INTELIGENCIA ARTIFICIAL

PROYECTO SEMESTRAL

PROYECTO FINAL

Realizado por:

Neyder Stiven Arroyave Monsalve

Santiago Franco Hernández

Maria Gabriela Alvarez Chaves

Inteligencia Artificial

Facultad de Ingeniería

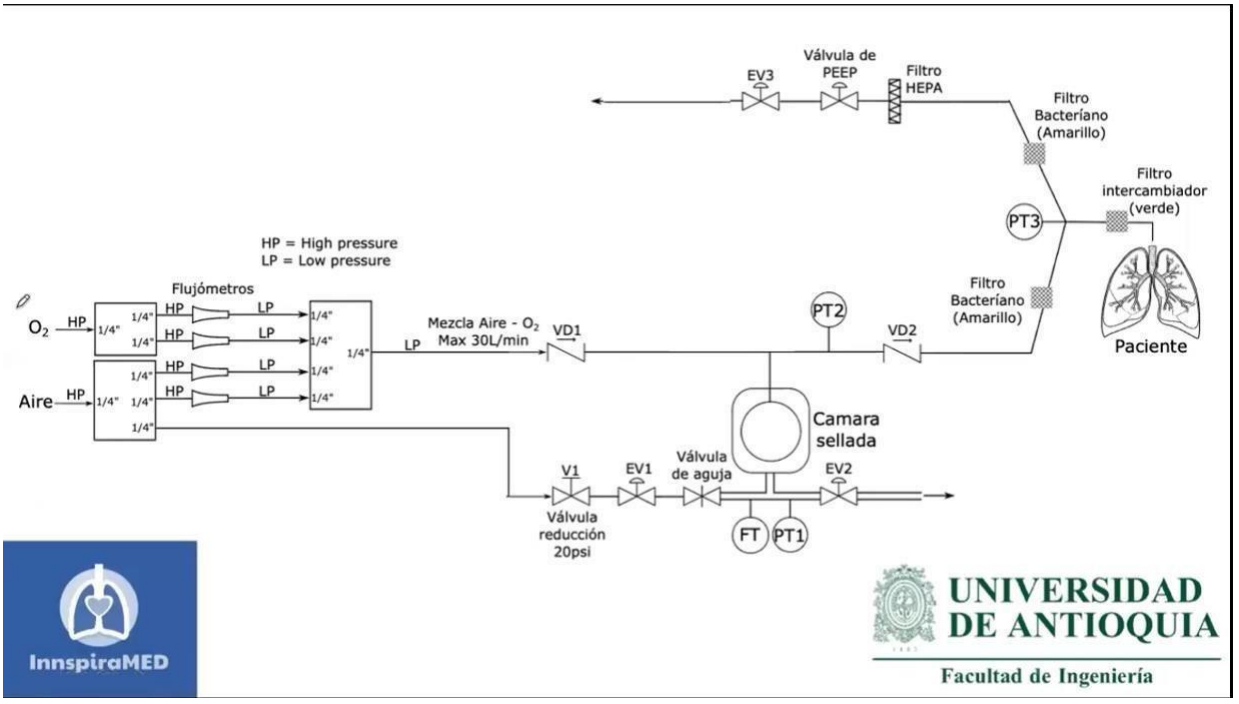
Universidad de Antioquia

Semestre 2022-1, Medellín, 22 de agosto de 2022

1. INTRODUCCIÓN

La ventilación mecánica es un procedimiento intensivo para el clínico que tuvo una gran presencia durante los primeros días de la pandemia de COVID-19. El desarrollo de nuevos métodos para controlar los ventiladores mecánicos tiene un coste prohibitivo, incluso antes de llegar a los ensayos clínicos. Los simuladores de alta calidad podrían reducir esta barrera.

Los simuladores actuales se entrenan como un conjunto, en el que cada modelo simula una única configuración pulmonar. Sin embargo, los pulmones y sus atributos forman un espacio continuo, por lo que hay que explorar un enfoque paramétrico que tenga en cuenta las diferencias de los pulmones de los pacientes. En esta competición, se simulará un ventilador conectado al pulmón de un paciente sedado. Se tendrán en cuenta los atributos de los pulmones, la conformidad y la resistencia.

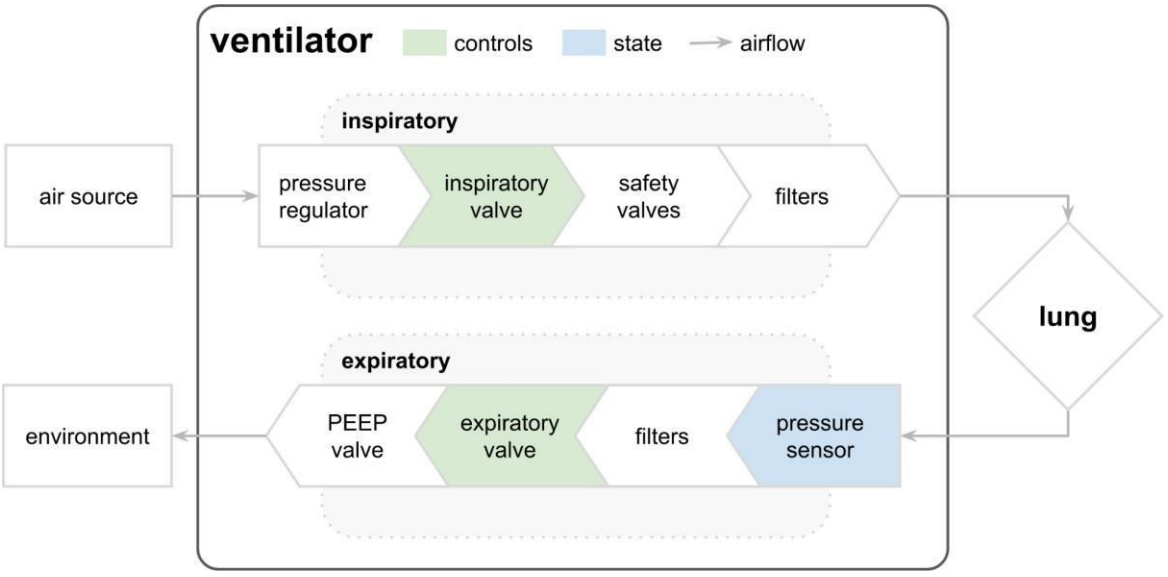


2. EXPLORACIÓN DESCRIPTIVA DEL DATASET

Los datos del ventilador utilizados en esta competición se produjeron utilizando un ventilador de código abierto modificado conectado a un pulmón de prueba de fuelle artificial a través de un circuito respiratorio. El diagrama siguiente ilustra la configuración, con las dos entradas de control resaltadas en verde y la variable de estado (presión de las vías respiratorias) a predecir en azul. La primera entrada de control es una variable continua de 0 a 100 que representa el porcentaje de apertura de la válvula solenoide inspiratoria para dejar entrar el aire en el pulmón (es decir, 0 está completamente cerrado y no deja entrar el aire y 100 está completamente abierto). La segunda entrada de control es una variable binaria que representa si la válvula exploratoria está abierta (1) o cerrada (0) para dejar salir el aire.

En esta competición, los participantes reciben numerosas series temporales de respiraciones y aprenderán a predecir la presión de las vías respiratorias en el circuito respiratorio durante la respiración, dada la serie temporal de entradas de control.

Cada serie temporal representa una respiración de aproximadamente 3 segundos. Los archivos están organizados de tal manera que cada fila es un paso de tiempo en una respiración y da las dos señales de control, la presión de las vías respiratorias resultantes y los atributos relevantes del pulmón, descritos a continuación.



Archivos

- **train.csv** - datos de entrenamiento
- **test.csv** - datos de prueba
- **sample_submissio**

- **n.csv** - un ejemplo de archivo de presentación en el formato correcto

Variables

- **id** - identificador de paso de tiempo globalmente único en todo el archivo
- **breath_id** - paso de tiempo globalmente único para las respiraciones
- **R** - Atributo pulmonar que indica el grado de restricción de las vías respiratorias (en cmH2O/L/S). Físicamente, es el cambio de presión por cambio de flujo (volumen de aire por tiempo). Intuitivamente, se puede imaginar que se infla un globo a través de una pajita. Podemos modificar R cambiando el diámetro de la pajita, siendo más difícil soplar con R.
- **C** - Atributo pulmonar que indica el grado de complacencia del pulmón (en mL/cmH2O). Físicamente, es el cambio de volumen por cambio de presión. Intuitivamente, podemos imaginar el mismo ejemplo del globo. Cambiando el grosor del látex del globo, con un C más alto teniendo un látex más fino y más fácil de soplar.
- **time_step** - la marca de tiempo real.
- **u_in** - la entrada de control para la válvula solenoide inspiratoria. Va de 0 a 100.
- **u_out** - la entrada de control para la electroválvula de exploración. Puede ser 0 o 1.
- **pressure** - la presión de las vías respiratorias medida en el circuito respiratorio, medida en cmH2O.

Para analizar el comportamiento de los datos y realizar los respectivos modelos se procedió a ubicar los datos desde Kaggle los archivos test y train; luego continuamos con el procesamiento de los datos. Se establecieron gráficos que relacionan el identificador de paso de tiempo globalmente único en todo el archivo y la presión de las vías respiratorias medidas en el círculo respiratorio para relacionar los datos y su comportamiento. Al ser una gran cantidad de datos, se realizaron filtraciones según el paso de tiempo globalmente único para las respiraciones, a fin de visualizar de manera más fácil la relación de las variables entre sí.

Luego de dicho procesado se analizó el comportamiento de cada variable de manera independiente, verificando la existencia de datos atípicos o fuera de lo establecido.

Para seleccionar la entrada y características de destino (datos a predecir), encontramos que:

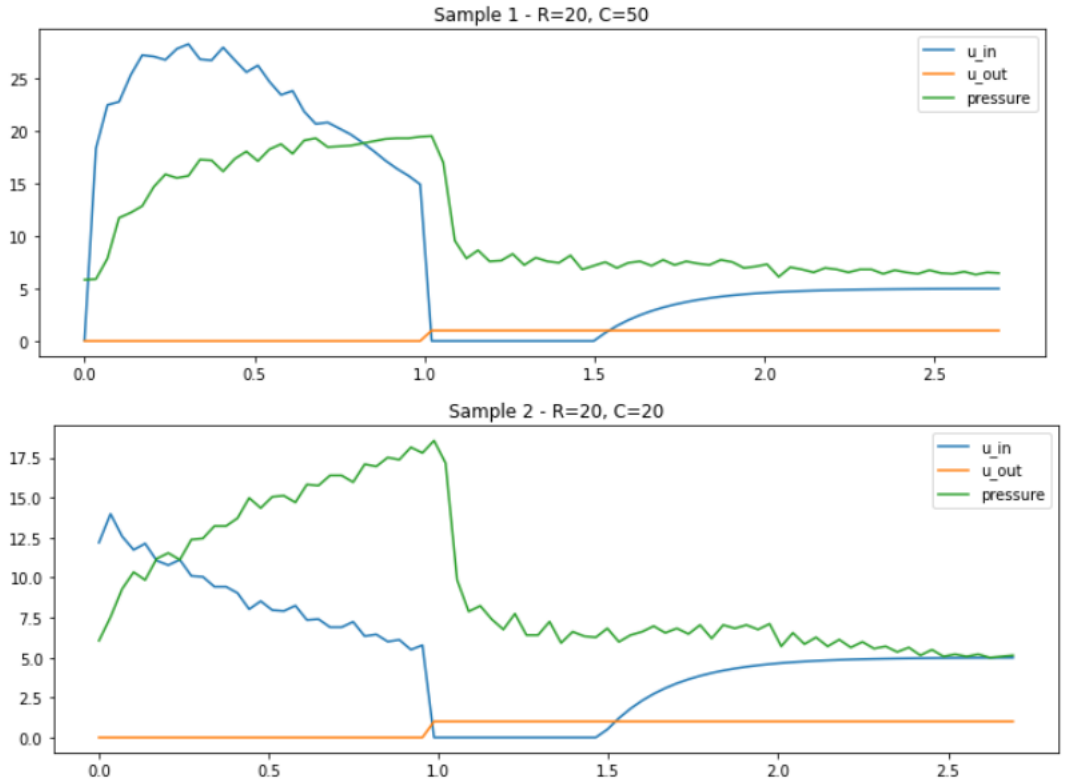
- id : Esta característica es irrelevante en la predicción de presión
- aliento_id: Ayuda a identificar los eventos, pero no es relevante para la predicción de la presión

- R: Para cada ciclo de respiración la R permanece constante, no se ha rastreado ninguna variación, por lo que no debemos considerarla como característica necesaria.
- C: La parte similar observada como en R, por lo tanto, no se considera como las características requeridas para el.
- time_step: Esta función tampoco es significativa para la predicción del objetivo
- u_in: La única característica importante que se correlaciona con la presión.
- u_out: Se comporta como un interruptor en la generación de presión.

Se encontraron relaciones luego de realizar ajustes de tamaños, como que la entrada de control para la válvula solenoide inspiratoria después de estar con una presión medida en cmH2O de 2520 tiende a suavizarse y mantener un comportamiento muy cercano a la constante.
Además, encontramos que cada ciclo de respiración tiene una duración de 80 unidades. En cada ciclo, u_in aumenta la forma en 0 y comienza a disminuir. En 30 cae bruscamente a 0, permanece igual de 30 a 45 unidades y luego aumenta exponencialmente y se vuelve constante hasta el final de cada ciclo de respiración. Aumento de presión cuando se activa u_in. Cae bruscamente después de la caída de u_in a 0. La presión no cae a 0 sino que mantiene la memoria hasta el próximo disparo de u_in.

3. ITERACIONES DE DESARROLLO

3.1 Preprocesado de datos:



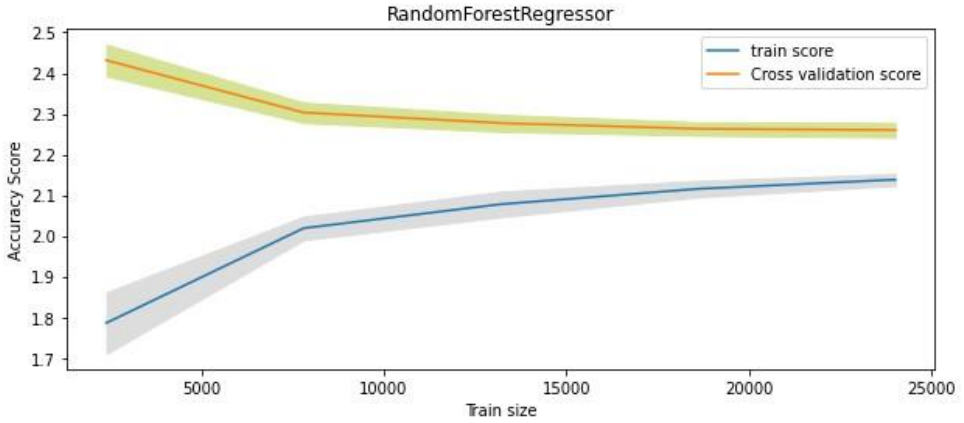
Podemos observar que:

- la variable u_out se comporta similar en ambas graficas
- La variable u_in inicia en diferentes puntos del eje y lo cual hace que tengan diferentes tendencias es decir que con los parámetros r= 20 y C = 20 esta decrezca y con los parámetros r = 20, c = 50 esta tenga una tendencia alcista y cuando está cerca al 0.5 en eje x esta empieza a decrecer.
- En ambas graficas existe una tendencia creciente de la variable pressure los cuales pueden coincidir en los picos a la hora de decrecer en el eje x con un valor de 1.

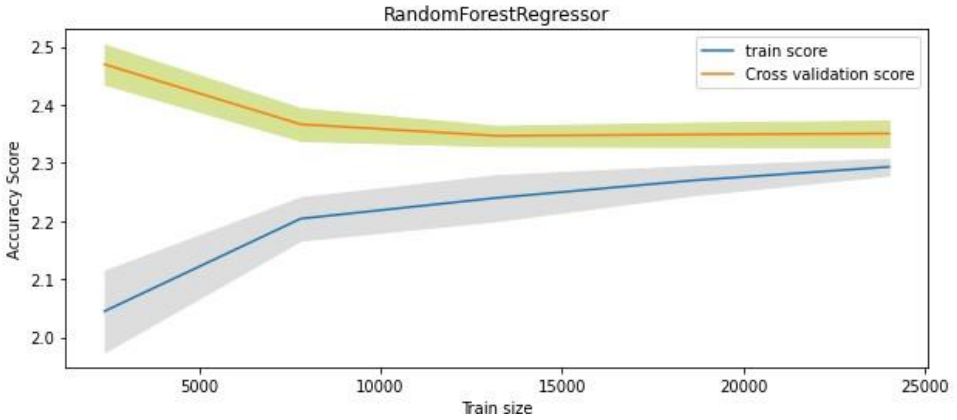
3.2 Modelos supervisados

Para realizar predicciones relacionadas con el comportamiento futuro de los datos, es decir conocer la presión de las vías respiratorias en el circuito respiratorio durante la respiración, dada la serie temporal de entradas de control, se realizaron múltiples modelos de los cuales podemos destacar los siguientes:

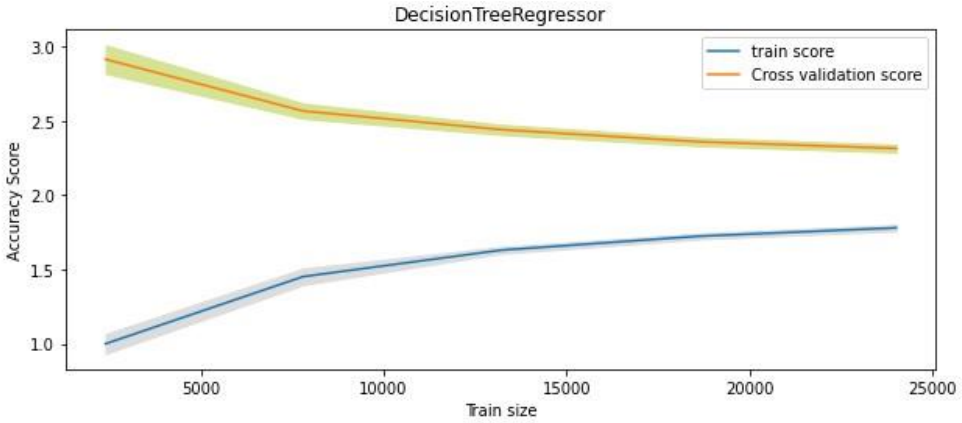
- Random forest regressor evaluation: Random forest es un algoritmo de aprendizaje automático supervisado que se usa ampliamente en problemas de clasificación y regresión. Construye árboles de decisión en diferentes muestras y toma su voto mayoritario para la clasificación y el promedio en caso de regresión.Evaluamos el Random Forest Regressor y evaluamos max_depth desde 3 hasta 11, encontramos que el mejor precision con menos bias y overfitting está entre 7 y 8, pero elegiremos el 8 porque parece tener menor bias. En este caso: los modelos con menos max_depth que 8 tienen sesgo (bias) y los modelos con más de 8 tienen sobreajusten (overfitting).



Encontramos que valor de n_estimators tiene mejor desempeño. Notamos que no tienen una diferencia significativa, por lo que decidimos trabajar con 200 para tener buen rendimiento de computación.



- Desicion Tree Evaluation: Evaluamos el Decision tree y el max_depth desde 3 hasta 11, encontramos que el mejor rendimiento con menos bias y overfitting está en 9. En este caso: Los modelos con menos max_depth que 9 tienen bias y los modelos que tienen más de 9 tienen overfitting.

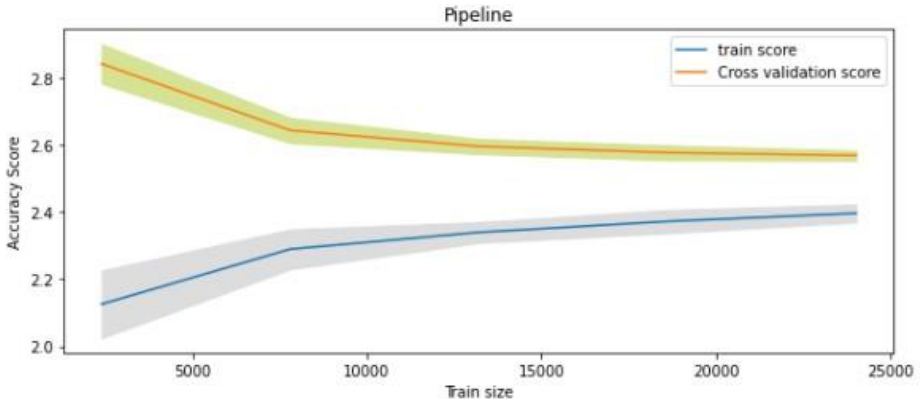


Realizamos evaluaciones correspondientes a cada uno de los modelos anteriores respecto a sus resultados estimados, mediante estas pudimos identificar que para nuestro trabajo era más conveniente analizar y trabajar con Random Forest Regressor con profundidad máxima = 8, n_estimadores = 200, teniendo en cuenta que en este caso: los modelos con menos max_depth que 8 tienen sesgo, los modelos con más de 8 tienen sobreajuste.

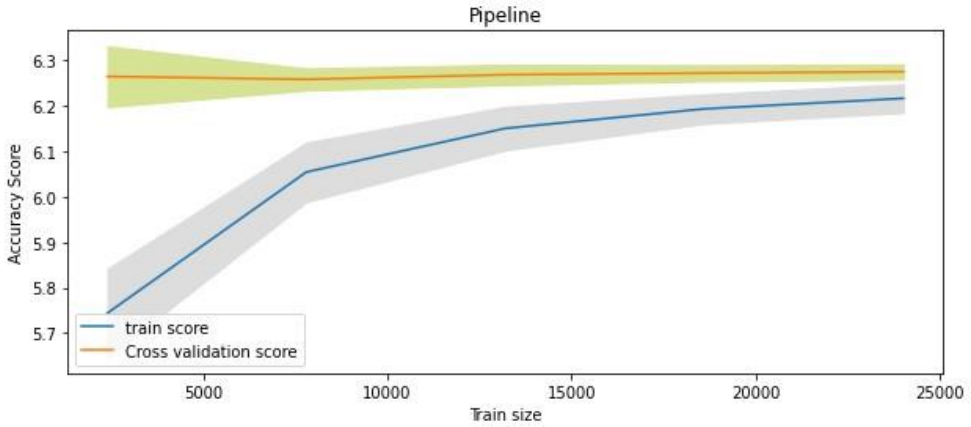
3.3 Modelos no supervisados

En la ejecución y búsqueda de soluciones para realizar las predicciones, encontramos problemas para trabajar con cierta cantidad de datos, ya que los modelos con algoritmos supervisados demandaban mucho tiempo, horas o incluso días, por ende, se hizo uso de algoritmos no supervisados para simplificar el trabajo y reducir las dimensiones de los datos.
Luego de realizar la aplicación anterior de estos algoritmos, para analizar qué cantidad de datos es óptima y ver si hay bias u overfitting evaluamos cada modelo con sus diferentes algoritmos, específicamente el n_components.

- Evaluamos PCA con diferentes componentes (de 1 a 6) y notamos que ninguno presenta problemas de overfitting o bias, sin embargo, reduciendo la cantidad de columnas se pierde accuracy de forma significativa. El de mejor rendimiento es el n_componenets = 6, que es igual a no aplicar el método. Por lo que no es recomendable en este caso. De todas maneras, si fuera necesario usarlo, escogeremos en n_components = 5, que es el que pierde menos accuracy.



- Evaluamos NMF con diferentes n_components (de 1 a 6). Se obtuvieron resultados muy poco útiles, ya que se tiene un bias muy grande en los que tienen de 2 a 6 componentes. EL único sin bias es el 1 y tiene un accuracy poco deseable. No se recomienda el uso de dicho método para este caso.



Analizando los dos algoritmos supervisados. El mejor estimador que evalúa el sesgo, el sobreajuste, la precisión y el rendimiento entre random forest regressor y decision tree es: Random Forest Regressor con max_depth = 8, n_estimators=200. En este caso: los modelos con menos max_depth que 8 tienen sesgo (bias), los modelos con más de 8 tienen sobreajuste (overfitting). Sin embargo, solo usamos 60.000 filas (42.000 en proceso), debido al rendimiento del hardware y el conjunto de datos tiene un total de 6 millones de filas.

Evaluamos PCA con diferentes n_components (de 1 a 6) pero sin la columna 'breathe_id' en el dataset, se nota una mejoría clara en el rendimiento comparado con el PCA que incluye dicha columna. Todos pierden accuracy respecto a no usar PCA, pero se gana rendimiento computacional. Dependiendo del caso se podría usar un n_componenets = 4 o 5 sin perder mucho accuracy.

Evaluamos NMF con diferentes n_components (de 1 a 6) pero sin la columna 'breathe_id' en el dataset, se nota una mejoría clara en el rendimiento comparado con el NMF que incluye dicha columna. De todas maneras, todos sufren de alto bias, el que menos bias tiene es el n_components = 1 pero tiene peor accuracy.

Para establecer modelos de producción primero se aclara que solo usamos 120.000 filas debido al rendimiento del hardware y que el conjunto de datos tiene un total de 6 millones de filas. Por lo que, para trabajar con la totalidad de los datos o una cantidad mayor se debe contar con más tiempo para la ejecución de cada modelo u obtener un mejor hardware que permita tener más capacidad de procesamiento para poder establecer niveles de desempeño para una buena producción y tener unos buenos procesos de monitoreo en los modelos.

Se encontró que el mejor algoritmo No supervisado (que evalúa el sesgo, el sobreajuste, la precisión y el rendimiento) entre PCA y NMF para este caso en el que eliminamos breathe_id del conjunto de datos es: PCA con n_components = 5 o 4 (Dependiendo de la precisión deseada), debido a que esto mostró una mejora significativa en la ejecución de cada algoritmo.

Recomendamos usar este algoritmo para poder procesar más datos en menos tiempo sacrificando la precisión.

4. RETOS Y CONSIDERACIONES DE DESPLIEGUE

Para llevar a cabo la ejecución o aplicación de este proyecto en un ámbito real, como en clínicas, hospitales y centros especializados de salud, se deben tener en cuenta diversos factores para considerar la viabilidad de este en los distintos espacios mencionados.
Esta inteligencia artificial se tendrá que almacenar en un hardware que vaya mostrando constantemente el rendimiento de la presión del ventilador mecánico, siendo esto un costo adicional que los centros de salud van a sumir. Entonces actualmente consideramos que el proyecto no es viable, ya que se debe contar con el equipo adecuado para esto y el personal profesional capacitado, lo que es poco probable de llevar a cabo en los países subdesarrollados.

5. CONCLUSIONES

- Se logró identificar que para nuestro trabajo era más conveniente analizar y trabajar con Random Forest Regressor con profundidad máxima = 8, n_estimadores = 200, teniendo en cuenta que en este caso: los modelos con menos max_depth que 8 tienen sesgo, los modelos con más de 8 tienen sobreajuste.
- Para los algoritmos no supervisados La eliminación de 'breathe_id' del conjunto de datos mostró una mejora significativa en los algoritmos PCA y NMF.
- Se recomienda usar el algoritmo no supervisado para poder procesar más datos en menos tiempo sacrificando la precisión.
- Se utiliza un numero de filas menor debido al rendimiento del hardward. Si se quiere usar más datos se debe esperar más tiempo para la ejecución de los modelos.
- El proyecto no es viable, ya que se debe contar con el equipo adecuado y el personal profesional capacitado, lo que es poco probable de llevar a cabo en los países subdesarrollados por los recursos de estos.