Trabajo final del itinerario de Data Science de la IT Academy

Esteban Piliponsky

Mentora:

Aina Palacios

Presentación

El objetivo de este proyecto es predecir si pacientes internados (ingresados en España) en Terapia Intensiva estarán allí más de una semana, o antes de los siete días dejarán esa condición, ya sea por alta médica, defunción u otra causa. Esa predicción se hará en base a la edad del paciente, el sexo y las mediciones de signos vitales que se le toman al mismo durante los dos primeros días de internación.

La base de datos elegida para realizar el proyecto final al itinerario de Data Science contiene 615.584 registros de medición de signos vitales a pacientes internados en terapia intensiva (UCI en España) de una clínica de la provincia de Tucumán, Argentina, entre enero y noviembre de 2022.

La base de datos pertenece a la empresa Medexis, encargada de crear y manejar sistemas informáticos para la gestión de empresas de salud. El objetivo principal de esta empresa es facilitar la carga, archivo y consulta de información médica, pero, como sucede en prácticamente todos los ámbitos de la sociedad actual, esta tarea genera una gran cantidad de datos que pueden servir para su estudio y análisis, como se ha buscado demostrar aquí.

Descripción de los datos y metodología

El Data Set con el que se cuenta permite realizar diferentes tipos de análisis y aplicar diversos modelos predictivos. De las posibilidades existente se ha elegido, luego de una charla con el cliente –Medexis, en este caso– un estudio de los parámetros vitales medidos a pacientes en Terapia Intensiva a partir de dos dudas que sirven de disparadores: ¿Existe una correlación entre los parámetros básicos (o alguno/s de ellos) del paciente y la cantidad de días que este permanece internado? Y de ser así, ¿puede predecirse a partir de estos números cuántos días estará en esta situación?

Para realizar este trabajo se utilizó Python 3.0 en el entorno de Jupyter Notebook. La lectura y manipulación de los datos se realizó con las librerías Numpy y Pandas, los gráficos con Matplotlib y Seasborn, y los modelos predictivos con Scikit-learn.

La base de datos de la que se dispone es considerada un muestreo representativo de una comunidad a lo largo de prácticamente un año. Cada registro cuenta con nueve columnas: el ID del paciente (que por lógicas razones de privacidad es etiquetado con un número), fecha de ingreso a la internación, fecha de nacimiento, sexo, días de internación, condición al cierra de la internación, hora y fecha en que se toma el registro, tipo de registro y resultado de este.

Un primer análisis arroja como resultados que la base de datos contiene la información de 5679 pacientes que estuvieron internados entre 1 y 136 días. El 63,9% (3629) está entre 1 y 2 días en esa condición. Sin embargo, la experiencia muestra que el estado general de un paciente que permanece hasta 2 días en la terapia no es complejo de predecir por el personal médico. El interés radica principalmente en poder anticipar la cantidad de días que permanecerá si supera las dos primeras jornadas de internación.

Es por ello que se decide hacer un recorte importante de los datos para hacer este análisis. Se tomó sólo los pacientes que hayan estado internados por 3 o más días y se analizó los resultados de los parámetros vitales que se les ha tomado la primera vez y luego las 3 primeras muestras que se tomó durante el segundo día de internación.

Con esa información se desarrolló la técnica de Machine Learning supervisado, probando diferentes modelos predictivos de clasificación. Se dividió a la población en dos grupos: los pacientes internados hasta por una semana (de 3 a 7 días), y los que superan ese tiempo. El objetivo fue poder anticipar a qué grupo pertenece cada caso.

Metodológicamente el estudio se desarrolló en tres archivos de Jupyter Notebook diferentes: Primero uno de limpieza, preparación y reacomodación de los datos para ser analizados luego con modelos de entrenamiento y predicción de IA.

El primer paso se muestra dentro del proceso común de análisis de un Data Set: los datos contienen errores de llenado que se deben limpiar. A su vez, los tipos de datos fueron reconvertidos según su formato natural, como las fechas, o según las necesidades de nuestro análisis, como transformar de tipo objeto a tipo categórico el sexo del paciente, o de tipo numérico a tipo objeto el ID del mismo. Además, fue necesario crear nuevas columnas, una de edad y otra de día de internación en la que se produce la medición del parámetro.

El segundo paso se muestra como específico de la forma en la que está organizado el Data Set, que demanda un proceso de reorganización de los datos para su análisis. Cada registro del Data Set original representa el resultado de la toma de uno de los siete parámetros medidos ('Frecuencia Cardiaca', 'Frecuencia respiratoria', 'Presión Arterial

Máxima', 'Presión Arterial Mínima', 'Temperatura', 'Saturación O2', 'Presión Arterial Media') en un momento determinado.

Para poder aplicar el proceso deseado fue necesario crear un nuevo Data Set en el que cada registro pertenezca a cada paciente y en el que se sumen, en forma de columnas, cada uno de los registros que se le han realizado. Esta etapa implicó también hacer el recorte para el análisis: limitar los datos a los pacientes que hayan estado internados más de tres días y tomar los resultados de los parámetros que se les han medido en las dos primeras jornadas de su internación.

Por último, se definieron otros aspectos para terminar de crear el Data Set que se empleará para el modelo predictivo de clasificación: eliminar a los pacientes que no tengan ninguna medición en alguno de los signos vitales analizados y rellenar los datos faltantes de alguna medición con los resultados que obtuvo el paciente en la medición inmediatamente anterior.

El nuevo Data Set queda entonces con 1831 filas (una para cada paciente) y 29 columnas. En las columnas, 5 responden a los datos personales del paciente (ID, edad, feche de nacimiento, sexo y situación en la que sale de la Terapia Intensiva) y las otras 24 a las mediciones de sus signos vitales. Dentro de esas 24 hay 4 columnas para cada uno de los 6 signos vitales medidos. Las 4 columnas son una para la primera medición y las otras 3 para las 3 primeras medición del segundo día de internación.

El Data Set original no ha sido modificado, pudiéndose entonces utilizar la gran cantidad de información que se ha descartado para este proyecto en otros análisis.

Por último, el tercer paso implicó realizar tareas de preprocesamiento de datos, entrenamiento de modelos y análisis de resultado de los mismos. Esto se realizó en un tercer Notebook que ha sido solo destinado a la búsqueda y desarrollo del mejor modelo de Machine Learning supervisado de clasificación

Resultados y conclusiones

Sobre un Data Set con 2047 pacientes que estuvieron en Terapia Intensiva más de dos días, y que contaba con datos de edad, sexo, y 24 tomas de signos vitales durante sus primeras dos jornadas de internación, se ha realizado un análisis y limpieza de la información por un lado, y se ha aplicado un modelo predictivo de clasificación por el otro, para predecir si en función de las variables dadas podría saberse si el paciente estaría internado menos de una semana (**clase 0**) o más de una semana (**clase 1**).

El objetivo principal del trabajo es obtener el mayor porcentaje de precisión posible, medibles con los puntajes de accuracy, F1 Score (ambas considerando porcentaje de aciertos) y Cross Validation (que hace un análisis similar pero, probando no uno sino diferentes particiones de los datos entre entrenamiento y prueba).

Sin embargo, dentro de esta meta el **objetivo** principal es mejorar la capacidad predictiva de la **clase 1** ya que se considera más pernicioso creer que un paciente se quedará menos de una semana internado y que se quede más, que al revés. La capacidad de predecir cada clase se mide con el recall, que puede verse en general o para cada clase en particular.

Se ha probado con diferentes modelos predictivos de la librería scikit-learn, y luego se han medido sus resultados. Los mejores resultados se obtuvieron con el modelo Random Forest. Este arrojó, por ejemplo, una precisión del 69%, aunque teniendo en cuanta que la muestra arrojaba poco más de 1/3 (67%) de los casos para una de las clases (la clase 0) resultaba aceptable, pero baja. De todos modos, se elige este modelo para buscar mejorarlo.

Se trabajó entonces para optimizar los resultados del modelo Random Forest, primero mejorando los parámetros del modelo, con el paquete RandomizedSearchCV, de la librería ya mencionada, y luego creando nuevos casos de la clase 1 utilizando la librería imblearn para balancear la muestra.

Mediante este proceso se pasó a una precisión del 70%. El número de falsos positivos de la clase 0 aumentó, pero disminuyó el de la clase 1 que son las fallas que más se busca evitar ya que es preferible suponer que un paciente se quedará más de una semana y que luego no lo haga que al revés.

Una última propueste es la aplicación de un Smote, es decir una técnica que disminuye los casos de la clase mayoritaria, la clase 0 en este caso, y aumento lo de la otra. Sobre ese proceso se aplica un modelo de Regresión Logística. Este proceso disminuye la precisión general, pero mejora mucho el acierto para predecir la clase 1 y, sobre todo, baja los errores (falsos positivos) de esa clase.

En síntesis, el mejor modelo de Random Forest arroja una precisión del 70% con un recall (posibilidad de no arrojar falsos negativos) para la clase 1 de 47%. El modelo de Regresión Logística que penaliza la clase 0 por mayoritaria tiene un 67% de presición general, con un recall para la clase 1 de 62%.

Se han podido determinar dos aspectos útiles: a nivel médico se ha observado que no hay una correlación directa entre los parámetros vitales y los días de internación, pero sí que estos marcan una tendencia que, complementado con otros datos como diagnóstico,

resultados de laboratorio, etc., podrían ayudar a entender mejor cómo medir los días en que un paciente necesitará internación.

A nivel organizativo de la entidad de salud, poder anticiparse a la cantidad de tiempo que un paciente hará uso de una plaza de Terapia Intensiva puede mejorar la organización y la preparación del mismo para brindar una atención más organizada y eficiente. Si bien las predicciones no son categóricas, muestran una tendencia que de ser completada puede tener un importante nivel de precisión.

Se sugiere proponer al interesado ambos modelos: Random Forest con mejores parámetros y balanceado de la muestra, y Regresión logística con balanceado de la muestra para que elija entre sus necesidades predictivas.

Recursos:

Médicos:

Parámetros vitales en un monitor de UCI

Pandas:

Combinando DataFrames con Pandas

Guías para aplicar modelos:

Guide to the K-Nearest Neighbors Algorithm in Python and Scikit-Learn

Understanding Logistic Regression in Python Tutorial

How to Speed up Scikit-Learn Model Training

Evaluación de Modelos:

Evaluating Classification Models

Evaluating a Classification Model 2

Improve Your Model Performance using Cross Validation (in Python and R)

sklearn.model_selection.RepeatedKFold

K-Fold Cross Validation – Python Example

K-Fold Cross Validation Example

Mejora de parámetros de los modelos:

Hyperparameter Tuning the Random Forest in Python

Balanceo de la muestra:

Random Oversampling and Undersampling for Imbalanced Classification

Clasificación con datos desbalanceados

How to balance a dataset in Python

The right way of using SMOTE with Cross-validation

Repositorios de ejercicios similares:

Machine Learning Classification

Lead Prediction for On-line Training