Proyecto 1: Aprendizaje Supervisado

Andres Saldaña Rodriguez
A01721193@tec.mx
Tecnológico de Monterrey
Ingeniería en Ciencias de Datos y Matemáticas
Monterrey, Nuevo León, México

Lester Santiago García Zavala
A01721128@tec.mx
Tecnológico de Monterrey
Ingeniería en Ciencias de Datos y Matemáticas
Monterrey, Nuevo León, México

Antonio Jose Zarate Lozano
A01720990@tec.mx
Tecnológico de Monterrey
Ingeniería en Ciencias de Datos y Matemáticas
Monterrey, Nuevo León, México

ABSTRACT

En este artículo nos encontramos con la implementación de modelos de aprendizaje supervisado para poder clasificar si una persona tiene diabetes o no. Lo que se hizo fue analizar la base de datos binaria de diabetes. Se entrenaron los modelos y se hicieron predicciones. Durante este artículo se verá el contexto del problema, la metodología usada, los conceptos importantes. Al igual que, los resultados de tanto el análisis exploratorio de la base de datos como de los modelos en sí, con sus debidas conclusiones.

ACM Reference Format:
Andres Saldaña Rodriguez, Lester Santiago García Zavala, and Antonio Jose Zarate Lozano. 2022. Proyecto 1: Aprendizaje Supervisado. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCCIÓN

El aprendizaje automático forma parte de la inteligencia artificial. Se basa en datos para ir aprendiendo y poder después realizar los procesos de forma automática.[7] Es decir si es un problema de clasificación, se le da los datos iniciales, el modelo aprende de ellos y después trata de predecir qué clasificación asignarle a los datos entrantes.

En este caso como se mostrará a continuación nos encontramos con un dataset extenso binario de diabetes. Por lo que el problema con el que nos encontramos es poder predecir en base a las demás variables que contiene el dataset si un paciente tiene o no tiene diabetes. Las demás variables nos platican de información general del paciente. Por ende, es de suma importancia tener un modelo que teniendo esos datos generales pueda determinar si una persona tiene o no tiene diabetes. Ya que pudiera servir para detectar el diabetes en pacientes solo en base a cierta información como su BMI, si es fumador o no, si tiene presión alta, si tiene colesterol alto, si le ha dado un infarto, y si realiza actividad física o no.

Con anterioridad un estudio para predecir la diabetes tipo 2 en adultos los investigadores ecuatorianos: Belkis Sánchez Martínez, Vladimir Vega Falcón, Nairovys Gómez Martínez realizaron una regresión logística para poder predecir el diabetes tipo 2. Con este método obtuvieron resultados positivos ya que se cumplieron ambas hipótesis propuestas.[6]. En nuestro caso se emplearán modelos de clasificación con aprendizaje automático.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.1 Contexto

La diabetes es una enfermedad crónica, la cual los afectados pierden la capacidad de regular los niveles de glucosa en su sangre. Esto se debe a que pierden en control parcial o total de su producción de insulina. La insulina es una hormona, que se produce en nuestro cuerpo naturalmente. La insulina ayuda a que la glucosa entre nuestras células nos ayude a producir energía. Es como si la insulina fuera una llave que abre una puerta hacia nuestras células, dejando pasar a la glucosa y ayudándonos a producir energía. [8]

Existen 2 tipos de personas diabéticas. Las personas diabéticas tipo 1, las cuales no pueden producir nada de insulina, es decir no tiene llaves para dejar pasar a la glucosa. Las personas diabéticas tipo 2 sí producen insulina, pero esta no funciona de manera adecuada. Es decir es como si el cuerpo produjera llaves pero estas estuvieran dobladas y no pudieran abrir las puertas. También existen distintas variaciones de diabetes tipo 2, pero estos ya son más específicos y poco comunes.

El exceso de glucosa que se encuentra en la sangre a causa del déficit de insulina, puede causar muchos problemas. Algunos pueden ser leves, como sentirse cansado, estar sediento, perder peso, etc. Sin embargo, tiene complicaciones más severas que pueden llevar a afectar el corazón, los riñones, los ojos, pies, etc. Estas complicaciones pueden causar la muerte.

El "dataset" que estaremos utilizando fue creado por el gobierno de gran bretaña. Ellos marcaban a personas del país, y les hacían preguntas acerca de su salud, las cuales anotaban en el "dataset". Muchas preguntas estaban orientadas hacia la diabetes, como por ejemplo si tienen el colesterol alto o no, pero otras preguntas son más generales, por ejemplo; si eran fumadores.

2 CONCEPTOS PREVIOS Y SU DESCRIPCIÓN

- **Análisis exploratorio de datos:** El análisis exploratorio de datos consiste en mostrar los estadísticos de cada atributo de la base de datos.
  - **Conteo de datos:** Cantidad de instancias de cada atributo.
  - **Promedio o Media:** Se suman todos todos los valores y se divide en el conteo.
  - **Desviación estándar:** Cuanto varían los datos de la media.

  - **Mínimo:** Valor mínimo.
  - **Máximo:** Valor máximo.
  - **1er Cuartil:** Determinado por el valor que indica que los valores menores a ese pertenecen al 25 por ciento de los datos.
  - **2o Cuartil:** Determinado por el valor que indica que los valores menores a ese pertenecen al 50 por ciento de los datos.
  - **3er Cuartil:** Determinado por el valor que indica que los valores menores a ese pertenecen al 75 por ciento de los datos.
- **Correlación:** Indica que tan fuerte es una tendencia lineal entre 2 variables. También si es positiva o negativa.
- **K-Nearest Neighbors (KNN):** Método de aprendizaje supervisado para clasificar datos. Se le da el número "k" que indica cuántos vecinos se tomarán. El método KNN lo que hace es que con los valores de entrenamiento se le saca la distancia euclidiana al nuevo valor a clasificar. Después se acomoda de menor a mayor dsitancia. Se checan los "k" vecinos más cercanos (menor distancia) y la clase que más se repita en los datos de entrenamiento es la clase que se le asignará al valor que se quiere clasificar.
- **Support Vector Machine (SVM):** Un método de aprendizaje supervisado que busca encontrar el hiper plano que maximice la distancia mínima entre los puntos del conjunto de entrenamiento. Esto para que con la ecuación del hiper plano obtenida se le inserta el valor a clasificar y dependiendo de si da positivo o negativo se le asigna clasificación.
- **Random Forest:** Es un método de clasificación que usa un árbol sencillo. Es un árbol binario. Lo que se es busca es pasar data del nodo padre a los nodos hijos con la intención de mejorar la homogeneidad.[5]
- **Redes Neuronales:** Las redes neuronales buscan clasificar mediante una función de actuación y una de activación. Como funciona es que recibe los datos de entrada luego en las capas ocultas se encuentra la función de actuación y de salida sale la función de activación con la cual le asigna la clasificación al nuevo dato.
- **Árboles de decisión:** Método para clasificar basado en condiciones. Es decir empieza con una condición si la condición se cumple y es Verdadero, se va a la derecha en el siguiente nivel, si es falso se va hacia la izquierda en el siguiente nivel. Así hasta llegar al máximo nivel de profundidad. Cada nivel tiene una clasificación que indica que clase asignarle al valor que se quiere probar.
- **Matriz de confusión:** Forma de evaluar el modelo escogido de manera gráfica para ver los falsos positivos y falsos negativos al igual que los verdaderos positivos y los verdaderos negativos. Muestra los errores de tipo I (falso positivo) y de tipo II(falso negativo).
- **Asertividad:** Métrica para ver que tan correcto es el modelo.
- **Presición:** Métrica para evaluar que tanto se ajusta el modelo a los datos.
- **Sensibilidad y Especificidad:** La sensibilidad mide que tan probable es que un resultado sea positivo cuando en realidad es positivo, mientras que la especificidad mide que tan probable es que un resultado sea negativo cuando en realidad es negativo.[1]
- **Curva ROC:** La curva ROC grafica en el eje vertical la Sensibilidad y en el eje horizontal la Especificidad de los modelos.[4]
- **Validación cruzada:** La validación cruzada sirve para medir un modelo de machine learning en el cual se comparan los resultados obtenidos con los datos de prueba contra unos datos que se usaron para entrenar el modelo.[2]
- **Índice de GINI:** Es un índice usado para medir la pureza de los nodos hijos en un árbol de decisión. Esto quiere decir que se calcula con la probabilidad de que si dos datos de la misma clase al ir bajando en el árbol de decisión siguen perteneciendo a la misma clase o cambia. Entre mayor el número más puro el árbol.[3]
- **Entropía:** Muestra de forma cuantitativa la incertidumbre de la muestra del árbol de decisión.

3 METODOLOGÍA

3.1 Descripción general del dataset

Nuestro "dataset" descargado de "Kaggle" es acerca de la enfermedad del diabetes. Por lo tanto contiene la columna binaria de diabetes el cual sirve como la clase y 21 atributos con 70692 instancias. Encontramos atributos que servirán para a la hora de ambos proyectos el de aprendizaje supervisado y del aprendizaje no supervisado elaborar modelos de inteligencia artificial que nos puedan predecir con alta exactitud si un paciente tiene o no tiene diabetes.

También como nos encontramos con muchos atributos se tendrá después que analizar cuales si tienen un impacto en el resultado de la clase y cuales no. Ya que a simple vista hay varios que no aportaran para catalogar al paciente como diabetico o no. Como lo son CholCheck, AnyHealthcare, NoDocbcCost, GenHlth, PhysHlth, MentHlth, Education, e Income.

Todo esto se analizará con más detalle en las siguientes fases de los retos.

3.2 Describir cada uno de los atributos

En nuestro "dataset" encontramos 21 atributos de los cuales creemos que solamente 13 de ellos son de vital importancia para realizar el proceso de aprendizaje ya que nos aportan información importante para determinar porque el encuestado tiene o no diabetes. Los otros 8 atributos sobrantes creemos que pueden ser descartados debido a que no nos aporta información necesaria para determinar si una persona tiene o no diabetes.

Los 13 atributos que si utilizaremos para el proceso de aprendizaje son los siguientes: HighBP el cual nos indica si el encuestado tiene una presión arterial alta. HighChol para saber si el nivel de colesterol es alto. BMI es el índice de masa corporal el cual nos puede proporcionar si tienen obesidad o no. Smoker el cual nos indica si la persona ha fumado más de 5 cajetillas en toda su vida. Stroke para saber si el encuestado ha tenido alguna vez en su vida un derrame cerebral. HeartDiseaseorAttack para determinar si han tenido algún ataque al corazón o algún problema relacionado con

su corazón durante su vida. Los atributos de Fruits y Veggies nos indican si los encuestados consumen frutas y verduras diariamente es decir si comen de manera saludable. PhysActivity es un atributo que nos indica si han realizado algún tipo de actividad física en el último mes. HvyAlcoholConsump este atributo nos indica si consumen grandes cantidades de alcohol y para determinar esto en el caso de hombres son 14 vasos a la semana mientras que en mujeres son 7. DiffWalk en donde nos hacen saber los encuestados si acaso tienen una dificultad para caminar o para subir las escaleras. Los últimos dos atributos Sex y Age son de gran importancia ya que nos indica si el encuestado es hombre o mujer y la edad que tienen por lo que nos ayudara a determinar en que rango de edad y en que sexo la diabetes es mas común. Todos los atributos que tienen que ver con enfermedades nos harán saber que tan relacionadas están con la diabetes es por eso que son de gran importancia.

Por otro lado los 8 atributos que serán descartados son los siguientes: Education el cual nos dice el nivel de educación de los encuestados. Income el cual nos indica el ingreso de cada uno de los encuestados. Ambos casos mencionados anteriormente no tienen relación con la diabetes ya que es solo el nivel social y económico del encuestado. Los atributos de salud física, salud mental y salud general se descartaran ya que no confiamos en el auto diagnóstico de los encuestados respecto a estos temas que son de gran importancia por lo que nos fijamos en problemas que sí pueden ser medibles como las enfermedades. NoDocbcCost el cual indica si las personas no pudieron ver a un doctor debido al costo. AnyHealthcare si tienen algún seguro de gastos médicos por lo que nuevamente ambos atributos solo nos muestran el nivel económico de los encuestados. Por último CholCheck el cual indica si se han realizado un chequeo de colesterol en los últimos 5 años el cual se descarta ya que ir a checar su nivel de colesterol no tiene un impacto para determinar si tienen diabetes aunque el resultado de dicho chequeo si lo tiene.

3.3 ¿Qué conocimiento queremos extraer?

Nuestra meta como se mencionó con anterioridad seria poder predecir si una persona es diabética o no, aunque estaríamos agrupando "tener diabetes" con ser "pre-diabético". Queremos ver si podemos predecir la diabetes utilizando estas medidas de salud, o indicadores de salud. Como lo son las preguntas de; "¿Haz tenido un ataque al corazón?" , "¿Fumas?, "¿Cual es tu índice de masa corporal?", "Sexo" etc. Si logramos ser exitosos, estaríamos creando una herramienta que pudiera predecir la diabetes sin necesidad de ninguna prueba, simplemente respondiendo preguntas de el estilo de vida de una persona.

Con esta idea, podríamos formar un cuestionario en una pagina en linea, en el cual, en cuestión de minutos, te pudiera decir si es posible que seas diabético o pre-diabético, basado en tu estilo de vida. Considerado que millones de personas padecen de esta terrible enfermedad, seria muy conveniente poder averiguar la disposición de uno a ella, con un simple cuestionario en linea.

3.4 Proceso a seguir para la extracción del conocimiento.

(1) Investigar la base de datos en Kaggle.
(2) Descargar la base de datos.
(3) Definir la problemática a solucionar.
(4) Cargar las librerías necesarias.
(5) Cargar la base de datos en una libreta de Jupyter.
(6) Eliminar las columnas no deseadas.
(7) Realizar análisis exploratorio de datos para conocer la base de datos.
(8) Entrenar los modelos.
(9) Probar los modelos.
(10) Medir los modelos.
(11) Observar los resultados.
(12) Concluir.

3.5 Metodología CRISP en el proyecto

La metodología CRISP-DM, es el modelo analítico más usado dentro del área de la minería de datos. Ya que explica los pasos a seguir, desde que se llega con una organización o empresa, hasta que se entrega un modelo de "machine learning" o inteligencia artificial.

Se empieza con el entendimiento de los datos, y el entendimiento de la empresa. Es decir que tenemos que lograr comprender cómo es que opera la empresa, cuál es el trabajo que hace,y como los datos que nos están proporcionando se relacionan a esto. Esta etapa es muy importante, y de esta depende que podamos establecer objetivos y metas claras y alcanzables. En pocas palabras les da contexto a los datos.

Luego pasamos a la fase de preparación de datos, en esta fase es una de las más tardadas, ya que nos tenemos que asegurar de limpiar los valores nulos que encontremos, analizar la distribución de los datos, ver que tan balanceados están, eliminar datos que no nos vayan a servir, escalar los datos, etc. Básicamente nos tenemos que asegurar de que los datos que nos dio la empresa sean aptos para los modelos que vayamos a usar en los siguientes pasos. Aquí tiene mucho que ver el concepto "garbage in, garbage out", es decir que si le metes basura a los modelos, estos van a modelar basura. Con esto en mente es que debemos de explorar y asegurarnos de que los datos que estaremos usando en los modelos sean de la mejor calidad posible.

Finalmente pasamos a la fase de la modelación de los datos. Ahora si los datos que estuvimos limpiando y preparando anteriormente se pasan a los modelos. Se entrenan con estos datos, y luego los probamos para poder evaluar el desempeño de los modelos.

Tras conseguir las métricas adecuadas para evaluar los modelos, es decir no vamos a utilizar métricas para modelos de clasificación que para modelos de regresión. Una vez evaluados los modelos, vemos si estos resultados son los que esperábamos, en caso de que no lo sean, es posible que nos tengamos que regresar a la fase de modelaje y optimizar los hiper-parámetros, o incluso hasta el principio hasta el entendimiento del negocio, en caso de que los resultados que nos arrojen los modelos sean terribles.

Finalmente tenemos la fase de la entrega del modelo. En esta fase se puede entregar algún tipo de aplicación al cliente, o algo similar, para que ellos puedan utilizar el modelo cuantas veces quieran, con nuevos datos. Es muy importante estar seguros que el modelo pase todas las métricas establecidas en la evaluación y que cumpla con el propósito original antes de entregárselo al cliente. Después se le puede hacer mantenimiento o incluso seguir mejorándolo, pero esto ya no es parte de la metodología CRISP-DM.

4 RESULTADOS

En esta pequeña tabla a continuación podemos ver las diferentes métricas previamente explicadas en la sección de conceptos. Aunque para este proyecto nos estaremos enfocando en la métrica de exactitud "accuracy", podemos ver otras métricas para poder comparar a los modelos de diferentes perspectivas. Hasta abajo podemos ver un renglón que dice: "total", esta es la suma de todas las métricas para poder tener una idea más general del desempeño del modelo, entre más cerca de 5 este, mejor desempeño tiene ese modelo. Importante notar que aunque un modelo tenga un mal desempeño en general puede que sea muy bueno para algo en específico como lo son los árboles de decisión para la "especificidad", ya que ellos logran tener la más alta de todos los modelos. Desde esta fase preliminar podemos hacer supuestos de que el algoritmo de redes neuronales y el de árboles de decisión serán los mejores para lo que queremos lograr, lo cual es conseguir el algoritmo con la exactitud más alta. cabe mencionar que estas métricas son con los hiperparametros "default" dentro de los modelos, y que pudieran mejorar de optimizar los modelos, como se hace más adelante.

	KNN	SVM	RandF	RedesN	Dtree
Accuracy:	0.691	0.707	0.671	0.712	0.661
Precision:	0.683	0.692	0.67	0.695	0.677
Sensitivity:	0.707	0.744	0.673	0.751	0.611
Specificity:	0.674	0.671	0.67	0.672	0.71
Roc Auc Score:	0.747	0.769	0.736	0.781	0.676
Total:	3.502	3.583	3.42	3.611	3.335

Figure 1: Tabla con las métricas de de evaluación para todos los modelos.

A continuación se muestra la curva ROC. Esta curva ROC es una representación visual del puntaje ROC. Mientras más hacia arriba y hacia la derecha este la curva, como formando un especie de ángulo recto en la esquina superior izquierda, mucho mejor la capacidad del algoritmo para poder distinguir las diferente clases de la variable predictora. Como podemos ver muchos algoritmos tienen una curva similar, a excepción de árboles de decisión, el cual podemos ver que batalla un poco más y se queda corto. De igual manera podemos observar como Redes Neuronales y SVM son las que tuvieron ligeramente mejor desempeño por encima de KNN y árboles aleatorios.

Además de todas las métricas que empleamos para evaluar los modelos también obtuvimos la matriz de confusión la cual se muestra a continuación. En dicha matriz podemos observar la cantidad de verdaderos positivos y negativos además de la cantidad de falsos

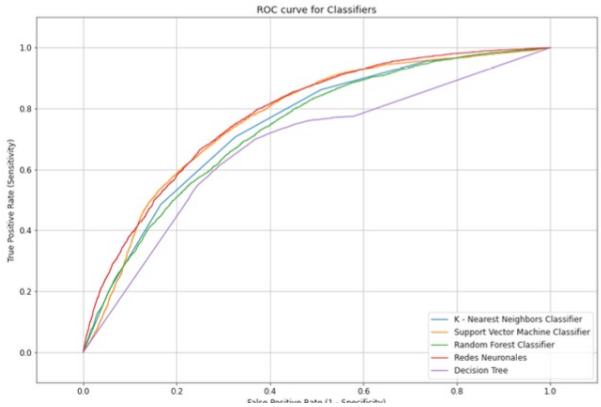

Figure 2: Gráfico de la Curva ROC para todos los modelos.

positivos y negativos de todos los modelos. En dicha matriz observamos que la cantidad de falsos positivos y negativos en los modelos son muy similares aunque hay unos que destacan. Lo que se busca obtener en este caso es la menor cantidad de falsos positivos y negativos además de tener la una distribución similar en cantidades. En este caso podemos observar que SVM y Redes Neuronales tienen cantidades bajas en falsos negativos a comparación de los demás modelos y promedio en el caso de falsos positivos pero en total son los dos modelos que menor cantidad de falsos tienen. Por lo que nuevamente se podría considerar que ambos modelos mencionados con anterioridad tienen un mejor desempeño que los demás.

```
K - Nearest Neighbors Classifier  - True Positive: 4982 True Negative:
4782 False Positivee: 2308 False Negative: 2067
Support Vector Machine Classifier - True Positive: 5244 True Negative:
4759 False Positivee: 2331 False Negative: 1805
Random Forest Classifier          - True Positive: 4741 True Negative:
4751 False Positivee: 2339 False Negative: 2308
Redes Neuronales                  - True Positive: 5292 True Negative:
4768 False Positivee: 2322 False Negative: 1757
Decision Tree                     - True Positive: 4310 True Negative:
5031 False Positivee: 2059 False Negative: 2739
```

Figure 3: Matriz de Confusión de todos los modelos.

Por último, cabe mencionar como después de optimizar los hiperparámetros y hacer la validación cruzada de nuestros algoritmo con mejor desempeño obtenemos:

Para SVM una exactitud de 0.69 con una desviación estándar de +/- 0.01 y para Redes neuronales exactitud de 0.71 con una desviación estándar de +/- 0.0.

Siendo estas las cifras finales que presentaremos al cliente, ya que son el promedio de exactitud con los algoritmos optimizados y tras haber hecho una validación cruzada de k = 5. Así que podemos estar seguros que no cambiarán significante mente.

5 CONCLUSIONES

Tras probar los 5 modelos, nos dimos cuenta que la mejor opción sería usar redes neuronales. Aunque SVM también tuvo un buen desempeño, este algoritmo es muy tardado , a comparación del resto.

Por lo mismo, fue más tardado encontrar los hiper-parámetros de este, porque se tardaba mucho, aunque fueran pocos datos. Por esto, terminamos limitando el número de datos a 5000, en la parte de optimización, de este algoritmo en particular. Aunque el SVM y KNN, tuvieron un beneficio notable de haber escalado los datos con "MinMax", el algoritmo de redes neuronales tuvo un mejor desempeño sin este escalamiento. Aunque técnicamente el algoritmo de redes neuronales fue el que mejor desempeño tuvo, con .71 de exactitud, creemos que hay espacio para mejora, e incluso se podría llegar a un .80 si se llegan a utilizar los hiper-parámetros adecuados y se le dispone tiempo suficiente.

Este proyecto también ayudó a ilustrar cómo los diferentes algoritmos son buenos para diferentes cosas, ya que por lo general los árboles decisión suelen ser muy buenos, pero para este conjunto de datos en particular no tuvieron buen desempeño. Fue muy interesante explorar y modelar esta base de datos, y fue un alivio ver que estaba tan limpia y balanceada, no nos encontramos con valores nulos. De igual manera creemos que hay espacio para mejorar en caso de que algún día se quiera volver a visitar el proyecto.

5.1 Reflexiones

5.1.1 ***Andrés Saldaña****.* - Aunque ya había hecho un proyecto similar en la clase de ciencia de datos, muchas cosas fueron diferentes. Tuve la oportunidad de trabajar con un "dataset" machismo más grande, ya que previamente había trabajado con uno de 5,000, y ahora pude trabajar con uno de 70,000. También aprendí cosas nuevas como la optimización de hiper-parámetros y las pruebas PCA, de reducción de dimensionalidad. Aunque se pudo haber sacado adelante el proyecto sin esas cosas, definitivamente aportan a la mejora del mismo, y es bueno saber cómo usarlas. Algo muy importante es que tenía una percepción errónea de cómo funcionaban las matrices de correlación, y sobre todo pude aclarar muchas dudas que me quedaron previamente en la clase anterior. Aunque todavía tengo algunas dudas acerca del tema, ahora tengo mucho mejor conocimiento de algunos conceptos los cuales pensaba que dominaba previamente. Me da gusto seguir haciendo proyectos y trabajos relacionados a el aprendizaje automático ya que siento que con cada proyecto o tarea que hago, aprendo un poco más, y me acerco a convertirme en un científico de datos completo.

5.1.2 ***Lester García****.* - A través de este proyecto logre utilizar todo el conocimiento aprendido en la clase y aplicarlo en un caso de la vida real. En el proyecto se trabajó con un dataset muy grande a comparación de los que había utilizado para otros proyectos ya que fueron más de 70,000 datos. Además realmente no tenía conocimiento de cómo los modelos de clasificación podían ser implementados en la vida real por lo que no entendía la importancia de todo lo que estaba aprendiendo durante la clase. Durante este proyecto pude verificar la importancia de cada uno de los modelos y que dependiendo de los datos que se tengan es el modelo que debemos de utilizar. En nuestro caso utilizamos 5 modelos los cuales aunque arrojaron valores similares al momento de evaluarlos, había un modelo mejor que los demás. Pudimos aplicar distintos algoritmos para evaluar los modelos lo cual fue bastante interesante ya que los peores modelos según un algoritmo de evaluación no siempre era el peor en otro por lo que me mostró la importancia de realizar todas las evaluaciones posibles a los modelos para obtener el mejor. Todo esto se que será de gran ayuda para crecer como estudiante y en un futuro aplicar todos mis conocimientos en el ámbito laboral.

5.1.3 ***Antonio Zárate****.* - Con este proyecto siento que me desarrolle más como científico de datos. Se trabajó con un dataset que contenía demasiados datos. Sobre todo lo más importante es poder ver los modelos de clasificación aplicados para un problema de la vida diaria, sobre todo un problema de salud humana. El poder ver cómo sirven los modelos de clasificación en el mundo real me sirvió mucho para poder apreciar lo que estamos aprendiendo, y para poder darle cuerda a mi imaginación y así ver en que más lo pudiera aplicar. Siento que aprendí mucho en el apartado de los modelos de aprendizaje supervisado, en especial a la hora de evaluar los modelos. Porque si bien se sabe que se pueden implementar muchos pero cada uno tiene resultados de asertividad distintos. Me ayudó a ver como funciona cada tipo de modelo de aprendizaje supervisado, como se entrenan, y cómo se evalúan para ya dependiendo de los datos que tenga a mi disposición escoger que modelo es el óptimo para ese dataset, compararlo con los otros y evaluarlos. Muchas de las formas de evaluar los modelos eran nuevas para mí y pude aprenderlos durante este proyecto. Como la validación cruzada, la curva ROC, la sensibilidad y especificidad.

REFERENCES

[1] Sensibilidad y especificidad.
[2] Validación cruzada.
[3] Qué son los árboles de decisión y para qué sirven, Apr 2022.
[4] Cerda, J., and Cifuentes, L. Uso de curvas ROC en investigaciÃclÃnica: Aspectos teÃ-prÃ¡cticos. *Revista chilena de infectologÃa 29* (04 2012), 138 – 141.
[5] Chen, X., and Ishwaran, H. Random forests for genomic data analysis. *Genomics 99*, 6 (2012), 323–329.
[6] SÃ¡nchez MartÃnez, B., Vega FalcÃ, V., and GÃMartÃnez, N. PredicciÃde la diabetes mellitus tipo 2 en pacientes adultos mediante regresiÃlogÃstica binaria. *Dilemas contemporÃ¡neos: educaciÃ, polÃtica y valores 8* (08 2021).
[7] TechTarget, C. d. ¿qué es aprendizaje automático (machine learning)? - definición en whatis.com, Jan 2017.
[8] UK, D. Diabetes: The basics.