Inteligencia Artificial y Aprendizaje Automático

Actividad semana 9: Taxonomía de métricas de clasificación

Maestría en Inteligencia Artificial Aplicada

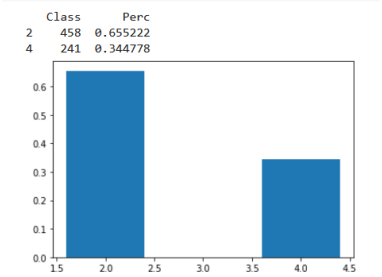
Prof. Luis Eduardo Falcón Morales

Equipo:

- Genaro Rodríguez Vázquez A01150931
- Juan Pablo Acosta López A01794035
- Maria Nelly Porras Alcantar A01793828
- Juan Carlos Torres Luna A01163204

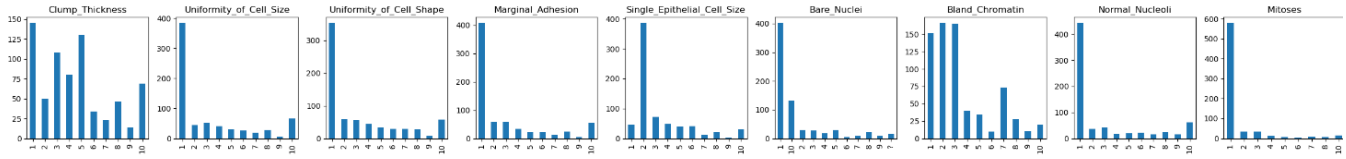
I) Inspección de Datos

1) Breast Cancer Wisconsin (Diagnostic) Data Set:



Se puede observar que en esta base de datos de cáncer de senos tiene clases de la variable independiente ligeramente desbalanceadas, por lo tanto, podría ser factible hacer técnicas de submuestreo o sobre muestreo, aunque podrían no ser necesarias.

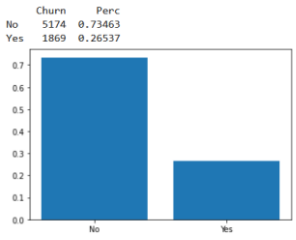
Respecto a la métrica, se entiende que la predicción de una enfermedad cae dentro de los casos en el cual es muy difícil ponerle un precio el hecho de diagnosticar a alguien con una enfermedad, por lo tanto, en este caso en específico es más importante medir el *recall* que la precisión. Sin embargo; dado que, por temas de tiempo para atender a todos los posibles pacientes, se recomienda maximizar el *recall* a un **k** de precisión que permita atender a varios pacientes al día, sin saber que tanto tiempo tardan las pruebas de diagnóstico no se puede calcular el **k** óptimo de la precisión.



Al evaluar su distribución, vemos que en 6 de 8 variables de entrada (*sample code number* debe de ser eliminada, ya que actúa como simple *index*), hay una distribución para nada normal (cargada a la izquierda). Dicho lo anterior, será necesario realizar un *pipeline* que escale estos valores según el máximo y mínimo de todo el conjunto de datos X.

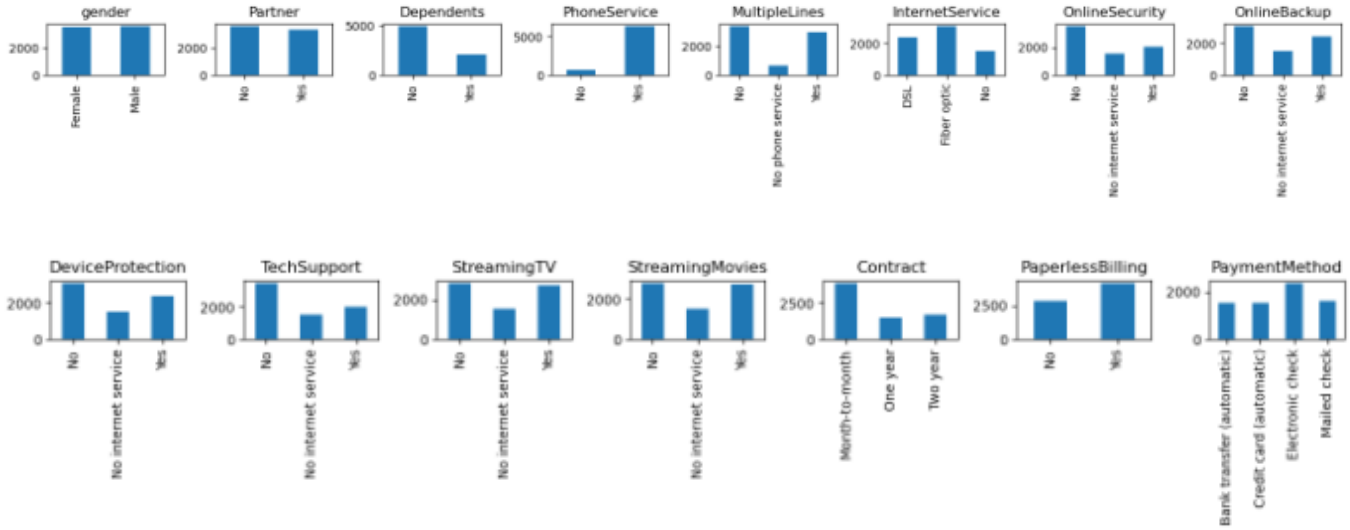
Debido a que la mayoría de las variables de entrada son de tipo cuantitativo, algún modelo de regresión lineal debería de ser el mejor en predecir de forma correcta la clase de salida, sin requerir demasiada capacidad de cómputo.

2) Telco customer churn IBM Data Set:

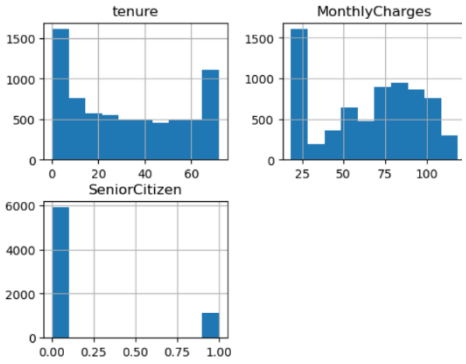


Se puede observar en esta base de datos clientes potenciales y no confiables se tienen clases de variable independiente desbalanceadas, por lo tanto, es recomendable pasar los datos a través de procesos de submuestreo y sobre muestreo.

Desde el punto de vista financiero, la literatura nos dice que mantener a un cliente es mucho más barato que obtener nuevos clientes, por lo tanto, es muy importante el mantenimiento de los clientes. Se requiere maximizar el *recall* para poder identificar la mayoría de los clientes que pudieran abandonar. De otro modo, a querer maximizar el *recall* se debería de analizar el costo de catalogar como posible abandono a alguien que realmente no va abandonar, dependiendo de cuál sea el mecanismo que se implemente para los posibles abandonos y se tendría que evaluar que tan caro podrían ser los falsos positivos.



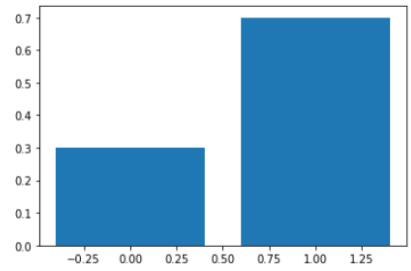
Para el estudio de Telco Churn, encontramos que la mayoría de las variables de entrada son de tipo categórico, mezcladas entre ordinales y nominales. Solo 4 de ellas son cuantitativas. A su vez, hemos verificado que agrupar algunas de las familias de variables de entrada categóricas, no es una opción.



Al observar la distribución de estas últimas, notamos que su distribución no es del todo normal, sus extremos tienden a tener picos. Para este estudio, se sugeriría probar con algún modelo de árbol de decisión, o bien de red neuronal perceptrón.

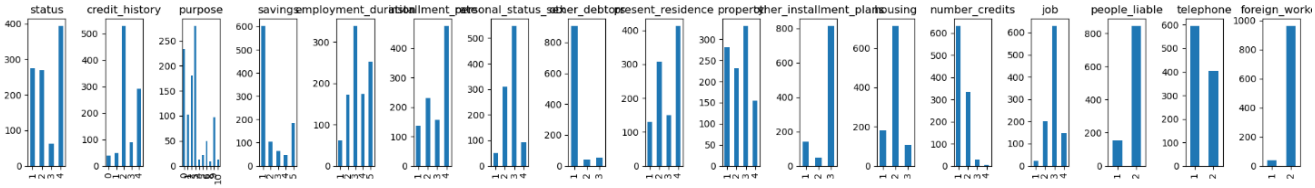
3) South German Credit Data Set:

	credit_risk	Perc
1	700	0.7
0	300	0.3

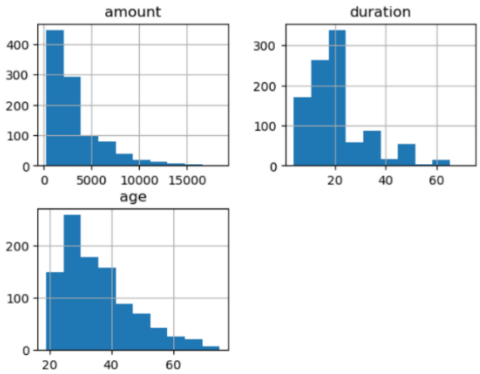


En este caso de la base de datos de riesgo de crédito de *South German Bank*, se puede observar clases de la variable independiente desbalanceadas. Por lo tanto, es recomendable pasar los datos a través de procesos de sobre muestreo y submuestreo.

Dado que 0 se considera un mal pagador de crédito y 1 un buen pagador de crédito se predecirá si el cliente será un mal pagador o no. Debido a que en este caso el otorgarle un crédito a alguien que podría potencialmente no pagar absolutamente nada, se considerará como alto riesgo los falsos negativos, y por ello lo que se quiere maximizar en este caso es la precisión, sin tener una referencia de alguna cuota de préstamos que se quiera dar, es difícil calcular el **k** óptimo de *recall*.



Lo primero que observamos en los datos de este caso de estudio diferente a los anteriores, es que aquí si podemos agrupar 2 de sus variables de entrada:

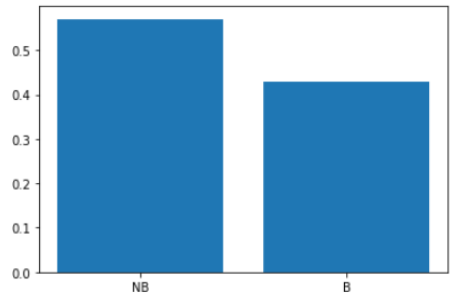


sex_other_debtors y *number_credits*. Esto si bien no disminuye mucho la capacidad de cómputo, puede que haga una diferencia significativa en las métricas de *recall* o precisión.

En cuanto a sus variables categóricas, todas ellas nuevamente se encuentran cargadas hacia un extremo, será necesario normalizarlas por medio de un *pipeline* de escalamiento *min_max*. Para este modelo, se recomienda utilizar un modelo de regresión logística, o red neuronal debido a la combinación de tipos de datos observados.

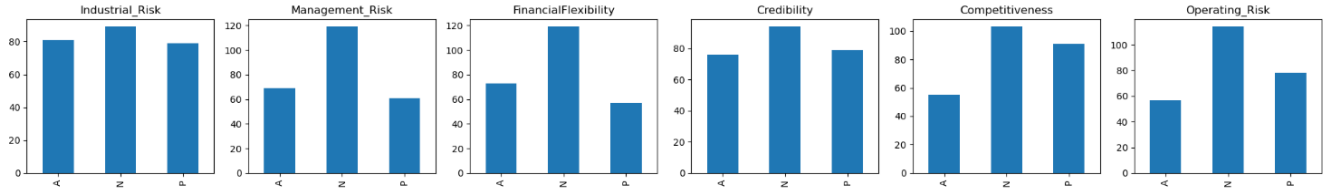
4) Qualitative Bankruptcy Data Set:

	Class	Perc
NB	142	0.570281
B	107	0.429719



En este caso, en la base de datos de riesgo de datos cualitativos de bancarrota, no se observan unas clases desbalanceadas, por lo tanto, no es necesario hacer sobre muestro o submuestreo.

Debido a esto debemos elegir que clase es la que se considerará positiva, para este caso *'bankruptcy'* se considerará como la clase positiva y *'non-bankruptcy'* como la clase negativa. Dado la naturaleza de predecir si una compañía estará o no estará en bancarrota implica mucho dinero y muchas inconveniencias para los empleados que se quedarían sin trabajos, quisiéramos minimizar las instancias en los cuales se deja a una empresa en banca rota, esto quiere decir que queremos maximizar el *recall*, sin embargo; queremos también mantener la precisión a un **k** alto debido a que un falso positivo también implica dinero ya que una falsa alarma de bancarrota puede llevar a tomar decisiones muy extremas.



Lo más notorio de este modelo respecto al resto, es que no cuenta con ninguna variable cuantitativa, de lo contrario cuenta con solo categóricas. La posibilidad de agrupar familias es imposible ya que se encuentran muy bien balanceadas todas ellas. Seguramente un modelo de árbol de decisiones será el mejor para este caso de estudio.

II) Resumen de Artículo

Una comparación experimental de las medidas de rendimiento para la clasificación

En nuestro caso nos enfocamos en resaltar lo que consideramos nos genera más valor y aporta en nuestro proceso de formación.

Teniendo en mente que el objetivo más importante de esta metodología es que nos permite evaluar la calidad de los métodos de aprendizaje y los modelos aprendidos, es fundamental no olvidar que debemos considerar que afuera hay una búsqueda constante en otros métodos que nos permitan hacer mejores elecciones más enfocadas en ares de espacialidad específicas, por ejemplo, IT o manufactura entre muchas otras.

Después de los análisis realizados con clústeres, relaciones, análisis de sensibilidad, umbral de clase, calidad de clasificación / separabilidad, rendimiento en C/S etc. , y de las correspondientes definiciones y experimentos, además de las relaciones entre métrica, una taxonomía y ordenación evidenciamos que en la mayoría de los casos lo que nos permite elegir la medida más adecuadas para la aplicación especifica de nuestro caso de estudio o análisis, apoyar estrategias de innovación en el momento de una toma de decisión .

Queremos resaltar la importancia de hacer una evaluación correcta de los modelos, identificar las diferencias que existen entre evaluar un modelo de regresión con error absoluto o con error cuadrático. y definitivamente lo más importantes es que contamos con un número importante de medidas para evaluar los clasificadores.

Para este caso se definió en tomar métricas para evaluar calificadores con precisión, Medida F, Tasa de rango, AUC, Brier entre otras y que se usaran 18 métricas diferentes clasificadas en 3 familias. Se buscaron métricas con las que buscaron: Un modelo que minimice el número de errores y que sean las más apropiadas para conjuntos de datos balanceados o desbalanceados, para detección de señales o fallas. Otras que fueran más útiles cuando queremos evaluar la fiabilidad de los clasificadores y que esto es fundamental para los modelos realicen correctamente una fusión ponderadora de los modelos. Entre muchas otras que nos permiten que lo clasificadores se utilicen para seleccionar los mejores Instancias de un conjunto de datos, los sistemas de recomendación, la detección de fraudes, el filtrado de *spam*, Al final de todo este análisis lo más importante es interpretar los resultados que en este caso nos muestran que la mayoría de estas métricas realmente miden cosas diferentes y en muchas situaciones la elección realizada con una métrica puede ser diferente de la elección realizada con otra , los resultados obtenidos en una medida podrían extrapolarse a otras medidas.

Otro aspecto por considerar es el análisis se completa con un conjunto de experimentos para cuantificar la sensibilidad a cuatro rasgos importantes que están presentes en algunas medidas, pero no están presentes en otras. Estos rasgos son la elección óptima del umbral de clase, la calidad de la clasificación / separabilidad, el rendimiento de la calibración y la sensibilidad (o, por el contrario, la solidez) a los cambios en la distribución de la clase anterior. A partir de este análisis, podemos cuantificar las relaciones sobre estas 'dimensiones', lo que es un complemento muy útil para los resultados del análisis de correlación.

Lo que se encontró en las conclusiones más interesantes de este primer trabajo experimental es que comparo las métricas de evaluación de los clasificadores más utilizadas, obteniendo conclusiones que involucran interdependencia y sensibilidad de las medidas, aunque sabemos que los estudios continúan constantemente, lo más importante es mantenernos actualizados con lo último que vamos teniendo como recurso.

Finalmente, frente a las conclusiones del estudio de relaciones, es que estructuraron un proceso progresivo y evolutivo para analizar las correlaciones que los llevo a validar la existencia de similitudes entre las medidas, sin embargo; las diferencias también son determinantes en el mismo. Lo que nos permite identificar que no hay nada concluyente, simplemente debemos buscar aplicar la que más convenga de acuerdo con la información y el caso de uso que vayamos a analizar.

C. Ferri, J. Hernández-Orallo, R. Modroiu, An experimental comparison of performance measures for classification, Pattern Recognition Letters, Volume 30, Issue 1, 2009, Pages 27-38.