínica [4]. Los modelos que se desarrollan aplican diferentes tipos de algoritmos, los cuales tienen un nivel de precisión que depende de la calidad de los datos y de la enfermedad asociada a la readmisión que se quiere predecir. En la tabla 1 se puede observar la popularidad de los algoritmos a la hora de abordar este tema.

Los problemas de readmisiones medicas son tan diversos como enfermedades se quiera predecir en el tema de readmisiones, a su vez, también están estrechamente ligados a la información que se posee sobre dicha enfermedad; ya que no todos los registros médicos electrónicos guardan la misma información y no toda la información está centralizada en una bodega de datos como la de Health Facts. Esto conlleva a que en algunos casos se prescinda, por ejemplo, de usar modelos de aprendizaje profundo debido a que la información disponible es demasiado superficial; también limita la posibilidad de hacer representaciones jerárquicas de dimensión alta que permitan tener un mejor panorama del problema y la forma en cómo se abordará [5].

TABLA 1

POPULARIDAD DE ALGORITMOS DE PARA PREDECIR READMISIÓN [6].

Familia de algoritmos	Algoritmo	Popularidad
Arboles	(1) Decision tree. (2) Random forest. (3) Boosted Tree Methods.	53%
Regresión logística regulada	(1) Lasso (L1 regularization) (2) Ridge regression (L2 regularization) (3) Elastic net.	28%
SVM		23%
Neural Networks:	(1) CNN. (2) RNN. (3) Deep stacking network. (4) Deep neural networks. (5) Ensemble DL methods.	33%
Otros algoritmos:	(1) Näive Bayes. (2) KNN. (3) Ensemble Methods. (4) Bayesian model Averagin	23%

3.1. Retos del dominio

En [1] se mencionan diferentes retos que afrontan los modelos de predicción de readmisiones hospitalarias. Los desafíos se dividen en dos grupos, el primero de ellos son los problemas asociados a los datos con los cuales se pretende modelar, y, el segundo hace alusión a los retos mismos del modelado.

3.1.1. Retos asociados a los datos

El desequilibrio y la localidad de los datos son dos sesgos comunes en los datos médicos, que se sabe que imponen un desafío significativo a los modelos predictivos.

El desequilibrio de datos se refiere a un fenómeno en el que los conjuntos de datos utilizados para entrenar un modelo predictivo tienen una distribución de clases desbalanceada. El desbalance de datos tiende a obligar a un clasificador a clasificar todas las muestras como normales, con el fin de satisfacer la función objetivo definida, como minimizar los errores de clasificación.

La solución comúnmente usada para el desequilibrio de los datos son los enfoques de muestreo, los cuales cambian las distribuciones de datos para equilibrar las muestras en diferentes. Normalmente se basan en eliminar muestras de la clase mayoritaria, repetir muestras de la clase minoritaria o crear muestras sintéticas para la clase minoritaria [1].

La localidad de los datos, por otro lado, está asociada a la forma como se distribuyen las muestras en el dataset con base al centro médico que aporta la información, al ser un conjunto de datos multicentro algunos hospitales pueden atender más individuos de cierta población étnica que otros, lo cual afecta el modelado y la implementación. A nivel de población, los datos para la predicción de readmisiones se pueden recopilar de un hospital local/regional. Sin embargo, el factor demográfico del cuerpo del paciente introduce naturalmente un sesgo de localidad. En [1] se propone como solución usar el aprendizaje federado, el cual permite que múltiples poseedores de datos entrenen un modelo en colaboración.

3.1.2. Retos asociados al modelado

El principal reto del modelado es la interpretabilidad del modelo, que se define como el grado en que la percepción humana puede explicar y comprender el comportamiento. Para aplicaciones médicas casi siempre se prefiere un modelo interpretable y transparente. En [1] proponen el esquema mostrado en la Fig. 1, que relaciona la interpretabilidad con la precisión de los modelos.

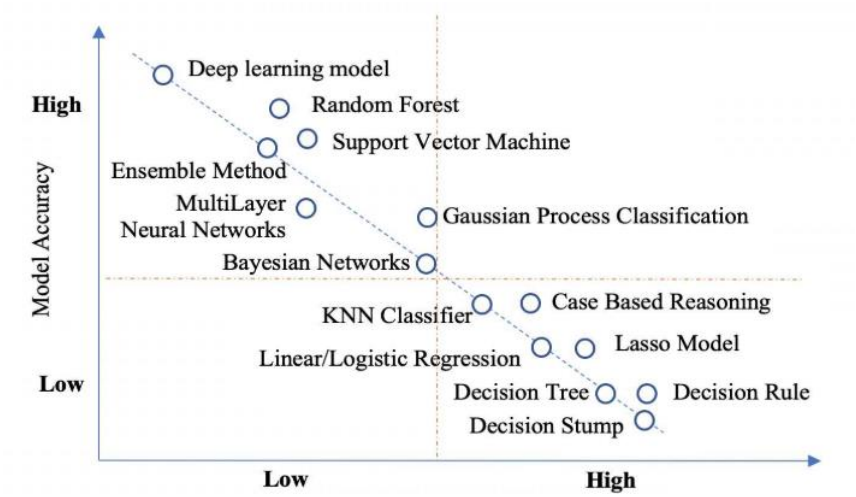


Fig. 1. Interpretabilidad de los modelos de machine learning [1]

En la Fig. 1 se evidencia que no hay un equilibrio entre la precisión y la interpretabilidad, lo que supone un reto para las partes involucradas en la toma de decisiones. Lo ideal es tener modelos transparentes y simples, estos son los que tienen una gran capacidad de interpretación y como se ve evidencia en la *Ilustración 1* los árboles/reglas de decisión y las regresiones lineales/logísticas, son los que entran en esta categoría y se utilizan con frecuencia en los dominios médicos [1].

3.2. Trabajos relacionados

Explorando la literatura, este dataset particular cuenta con numerosas y diferentes aproximaciones. Nos dimos cuenta en lecturas posteriores a la analítica de datos que coincidimos con muchos autores como [7][8][9] en la fase de exploración de los datos presentado en la sección 4, lo cual indica que el proceso de analítica realizado se abordó de manera correcta bajo las mismas inferencias sobre el dataset.

La mayoría coinciden en el uso de algoritmos como los presentados en la Tabla 1, pero difieren en el proceso exploratorio y creación de nuevas variables y esto se ve reflejado en los porcentajes de accuracy que obtienen en sus modelos.

4. **Análisis exploratorio de datos**

4.1. Búsqueda de valores nulos.

Al tomar el conjunto de datos lo primero que se realizó fue validar si existían valores nulos para poder usar la librería de python sweetviz, para tener una aproximación inicial de la estadística descriptiva del conjunto de datos.

Se encontró que gran parte de las columnas relacionadas a los medicamentos no contenían información relevante debido a que solo presentaban valores únicos o estaban altamente desbalanceadas, esto se puede ver en el anexo 1 donde se muestran las gráficas relacionadas al análisis exploratorio de los medicamentos. En consecuencia, como se muestra en la tabla 2 se eliminaron las columnas asociadas a los medicamentos.

TABLA 2

MEDICAMENTOS DESCARTADOS DEL DATASET

Medicamentos descartados			
metformin-pioglitazone	glyburide-metformin	miglitol	nateglinide
metformin-rosiglitazone	examide	acarbose	citoglipton
metformin-pioglitazone	troglitazone	tolbutamide	chlorpropamide
glipizide-metformin	tolazamide	acetohexamide	repaglinide
glimepiride-pioglitazone			

De igual forma, se encontró que las columnas *payer_code*, *encounter_id* y *number_diagnoses* no daban información importante para el modelo, ya que no entregan información relacionada al paciente.

4.2. Manipulación de datos

La edad y peso están representados en el dataset de forma categórica por medio de intervalos, para una mejor manipulación se decidió reemplazar por valores numéricos que haga referencia a

cada intervalo, se hizo como está en las tablas 3 y 4 usando un valor medio perteneciente a cada rango. Al ser pocos valores, y features se puede decir que hicimos un LabelEncoder manualmente.

TABLA 3

TRANSFORMACIÓN AGE.

Rango de edad	**Conversión**
[0-10)	5
[10-20)	15
[20-30)	25
[30-40)	35
[40-50)	45
[50-60)	55
[60-70)	65
[70-80)	75
[80-90)	85
[90-100)	95

TABLA 4

TRANSFORMACIÓN WEIGHT.

Rango de peso	**Conversión**
[0-25)	10
[25-50)	35
[50-75)	60
[75-100)	85
[100-125)	110
[125-150)	135
[150-175)	160
[175-200)	185
>200	205

De manera similar se procedió con la variable *Admission_type,* se reemplazaron los datos de texto por equivalentes a números para poder ser entregados al modelo.

TABLA 5

TRANSFORMACIÓN ADMISSION TYPE.

Admission_type	**Conversión**
Emergency	1
Emergency	2
Elective	3
New born	4
NaN	5
NaN	6
Trauma center	7
NaN	8

4.3. Análisis estadístico de variables

Cada variable se relacionó respecto a la variable objetivo. Como el problema se analizó en primera instancia como binario las gráficas y análisis mostrados a continuación corresponden a establecer si un usuario fue o no readmitido. Se encontró que se posee 54.864 usuarios no readmitidos y 46.902 que presentaron readmisión.

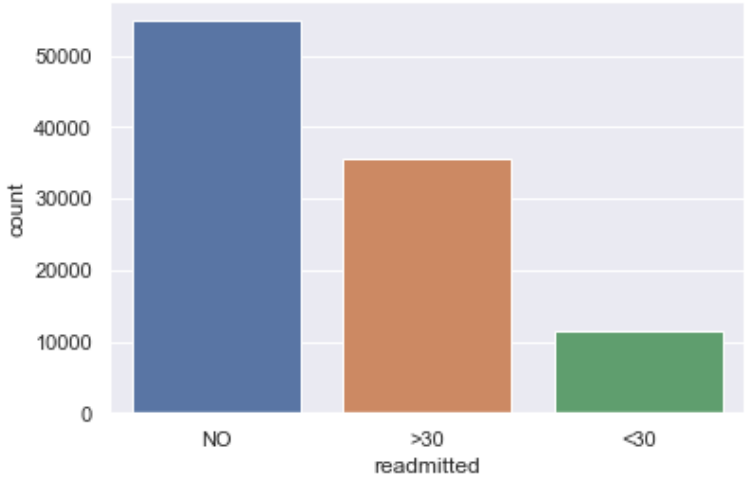


Fig. 2. Conteo de readmisiones.

El conjunto de datos es relativamente equilibrado respecto a las readmisiones como se puede ver en las Fig. 2 y 3. Incluso podría mejorarse con el aumento de datos útiles y comparar cómo es el rendimiento de los modelos predictivos con un conjunto de datos aumentado frente al conjunto de datos equilibrado original.

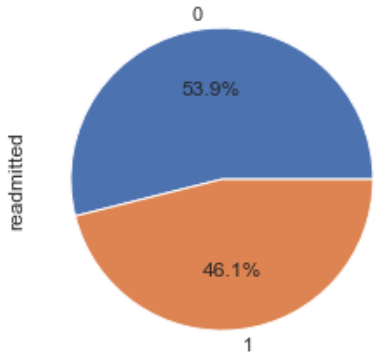


Fig. 3. Porcentaje de readmisiones.

La variable raza es de particular análisis ya que en EE. UU se posee gran diversidad racial y se desea inferir si la raza es un factor determinante para las readmisiones clínicas.

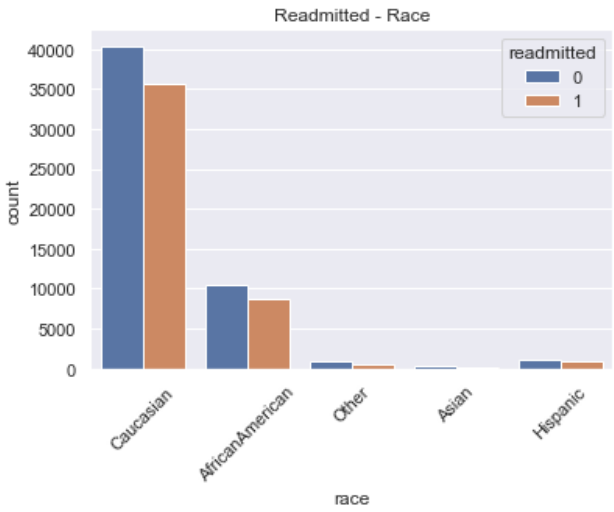


Fig. 4. Readmisiones por raza.

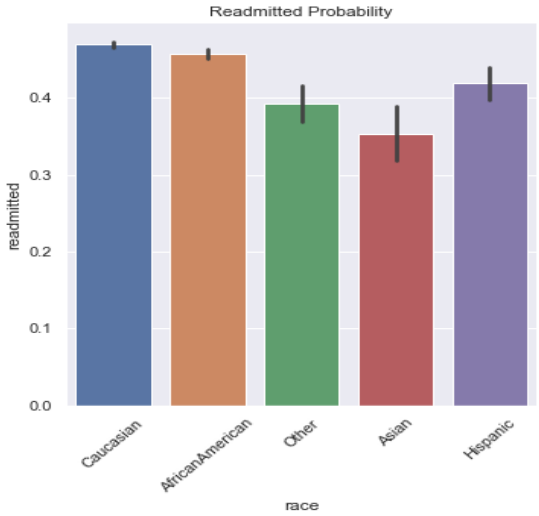


Fig. 5. Probabilidad de readmisiones por raza.

De las Fig. 4 y 5, se puede intuir que la raza quizás esté relacionada con la tasa de reingresos. La ilustración 5 muestra el cálculo de la probabilidad de ser readmitido en cada una de estas categorías, tomando como base el total de readmisiones de cada una.

Las Fig. 4 y 5 sugieren que, si eres asiático o perteneces a otra raza, podrías tener menos probabilidades de necesitar una nueva readmisión. Para ello se plantea la simulación de Montecarlo para validar esta hipótesis. Como hipótesis nula se define que toda raza tiene valores similares de readmisión, lo cual significa que la raza no afecta la probabilidad de readmisión.

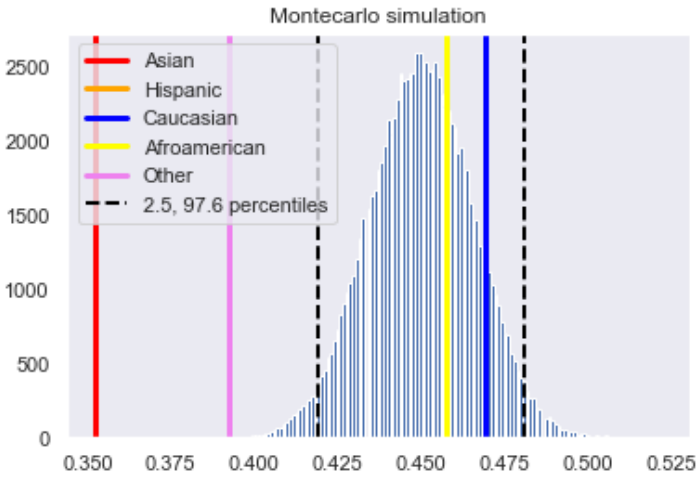


Fig. 6. Simulación de Montecarlo para la variable raza.

Como muestra la Fig. 6, los asiáticos y las otras razas están por fuera de las bandas de aceptación, por lo cual se rechaza la hipótesis nula. Esto significa que pacientes provenientes de estas razas tienen una menor probabilidad de ser readmitidos. Dicho resultado puede ser explicado desde los datos, ya que la proporción de razas dentro del conjunto de datos cuenta con mayor cantidad de información sobre pacientes caucásicos y afroamericanos. Se podría abordar desde un punto de vista cultural, pero para hacer dicha inferencia debería existir un mejor balanceo en esta variable.

Exploramos las demás variables para analizar qué tan balanceadas se encuentran respecto a la readmisión.

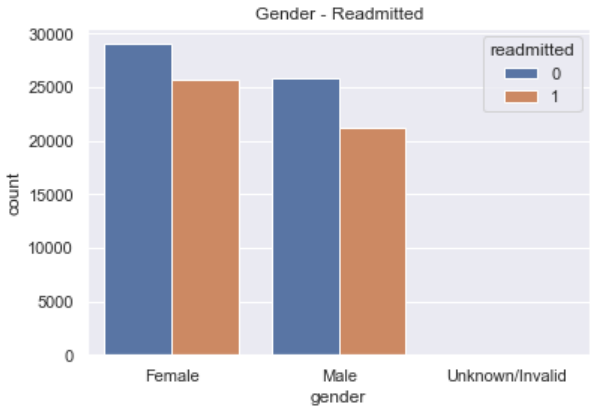


Fig. 7. Probabilidad de readmisión por género.

Encontramos que el género, abordado únicamente como masculino y femenino, no presenta diferencia significativa al momento de ser readmitido al hospital o no. La Fig. 7 evidencia que, aunque se observan menos hombres que mujeres, los valores son proporcionales a simple vista.

Tal como se ilustra en las Fig. 8 y 9, la probabilidad de readmisión es considerablemente inferior en las personas que se encuentran en el rango de edades [0-20). Así mismo, el peso (en libras) muestra que el sobrepeso o tener un bajo peso aumenta la probabilidad de ser readmitido.

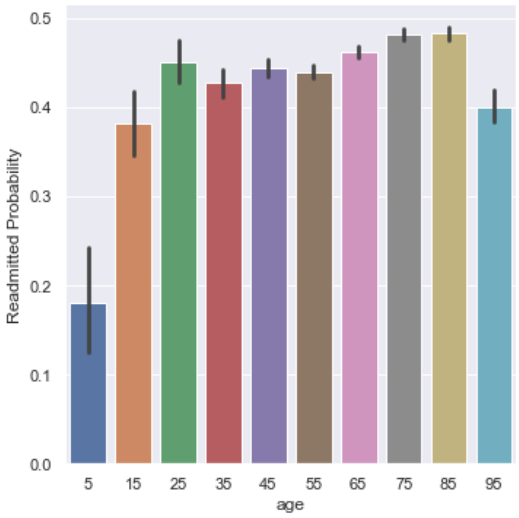


Fig. 8. Probabilidad de ser readmitido, edad.

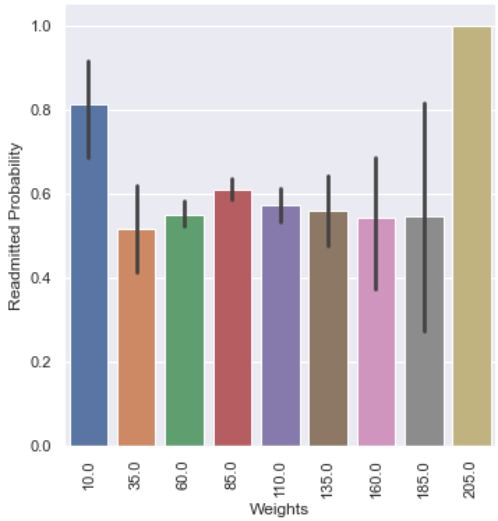


Fig. 9. Probabilidad de ser readmitido, peso

Se intuye que, si un paciente tiende a requerir más los servicios médicos (en el corto y en el largo plazo) este paciente también será más proclive a ser readmitido en comparación con otro paciente que haya requerido en menor medida los servicios médicos.

Con base a esto, se crearon dos variables nuevas, requests_1 y requests_10. La primera, es la suma de la cantidad de servicios hospitalarios solicitados por un usuario en 1 año, para la segunda se contabilizó la cantidad de apariciones del usuario en el dataset.

Las variables usadas para la construcción de la variable requests_1 son:

(1) *number_outpatient:* Número de visitas ambulatorias del paciente en el año anterior al registro.

(2) *number_emergency:* Número de visitas de emergencia del paciente en el año anterior al registro.

(3) *number_inpatient:* Número de visitas hospitalarias del paciente en el año anterior al registro.

En las Fig. 10 y 11, a la izquierda se puede ver una representación de la cantidad relativa de pacientes asociados a la cantidad de requerimientos de servicios hospitalarios.

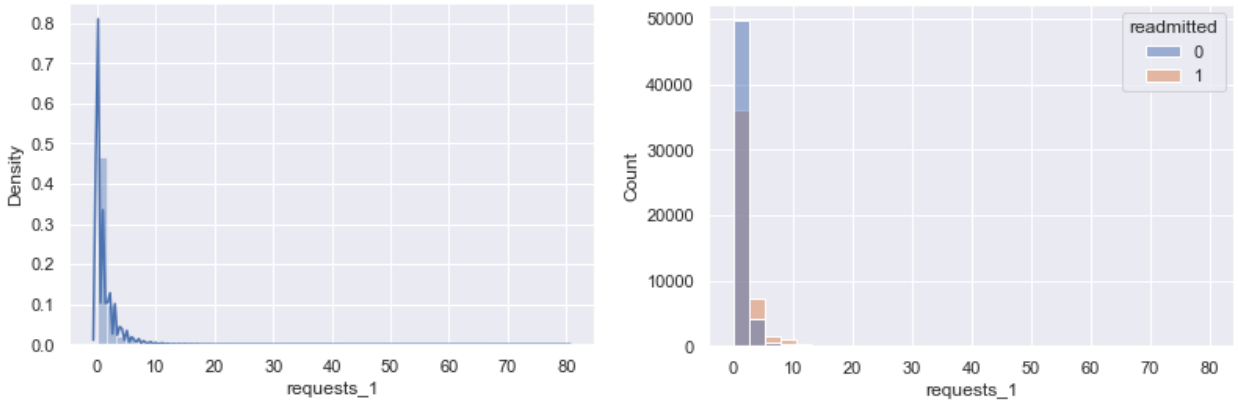


Fig. 10. Densidad y conteo de pacientes readmitidos en un año.

A la derecha se muestra la relación entre cantidad de servicios hospitalarios requeridos y readmisión. Se puede ver que hay una correlación positiva en este caso pues a mayor solicitud de servicios hospitalarios la tasa de readmitidos/no-admitidos crece de manera evidente. Esto indica que ambas variables podrían aportar información de cara a la toma de decisiones del modelo.

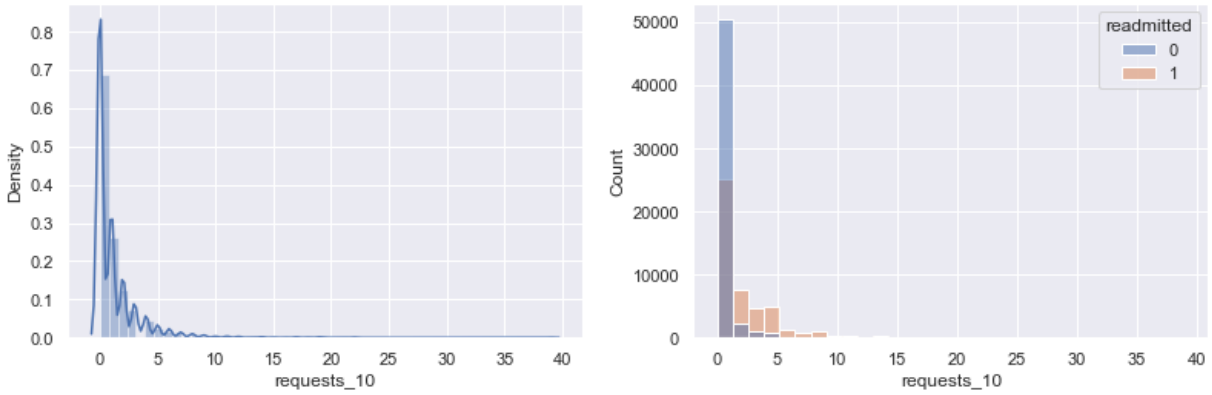


Fig. 11. Densidad y conteo de pacientes readmitidos en 10 años.

5. **Modelos de machine learning implementados**

Para entrenar los modelos se tuvo como estrategia implementar dos pipelines para transformar y reemplazar la información faltante a través del API de SimpleImputer de Sklearn, para los valores numéricos se usó la estrategia de la media y para los valores categóricos el valor más común.

Con base a la Tabla 1 se implementaron diferentes modelos de clasificación para dar solución al problema. Como se mencionó el problema se abordó en dos fases, en la primera se abordó problema binario para determinar si un usuario será readmitido o no.

5.1. Problema binario.

El primer escenario que se planteó fue si un paciente fuese readmitido o no en algún momento, un problema binario. Se plantearon algoritmos clásicos de clasificación haciendo uso de *scikitlearn* para evaluar el desempeño de cada uno. Para la evaluación se propusieron 3 métricas. el reporte de resultados se muestra en la Fig. 12.

	Model	Accuracy	Recall	F1
0	LogisticRegression	0.750952	0.617468	0.698964
1	RandomForest	0.797457	0.747568	0.775610
2	Decision Tree	0.728413	0.700595	0.707245
3	NN 1 Hidden Layer	0.752380	0.679832	0.719975
4	NN 3 Hidden Layers	0.765638	0.774866	0.755878
5	XGB	0.807112	0.717293	0.776913

Fig. 12. Modelos y métricas de desempeño.

Adicional a los resultados presentados en la Fig. 12, se incluyen análisis gráficos que respaldan estos resultados, como lo son las curvas de aprendizaje, curva ROC e importancia de las variables.

Para la interpretación de las curvas de aprendizaje es necesario prestar atención en 3 puntos que indican que el modelo se entrenó correctamente:

- La gráfica de entrenamiento se estabiliza, pues esto indica que las muestras fueron suficientes para entrenar el modelo de manera adecuada.
- La gráfica de validación se estabiliza y su valor estable es cercano al valor estable de la línea de entrenamiento.
- El periodo estable de ambas gráficas no excesivamente largo, de lo contrario sería un indicio de sobre entrenamiento.

Fig. 13. Learning Curve y performance del modelo XGBoost.

En este caso la gráfica de curvas de aprendizaje del modelo XGBoost indican que el modelo fue entrenado de manera confiable. Las curvas de aprendizaje de los otros modelos, para el caso binario pueden revisarse en el Anexo 2.

En el análisis de la curva ROC se busca un gráfico que sea lo más parecido posible a un gráfico de escalón. En todos los casos se obtuvo un resultado similar a un escalón, pero la mejor área bajo la curva se obtuvo con el modelo de XGBoost (Ilustración 14) con un AUC de 0.87. Esto indica, que este modelo tiene un ratio de verdaderos-positivos de 87%.

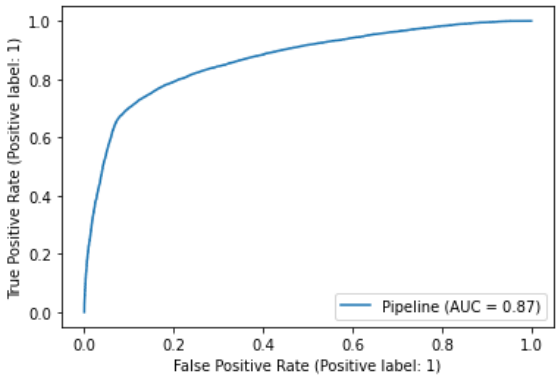


Fig. 14. ROC XGBoost.

También se extrae la importancia de las variables, en todos los modelos la ilustración 15 fue igual y se observa que la nueva variable creada *requests_1* tiene importancia significativa para el modelo, por lo cual la Fig. 10 es acertada al mostrar que entre mas servicios se soliciten a lo largo del año mas probabilidad de ser readmitido se tiene.

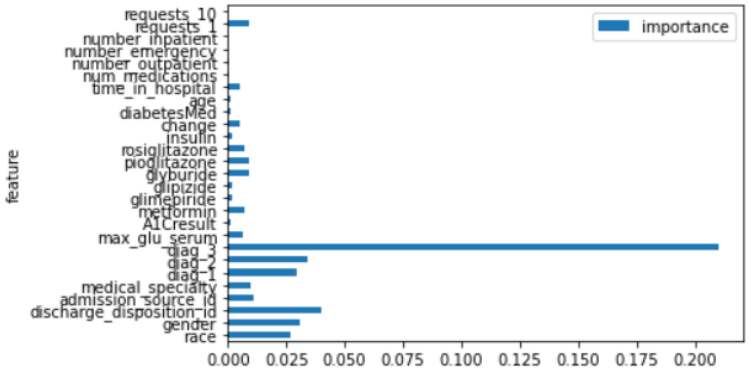


Fig. 15. Importancia de las variables.

Además, cobra relevancia las suposiciones hechas en el análisis exploratorio de la información, en el cual, los medicamentos que si se consideraron para el modelo, la raza, el género, los diagnósticos tienen importancia significativa al la hora de balancear la información.

5.2. Problema multiclase

Siguiendo una metodología similar al problema binario, se plantearon los mismos modelos para afrontar el problema multiclase.

	Model	Accuracy	Recall	F1
0	LogisticRegression	0.667154	0.667154	0.623715
1	RandomForest	0.707982	0.707982	0.668345
2	Decision Tree	0.621057	0.621057	0.616291
3	XGB	0.719404	0.719404	0.683599
4	NN 1 Hidden Layer	0.672423	0.672423	0.642421
5	NN 3 Hidden Layers	0.672423	0.672423	0.642421

Fig. 16. Modelos y métricas de desempeño problema multiclase.

De manera similar al problema binario, vemos en la Fig. 16 que el modelo basado en XGBoost fue entrenado de manera confiable, esto con base a los criterios mencionados para entender este gráfico en el problema de clasificación binaria.

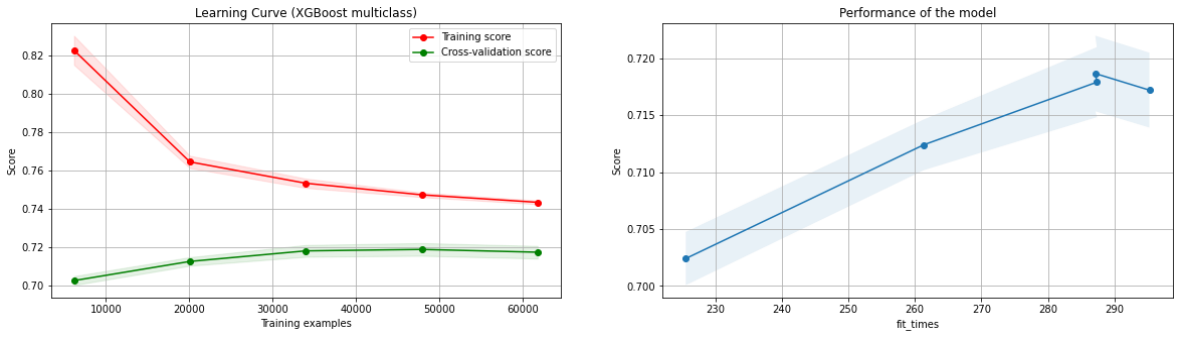


Fig. 17. Learning curve XGboost multiclase.

En el problema multiclase la escala de importancia de variables (Fig. 18) es mucho menor. A pesar de esto, se puede ver requests_1 tiene mucha mas importancia en este modelo y de igual manera, el porcentaje de importancia se distribuye mejor a lo largo de las features. Lo cual no se observa en la Fig. 15.

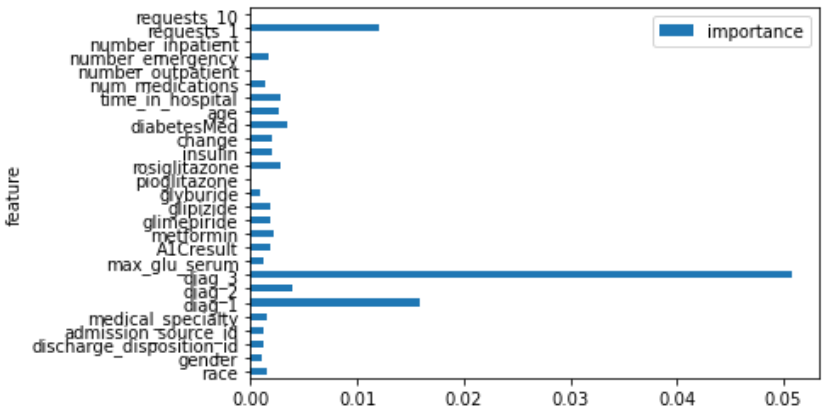


Fig. 18. Importancia de variables XGBoost Multiclase.

6. **Búsqueda de hiperparámetros.**

Se seleccionó el modelo basado en XGBoost para abordar este dominio en cuanto a un problema multiclase. Para realizar una búsqueda de los mejores valores de los hiperparámetros claves de este algoritmo. La selección de los valores propuestos se realizó basándose en reportes de modelos XGBoost en aplicaciones similares. A continuación, se relacionan los hiperparámetros con sus respectivos valores propuestos:

1. Máxima profundidad*:* [2, 4, 6, 8, 20]
2. Mínimo peso para nodo-hijo*:* [1, 2, 3, 4, 5]
3. Submuestras: [1, 3, 5]

Finalmente, el mejor modelo para el problema multiclase reporta los siguientes parámetros, respectivamente [6, 3, 1]. Las métricas de este modelo son:

	Model	Accuracy	Recall	F1
0	XGB_Tunned	0.720934	0.720934	0.685635

Fig. 19. Mejores parámetros XGBoost.

7. Entender las predicciones del modelo

El modelo XGBoost (multiclase) seleccionado tiene en cuenta diversas variables al momento de tomar la decisión de readmisión de un paciente como se muestra en la Fig. 18. La particularidad de este modelo es que reduce el sesgo demográfico en variables como el sexo y la raza, lo cual contribuye a disminuir el problema de la localidad de los datos.

Adicional este modelo, a diferencia de la iteración inicial y otros trabajos como [7] entrega como resultado si el paciente será readmitido en plazos inferiores a 30 días, mayores a 30 días o no será readmitido. A partir de la predicción del modelo, el personal encargado puede soportar sus decisiones respecto al dado de alta de un paciente, verificando las variables mostradas en la Fig.18. Valdría la pena tener una representación grafica a partir de una aplicación donde se muestre el estado de las variables mas representativas y facilitar el entendimiento y las decisiones del personal que se beneficia de esta aproximación que hemos realizado.

8. Conclusiones

En este trabajo se revisó el desempeño de diferentes algoritmos de machine learning después de procesos rigurosos de análisis y exploración de los datos. En la predicción de la readmisión en dos escenarios: un problema binario (paciente readmitido o no) y uno multiclase (paciente readmitido antes de 30 días, después o no readmitido). El desempeño en ambos escenarios fue bueno y, en general, los modelos tuvieron unos resultados cercanos entre sí; con técnicas como *gridsearch* para ajustar los hiper parámetros y un feature engineering cuidadoso el modelo escogido pudo mejorar su rendimiento. Lo que esto demuestra es que este se beneficia de ser tratado con técnicas de machine learning y ciencia de datos, y demuestra un camino claro para el estudio de otras poblaciones en este mismo contexto.

La creación de nuevos features a partir del dataset usado tiene un impacto considerable en el porcentaje de accuracy, en nuestro caso creamos la variable requests_1 asociada a la cantidad de servicios hospitalarios solicitados por un usuario. En trabajos como [7] crearon 3 variables

adicionales relacionadas a los medicamentos y la cantidad de cambios en los mismos; cabe aclarar que estos autores solo abordaron el problema desde la clasificación binaria; sin embargo lograron porcentajes de precisión por encima del 87% con todos los modelos que usaron.

8.1. Que se puede esperar del modelo

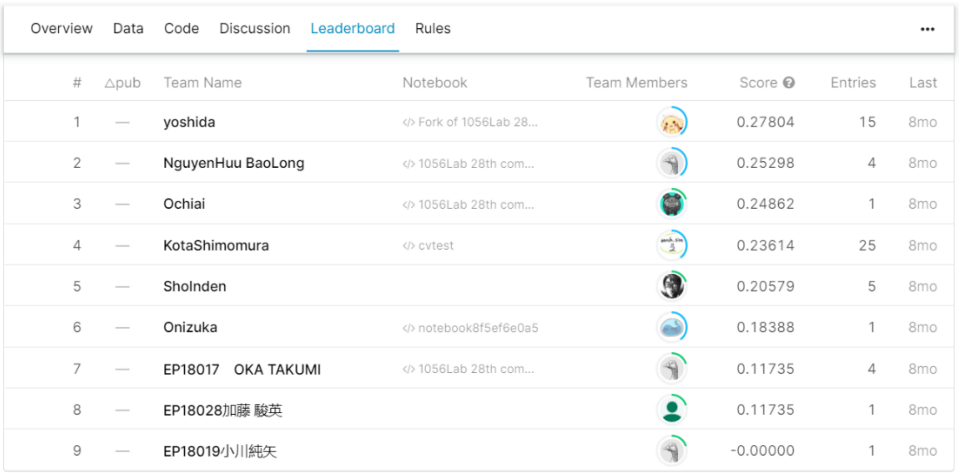
El modelo puede brindar información confiable sobre las posibles readmisiones clínicas en pacientes con diabetes. Sin embargo, este debe ser usado en compañía de médicos que permitan determinar si el resultado para X paciente es correcto basado en los datos que se le han suministrado. Solo la predicción del modelo no es un factor determinante para indicar si un paciente será readmitido, se debe tener siempre la ayuda del personal de salud encargado.

8.2. Limitaciones del modelo

La principal limitación del modelo, como en muchos es que está sesgado a un tipo de población. En este caso la información usada para el entrenamiento proviene de EE.UU, por lo cual dicho modelo no sería válido en un contexto como el de Colombia. Para poder implementarlo se requiere recolectar información lo más similar posible sobre pacientes a lo largo del territorio y hacer nuevamente el proceso de exploratorio de la información para así determinar que variables son las que tienen peso a la hora de entrenar el modelo para el contexto de Colombia.

8.3. Comparación de resultados

Las métricas resultantes de Accuracy, Recall y F1 para los modelos seleccionados y afinados con los métodos que aquí se describieron sugieren que son modelos candidatos para tratar el tema de readmisión con estas poblaciones en el contexto de EE. UU. Sin embargo, se quiso comparar de manera directa con otras soluciones que se han propuesto para el mismo problema. Nos remitimos a la página oficial de la competencia desde donde este dataset fue extraído y los usuarios con mejores métricas se presentan en la Fig. 20.



Overview Data Code Discussion Leaderboard Rules

#	△pub	Team Name	Notebook	Team Members	Score	Entries	Last
1	—	yoshida	Fork of 1056Lab 28...		0.27804	15	8mo
2	—	NguyenHuu BaoLong	1056Lab 28th com...		0.25298	4	8mo
3	—	Ochiai	1056Lab 28th com...		0.24862	1	8mo
4	—	KotaShimomura	cvtest		0.23614	25	8mo
5	—	ShoInden			0.20579	5	8mo
6	—	Onizuka	notebook8f5ef6e0a5		0.18388	1	8mo
7	—	EP18017 OKA TAKUMI	1056Lab 28th com...		0.11735	4	8mo
8	—	EP18028加藤 駿英			0.11735	1	8mo
9	—	EP18019小川純矢			-0.00000	1	8mo

Fig. 20. Resultados Kaggle.

La métrica de evaluación es *Cohen Kappa*. Esta métrica se enfoca en la confiabilidad del desempeño [7]. Debido a que la competencia estaba cerrada para el momento en el que se desarrolló este trabajo y por ende no se pudo someter directamente los modelos propuestos en Kaggle, se implementó el módulo <*metrics.cohen_kappa*> de scikit-learn para la evaluación del modelo. El valor Cohen Kappa para el modelo XGBoost afinado con búsqueda de hiperparámetros para el problema multiclase es: 0.463938.

Es superior a los mejores puntajes de la competencia. Sin embargo, hay que tener en cuenta que las métricas se calcularon con diferentes datasets de evaluación. En este trabajo se realizó con el 30% del dataset de test y los equipos que sometieron su modelo en Kaggle fueron evaluados con un dataset de evaluación que ahora se encuentra oculto.

9. Bibliografía

[1] S. Wang and X. Zhu, "Predictive Modeling of Hospital Readmission: Challenges and Solutions," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, pp. 1–20, 2021, doi: 10.1109/TCBB.2021.3089682.

[2] Y. W. Lin, Y. Zhou, F. Faghri, M. J. Shaw, and R. H. Campbell, "Analysis and Prediction of Unplanned Intensive Care Unit Readmission using Recurrent Neural Networks with Long Short-Term Memory," *bioRxiv*, vol. 742, pp. 1–22, 2018, doi: 10.1101/385518.

[3] B. Strack *et al.*, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *Biomed Res. Int.*, vol. 2014, 2014, doi: 10.1155/2014/781670.

[4] L. Grossman Liu *et al.*, "Published models that predict hospital readmission: A critical appraisal," *BMJ Open*, vol. 11, no. 8, pp. 1–15, 2021, doi: 10.1136/bmjopen-2020-044964.

[5] P. Wolff, M. Grana, S. A. Ríos, and M. B. Yarza, "Machine Learning Readmission Risk Modeling: A Pediatric Case Study," *Biomed Res. Int.*, vol. 2019, 2019, doi: 10.1155/2019/8532892.

[6] Y. Huang, A. Talwar, S. Chatterjee, and R. R. Aparasu, "Application of machine learning in predicting hospital readmissions: a scoping review of the literature," 2021.

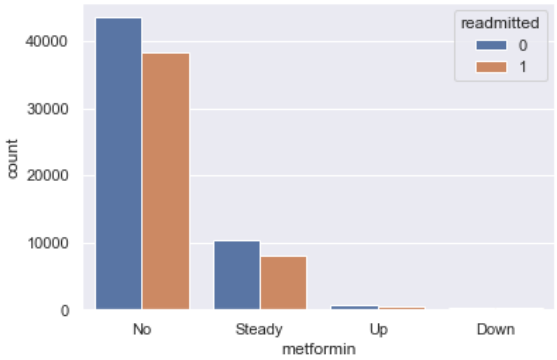
[7] L. Alturki, K. Aloraini, A. Aldughayshim, and S. Albahli, "Predictors of Readmissions and Length of Stay," *2019 IEEE/ACS 16th Int. Conf. Comput. Syst. Appl.*, pp. 1–8, 2019.

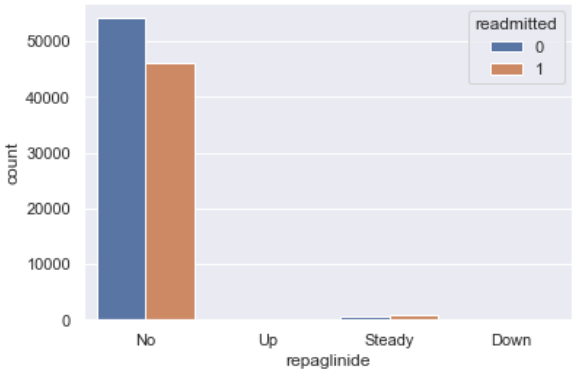
[8] M. Jia and F. Tian, "Readmission Prediction of Diabetic based on Convolutional Neural Networks," *2019 IEEE 5th Int. Conf. Comput. Commun. ICCC 2019*, pp. 1990–1994, 2019, doi: 10.1109/ICCC47050.2019.9064477.

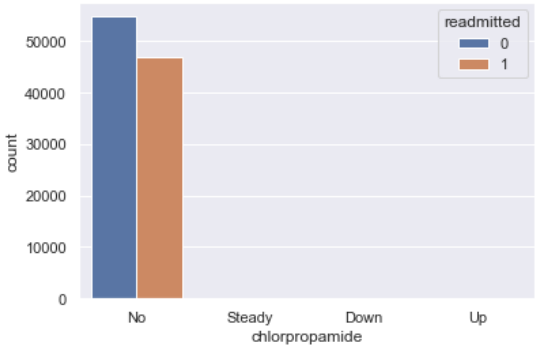
[9] R. Pawar *et al.*, "Diabetes Readmission Prediction using Distributed and Collaborative Paradigms," *1st Int. Conf. Data Sci. Anal. PuneCon 2018 - Proc.*, 2018, doi: 10.1109/PUNECON.2018.8745374.

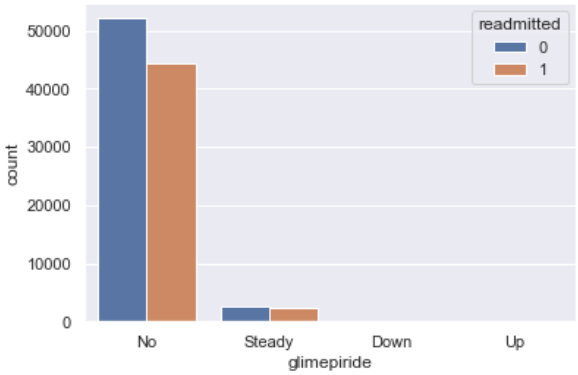
10. Anexos

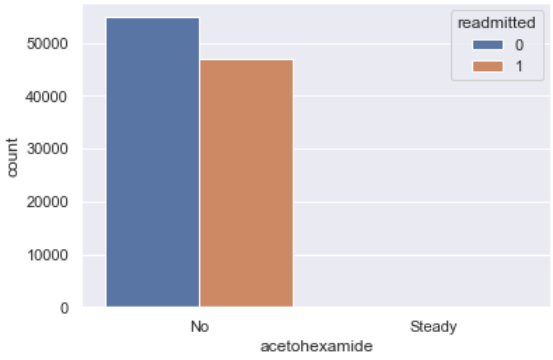
1. Gráficos de análisis exploratorio de medicamentos.

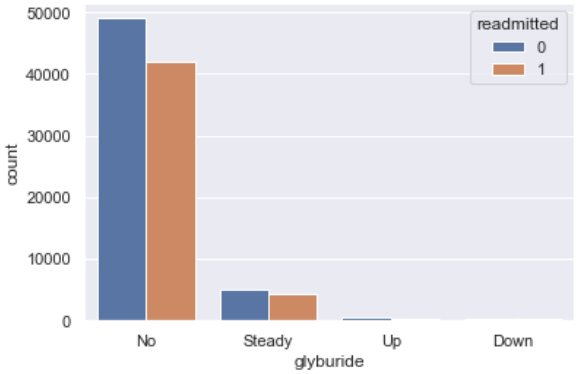


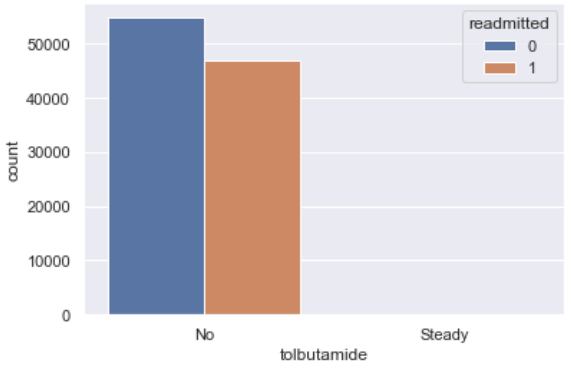


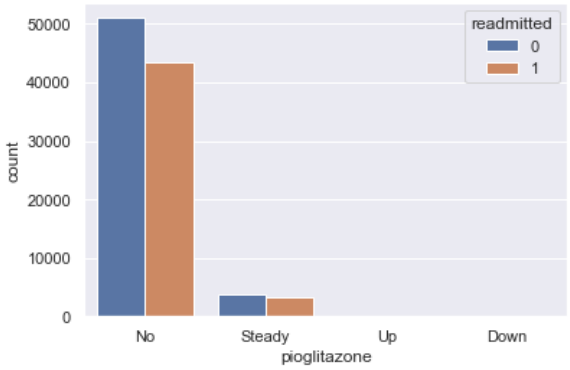


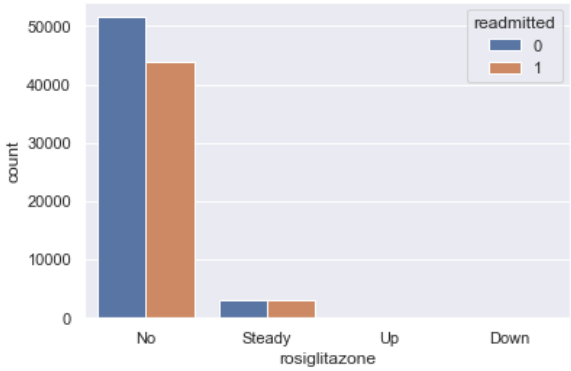


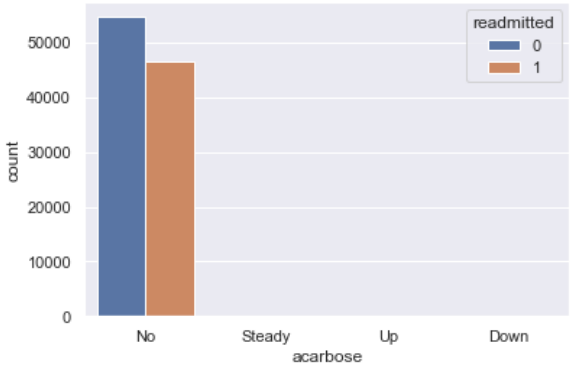


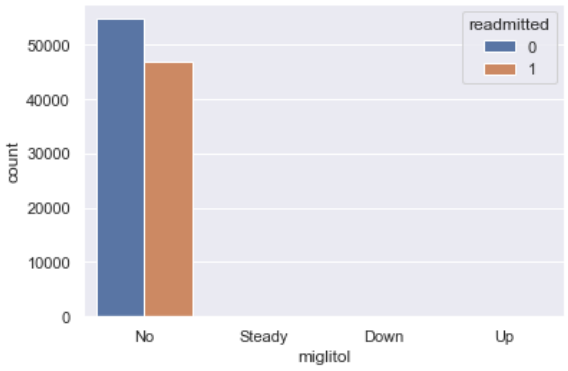


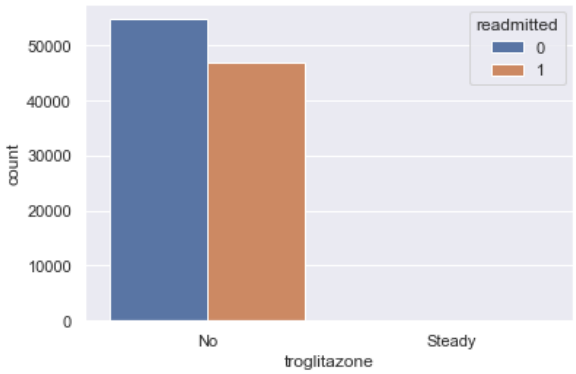

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
tolbutamide

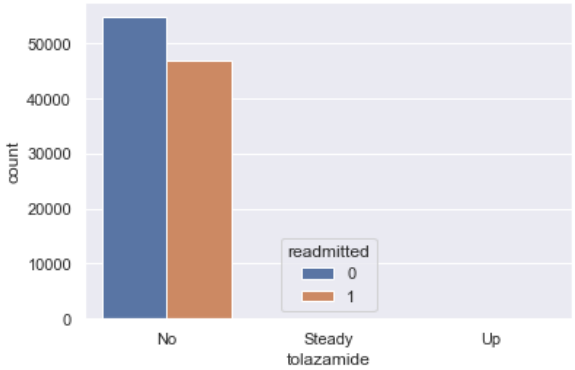

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
Up
Down
pioglitazone

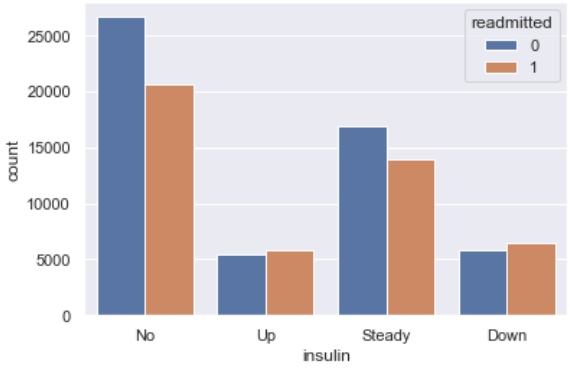

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
Up
Down
rosiglitazone

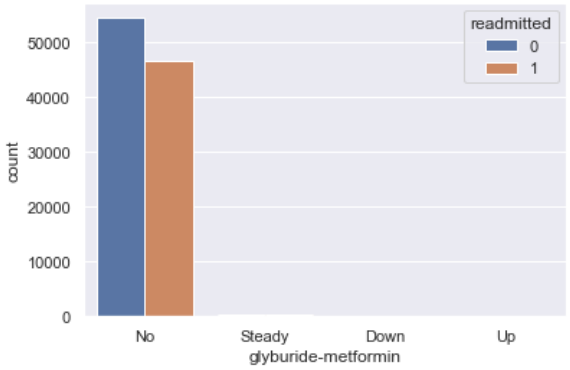

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
Up
Down
acarbose

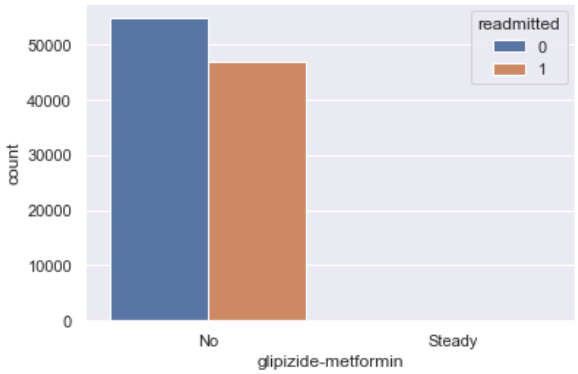

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
Down
Up
miglitol

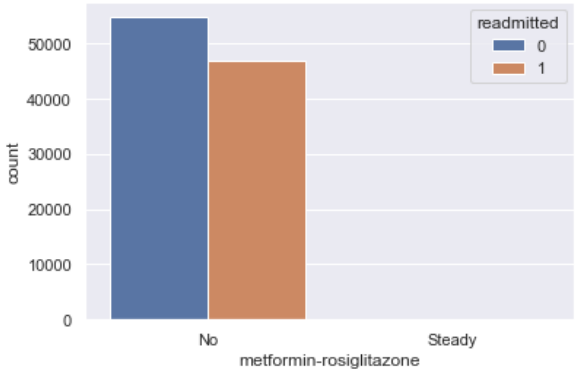

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
troglitazone

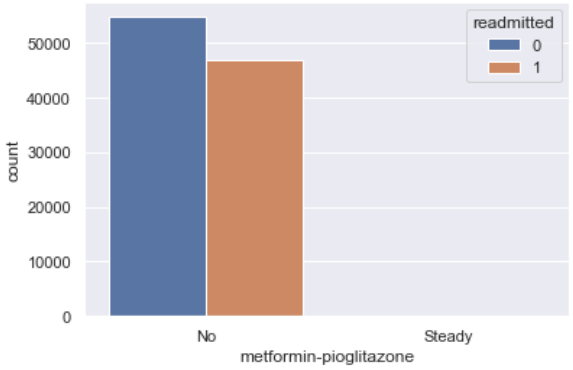

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
Up
tolazamide


readmitted
0
1
count
25000
20000
15000
10000
5000
0
No
Up
Steady
Down
insulin

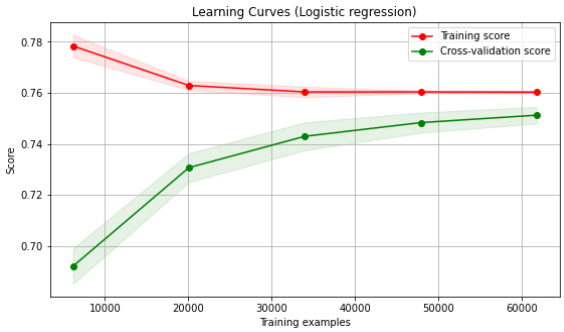

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
Down
Up
glyburide-metformin

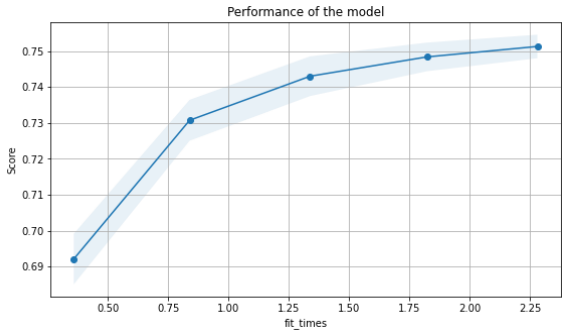

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
glipizide-metformin

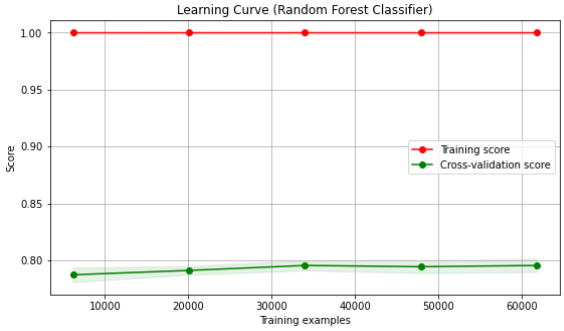

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
metformin-rosiglitazone

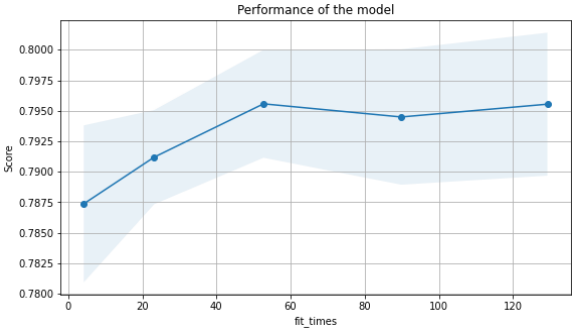

readmitted
0
1
count
50000
40000
30000
20000
10000
0
No
Steady
metformin-pioglitazone

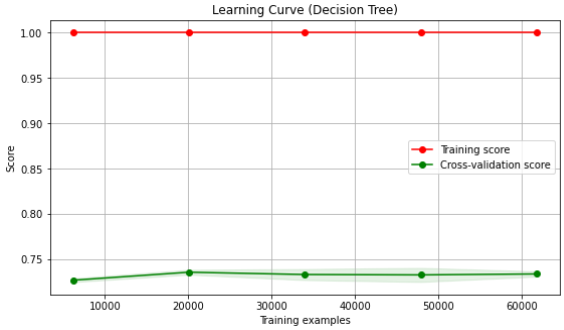
2. Learning curve otros modelos problema binario.

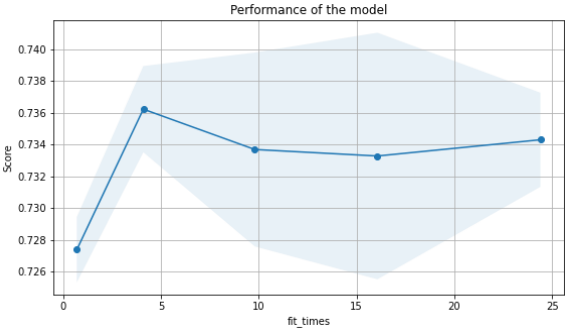




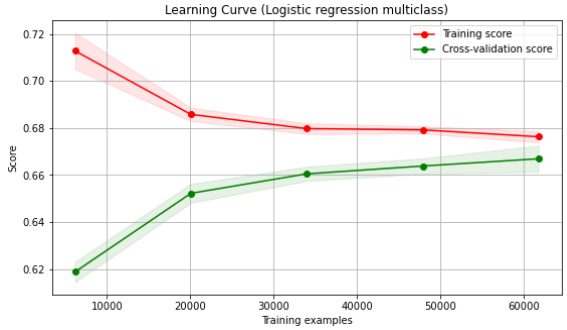


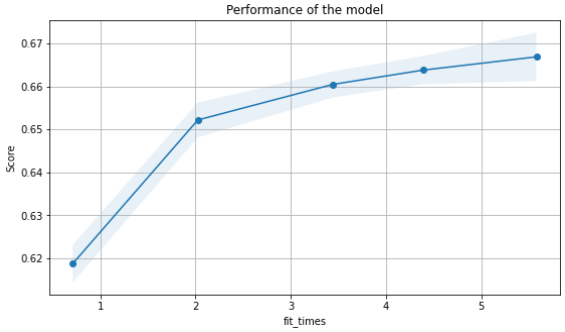


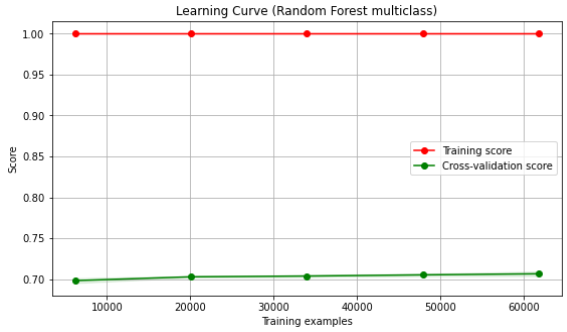


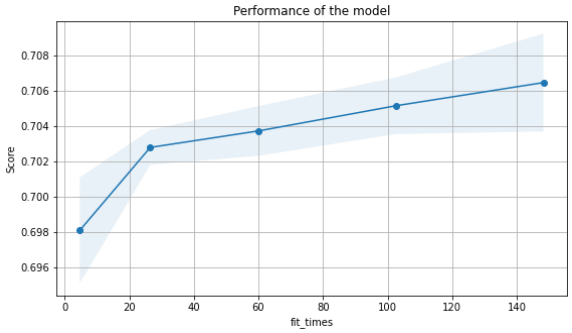


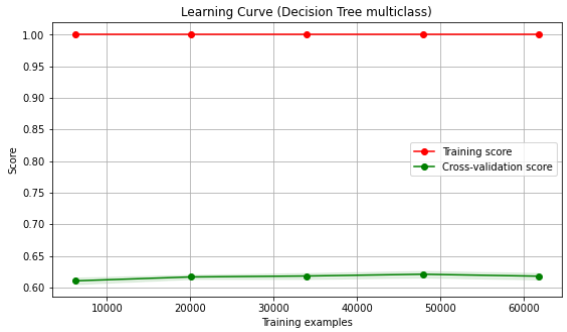
3. Learning curve otros modelos problema multiclase.

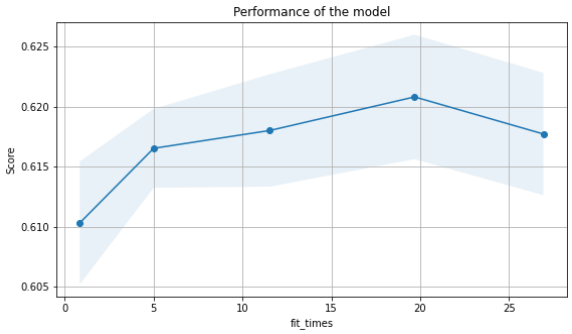


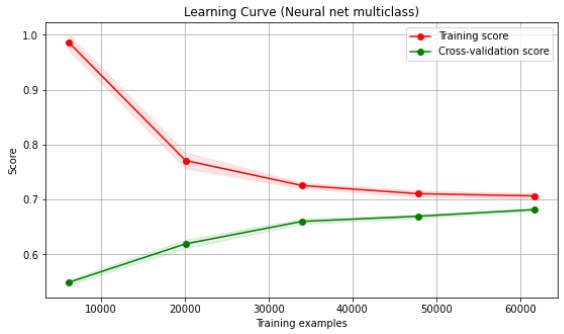


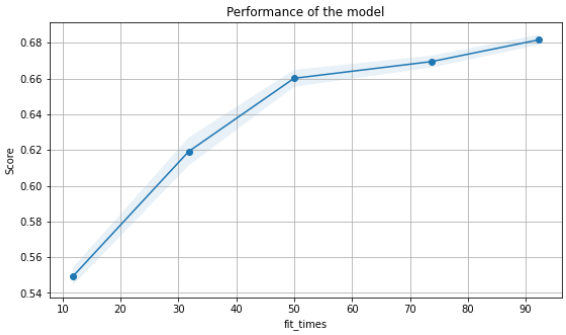




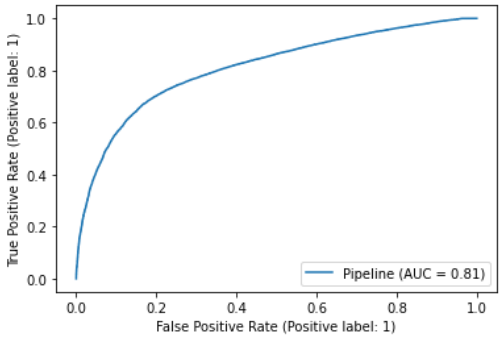




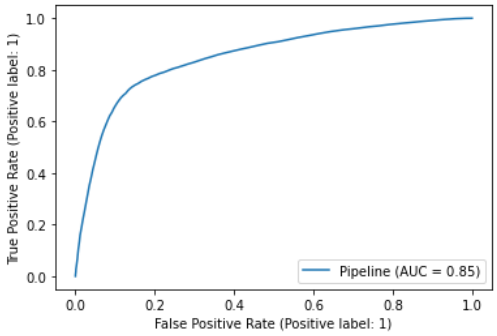




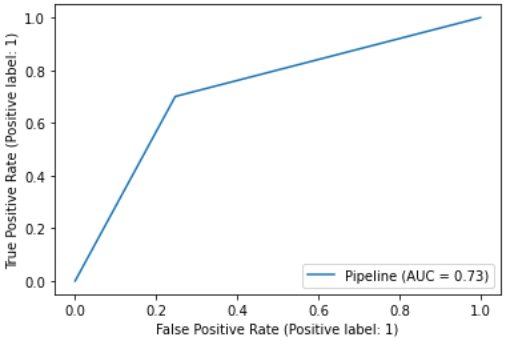
4. Curvas ROC modelos clasificación binaria.



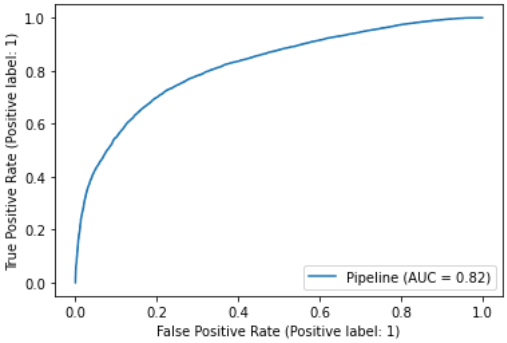
ROC Regresión logística.



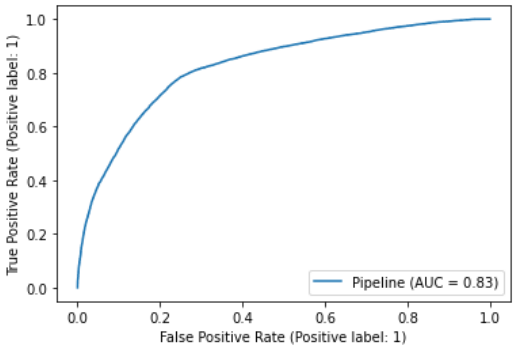
ROC Random forest.



ROC Desicion tree.



ROC Neural network 1 hidden layer.



ROC Neural network 3 hidden layers.