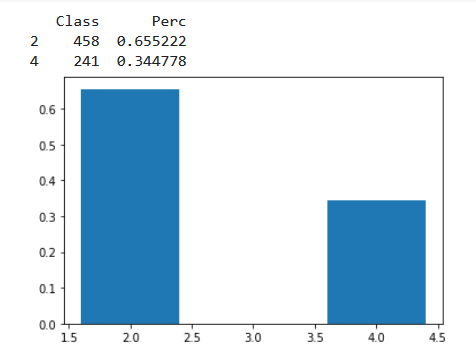
# 1er base de datos - Breast Cancer data



Podemos observar que 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 sub muestreo o sobre muestreo sin embargo 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.

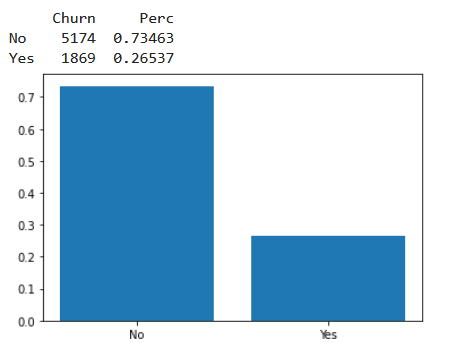
Chart, scatter chart

Description automatically generated

Al evaluar su distribucion, vemos que en 6 de lasa 8 variables de entrada (sample coude number debe de ser eliminada, ya que actua como simple index), hay una distribucion para nada normal (cargada a la izquierda), sera necesario realizar un pipeline que escale estos valores segun el maximo y minimo de todo el datset X.

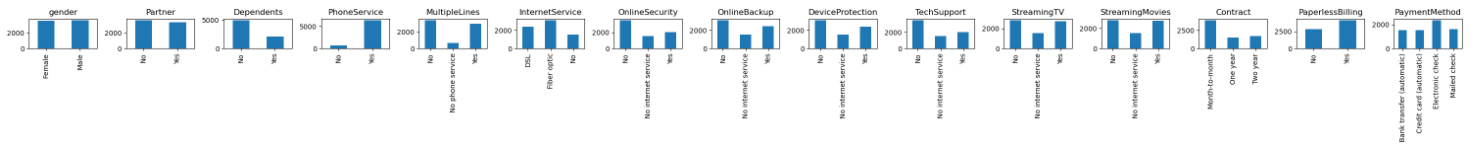
Debido a que la mayoria de las variables de entrada son de tipo cuantitativo, algun modelo de regresion lineal deberia de ser el mejor en predecir de forma correcta la clase de salida, sin requerir demasiada capacidad de computo.

**2da base de datos - Telco Customer Churn**



Podemos observar en esta base de datos de abandono de clientes de una base de telecomunicación tiene clases de variable independiente desbalanceadas, por lo tanto es recomendable pasar los datos a través de procesos de sobre muestreo y sub muestreo.

Desde el punto de vista de negocio, según la literatura mantener a un cliente es mucho más barato que obtener nuevos clientes, por lo tanto desde el punto de vista de negocio es muy importante el mantenimiento de los clientes, por lo tanto en este caso se quisiera maximizar el recall para poder identificar la mayoría de los clientes que pudieran abandonar. Como contrapeso 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 tendríamos que evaluar que tan caro podrían ser los falsos positivos.

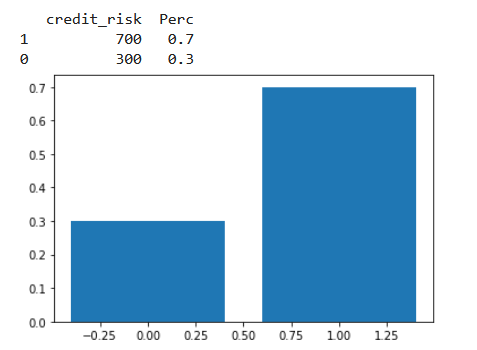


Chart, histogram

Description automatically generatedPara el estudio de Telco Churn, vemos que la mayoria 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 opcion.

Al observar la distribucion de estas ultimas, notamos que su distribucion no es del todo normal, sus extremos tienden a tener picos. Para este estudio, se sugeriria probar con algun modelo de arbol de decision, o bien de red neuronal perceptron.

# 3er base de datos - South German Credit



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 sub muestreo.

Dado que 0 se considera un mal pagador de crédito y 1 un buen pagador de crédito, vamos a querer predecir si el cliente será un mal pagador o no, dado 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, por lo tanto 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.

Chart, bar chart

Description automatically generated

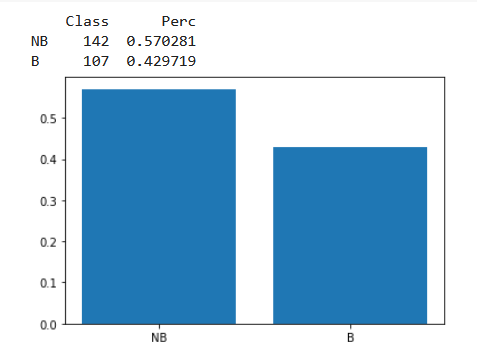
Chart, histogram

Description automatically generatedLo primero que observamos en los datos de este caso de estudio diferente a los anteriores, es que aqui si podemos agrupar 2 de sus variables de entrada:

sex\_other\_debtors y number\_credits. Esto si bien no disminuye mucho la capacidad de computo, puede que haga una diferencia significativa en las metricas de recall o precision.

En cuanto a sus variables categoricas, todas ellas nuevamente se encuentran cargadas hacia un extremo, sera necesario escalarlas (normalizarlas) por medio de un pipeline de escalamiento min\_max. Para este modelo, se recomienda utilizar un modelo de regresion logistica, o red neuronal debido a la combinacion de tipos de datos observados.

# 4ta base de datos Qualitative Bankruptcy data



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

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 naturelaza de predecir si una compañía estará o no estrá en banca rota implica mucho dinero y muchas incoveniencias 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 banca rota puede llevar a tomar decisiones muy extremas.

Chart, bar chart, histogram

Description automatically generated

Lo mas notorio de este modelo respecto al resto, es que no cuenta con ninguna variable cuantitativa, sino puras categoricas. La posibilidad de agrupar familias es imposible ya que se encuentran muy bien balanceadas todas ellas. Seguramente un modelo de arbol de decisiones sera el mejor para este caso de estudio.