**PROYECTO # AR-T1165**

**ELABORACIÓN DE ESPECIFICACIONES TÉCNICAS DEL SISTEMA DE ANTICIPACIÓN DE DEMANDA LABORAL**

**PRESTAMO BID 4648 OC/AR**

**MINISTERIO DE SALUD Y DESARROLLO SOCIAL DE LA NACIÓN**

**DIRECCIÓN GENERAL DE PROYECTOS ESPECIALES Y COOPERACIÓN INTERNACIONAL**

Informe Técnico

(Productos 1 y 2)

Análisis exploratorio de datos de Hacemos Futuro y modelos predictivos

**Índice**

[Objetivo 3](#_Toc5961240)

[Objetivo específico 3](#_Toc5961241)

[Los datos 3](#_Toc5961242)

[Prototipo 1: Pre procesamiento e integración de los datasets 4](#_Toc5961243)

[Prototipo 2: Visualizaciones y análisis de datos 5](#_Toc5961244)

[Gráficos de Barra 5](#_Toc5961245)

[Gráficos de líneas 10](#_Toc5961246)

[Histogramas o gráficos de densidades 11](#_Toc5961247)

[Prototipo 3: Modelos predictivos 13](#_Toc5961248)

[Resultados 16](#_Toc5961249)

[Conclusiones 18](#_Toc5961250)

[Bibliografía 18](#_Toc5961251)

## Objetivo

Realizar un prototipo de visualizaciones de datos y de modelos predictivos a partir de los datos proveídos por el equipo de Hacemos Futuro.

## Objetivo específico

1. Prototipo de pre procesamiento e Integración de los datasets proveídos por el equipo de Hacemos Futuro.
2. Prototipo de visualizaciones y análisis de datos.
3. Prototipo de modelos predictivos

## Los datos

Se cuenta con los siguientes archivos proveídos por el equipo de Hacemos Futuro:

* Dataset de Titulares: este dataset contiene los datos asociados a los beneficiarios del plan. El dataset tiene registros como el ID del beneficiario, sexo, edad, momento de ingreso al plan y varios registros declarativos como oficio y formación entre otros.
* Dataset de Empleo: este dataset contiene los datos asociados al empleo de los beneficiarios con empleo al momento de la toma de datos. El dataset tiene registros como el ID del beneficiario, la fecha de la toma de datos, el sector en el que se desempeña, el sueldo, etc.
* Anexo: este archivo cuenta con los códigos asignados a los distintos niveles educativos (Nivel Inicial: 1 - Primario incompleto: 2 - Primario completo: 3 - EGB incompleto: 4 - etc.) y a la asistencia educativa (Nunca asistió: 1 - No asiste, pero asistió: 2 - asiste: 3 - etc.)

## Prototipo 1: Pre procesamiento e integración de los datasets

Este prototipo consiste en la integración de ambos datasets, la limpieza de los datos y el pre procesamiento de los mismos. Aquí se busca poner a punto los datos para que posteriormente puedan ser accedidos por los métodos de visualización y predicción.

A continuación describiremos los pasos de pre procesamiento realizados:

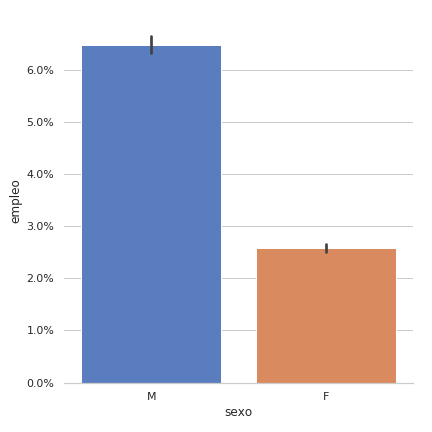
* Selección de registros: por cada titular se tienen registros correspondientes al mes de Mayo, Agosto y Diciembre 2018. Dado que se busca tener una única instancia (i.e. fila en el dataset) por titular, para cada titular se seleccionó el último registro no vacío en cada atributo del dataset de Titulares. Con respecto al dataset de Empleo, se tuvieron en cuenta únicamente los datos de diciembre, y en los casos de titulares con más de un trabajo, se seleccionó el trabajo con el sueldo de mayor monto.
* Integración de datasets: Se unieron ambos dataset según ID de titular. Los titulares que no se encontraban presentes en el dataset de trabajo fueron identificados como sin empleo.
* Reprocesamiento y limpieza de datos:
  + El registro de oficios se encuentra expresado en lenguaje natural, por ejemplo ["Artesano (ACTIVIDADES CULTURALES)"]. Para extraer el oficio se utilizó una expresión regular que identifique texto en mayúsculas que se encuentre entre paréntesis. En el ejemplo se extraería *ACTIVIDADES CULTURALES*
  + Se calculó la edad y los años bajo algún plan a partir de los atributos “FechaNac” y “MesAlta”
  + Se identificó como vivienda en barrio popular a los titulares que presentaron registros no vacíos en el atributo “leyenda\_renabap”
  + El estado de salud (atributo “estado\_saludo”) se transforme de variables categóricas (Muy mala, Mala, Regular, Buena, Muy buena, Excelente) a variable numéricas (1, 2, 3, 4 y 5).
  + los registros dicotómicos cuyas respuestas esperadas eran *No* y *Si* (como por ejemplo “tiene\_oficios” o “busco\_trabajo”) fueron reemplazados por booleanos, 0 y 1 respectivamente y por *Nan* en caso de aparecer otra respuesta posible. Esto unifica respuestas vacías con respuestas espurias como “[]”. La utilización de booleanos es necesario para el entrenamiento de modelos predictivos
  + Los valores de “MaxNivelEducativo” fueron agrupados asociándose por ejemplo “EGB incompleto” con “Primario incompleta” e identificados según los Id propuestos en el archivo Anexo.

## Prototipo 2: Visualizaciones y análisis de datos

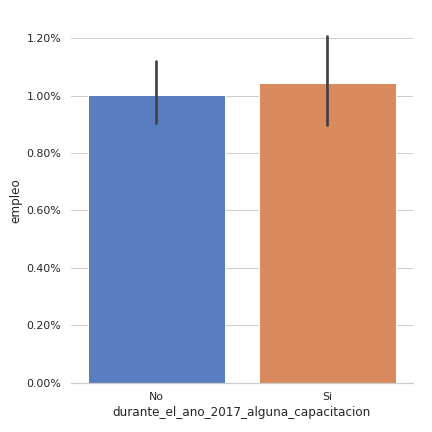
Este prototipo consiste en la generación de visualizaciones, que apunten a extraer conocimiento e intuición de las variables asociadas al empleo de los titulares. Este prototipo corresponde a visualizaciones de los datos pre-procesados en el *Prototipo 1.*

### Gráficos de Barra

Estos gráficos apuntan representar gráficamente un conjunto de datos. En este caso particular será de utilidad para comparar el porcentaje de titulares con empleo según los valores que tomen las variables categóricas de interés. Por ejemplo, en la Figura 1 comparamos el porcentaje de empleo según la variable *sexo.* Esta figura muestra que el porcentaje de mujeres con empleo es menos de la mitad de los hombres con empleoLas barras de error muestran el intervalo de confianza al 95%. Utilizando el mismo tipo de gráficos en la figura 2 se compara el porcentaje de empleo entre los titulares que realizaron o no alguna capacitación durante el 2017. Este grado evidencia que no hay una asociación apreciable entre el empleo y la realización de capacitaciones durante el 2017.

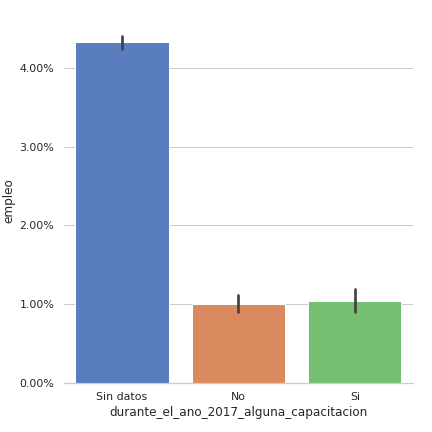


*Figura 1: Porcentaje de empleo según sexo*

**

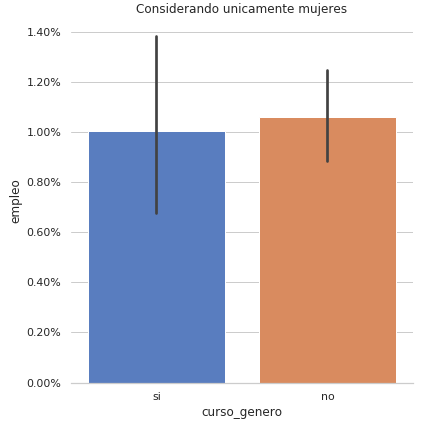
*Figura 2: Porcentaje de empleo de los titulares que realizaron alguna capacitación durante el 2017 comparado con los que no realizaron ninguna capacitación*

Resulta importante aclarar que no todos los titulares respondieron en la encuesta el ítem de la capacitación, y dado que se buscaba estudiar la asociación de esta variable con el empleo se excluyó de la Figura 2 las respuestas sin datos. En caso de incluir los beneficiarios con respuestas vacías en este campo se observa que los mismos cuentan con un porcentaje de empleo mucho mayor al de los titulares que contestaron.



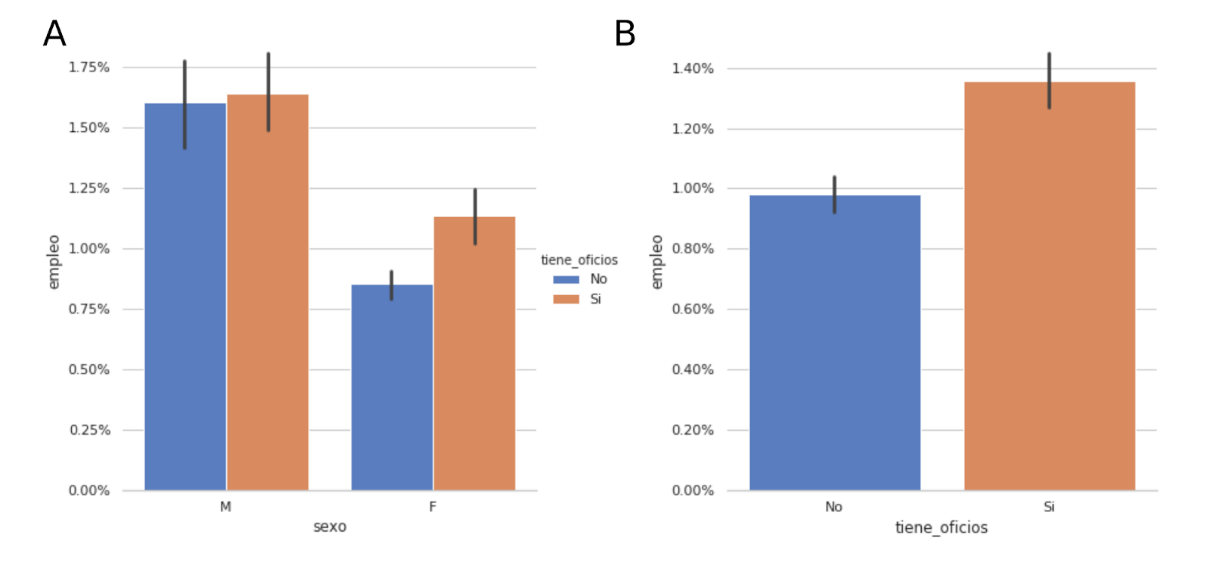
*Figura 3: Porcentaje de empleo de los titulares que realizaron alguna capacitación durante el 2017 comparado con los que no realizaron ninguna capacitación y los que no contestaron este ítem.*

Otro elemento que se puede considerar al realizar gráficos de barras es el de tomar subconjuntos del dataset. Por ejemplo para la Figura 4 consideramos únicamente a las mujeres y mostramos el porcentaje de empleo según si las beneficiarias tomaron o no una capacitación en temas de género.



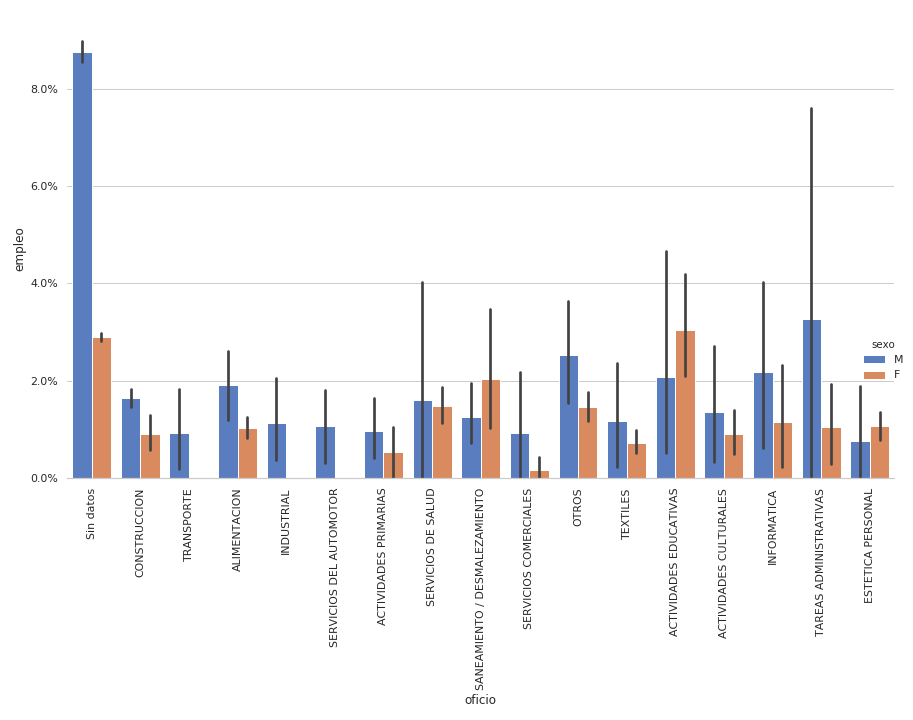
*Figura 4: Porcentaje de empleo de las titulares que realizaron la capacitación en temas de género comparado con las que no realizaron la capacitación. Para este gráfico sólo se consideraron las titulares de sexo femenino*

Este tipo de gráficos permite también hacer comparaciones en múltiples variables. Por ejemplo en la Figura 5A se observa el porcentaje de empleo según si tiene oficio y según sexo. Este gráfico muestra que las mujeres con oficios tienen un mayor porcentaje de empleo que las que no tienen oficio, mientras que en los hombres no se observa asociación entre el oficio y el empleo. Resulta relevante resaltar que si se hubiese separado únicamente por oficio (ver Figura 5b) se hubiese pasado por alto el hecho de que la asociación entre el oficio y el empleo es significativa únicamente en las mujeres.



*Figura 5: A) Porcentaje de empleo de los titulares según si tiene oficio y según sexo. B) Porcentaje de empleo de los titulares según si tiene oficio.*

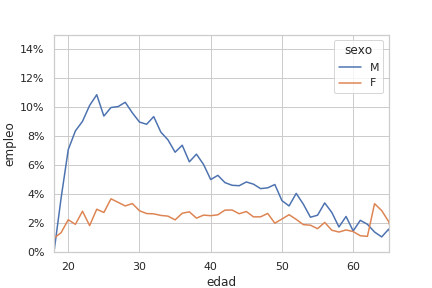
En la figura 6 mostramos otro ejemplo de análisis de múltiples variables, donde se graficó el porcentaje de empleo según el oficio declarado y el sexo. Aquí nuevamente se observa que los titulares con el registro vacío cuentan con un mayor porcentaje de empleo. Cabe aclarar que las categorías con barra de error amplias corresponden a casos con muy pocos datos. Por ejemplo solo hay 3 hombres con oficio de *Tareas Administrativas*, por lo que la comparación entre hombres y mujeres en este oficio resulta poco significativa



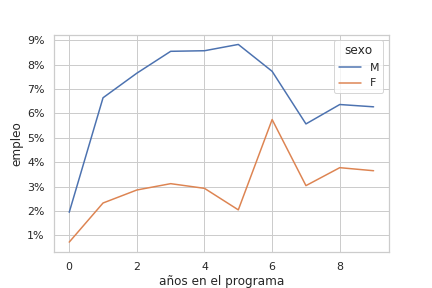
*Figura 6: Porcentaje de empleo de los titulares según oficio declarado y según sexo.*

### Gráficos de líneas

Estos gráficos apuntan representar gráficamente un conjunto de datos a partir de una variable con valores numéricos. Estos son particularmente útiles para representar series de tiempo. Por ejemplo, en la Figura 7 mostramos el porcentaje de empleo en función de la edad y por el sexo*.* Esta figura evidencia que alrededor de los 25 años se encuentra la mayor brecha de género en cuanto al porcentaje de empleo. La figura 8 es otro ejemplo de gráfico de líneas, donde se muestra el porcentaje de empleo en función de los años en el programa y el sexo.



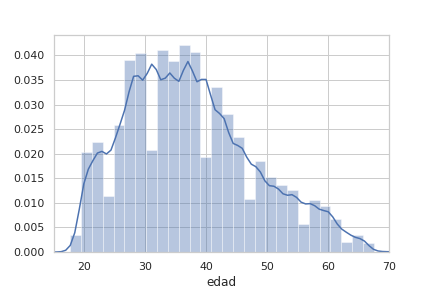
*Figura 7: Porcentaje de empleo de los titulares en función de la edad y por sexo.*



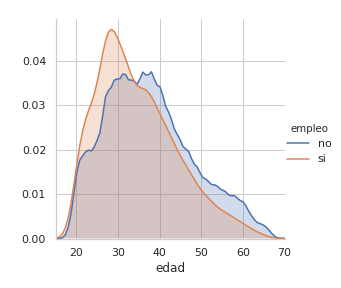
*Figura 8: Porcentaje de empleo de los titulares en función de los años en el programa y por sexo.*

### Histogramas o gráficos de densidades

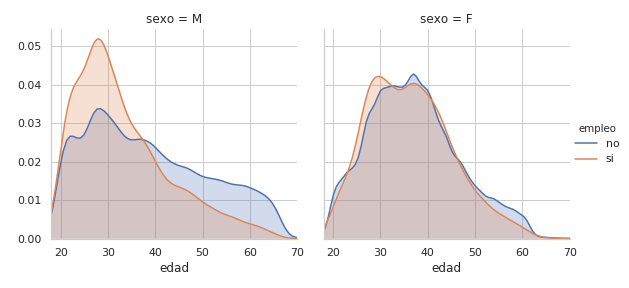
Estos gráficos sirven para representar cómo se distribuyen los datos a lo largo de una variable. Por ejemplo en la Figura 9 mostramos cómo se distribuye la edad de los titulares, donde se observa que las edades más frecuentes están alrededor de los 30-40 años. Esta figura muestra tanto el histograma (barras) como la densidad (línea). Los valores están normalizados de tal manera que el área debajo de la curva es 1. Esto resulta particularmente útil a la hora de comparar distribuciones. Por ejemplo en la figura 10 mostramos la diferencia entre las distribuciones de los titulares que tienen y que no tienen empleo. Aquí se puede observar una ligera diferencia entre las distribuciones, donde se evidencia que las edades más frecuentes entre los titulares empleados están alrededor de los 25-35 años. Más aún, si grafican las distribuciones de la Figura 10 separando por género (ver Figura 11), se puede observar que las diferencias en las edades de los que trabajan y los que no se ve fuertemente marcada en los hombres, a diferencia de las mujeres donde no se observa este comportamiento.



*Figura 9: distribución de la edad de los titulares*



*Figura 10: Comparación entre la distribución de la edad de los titulares que tienen y que no tienen empleo*

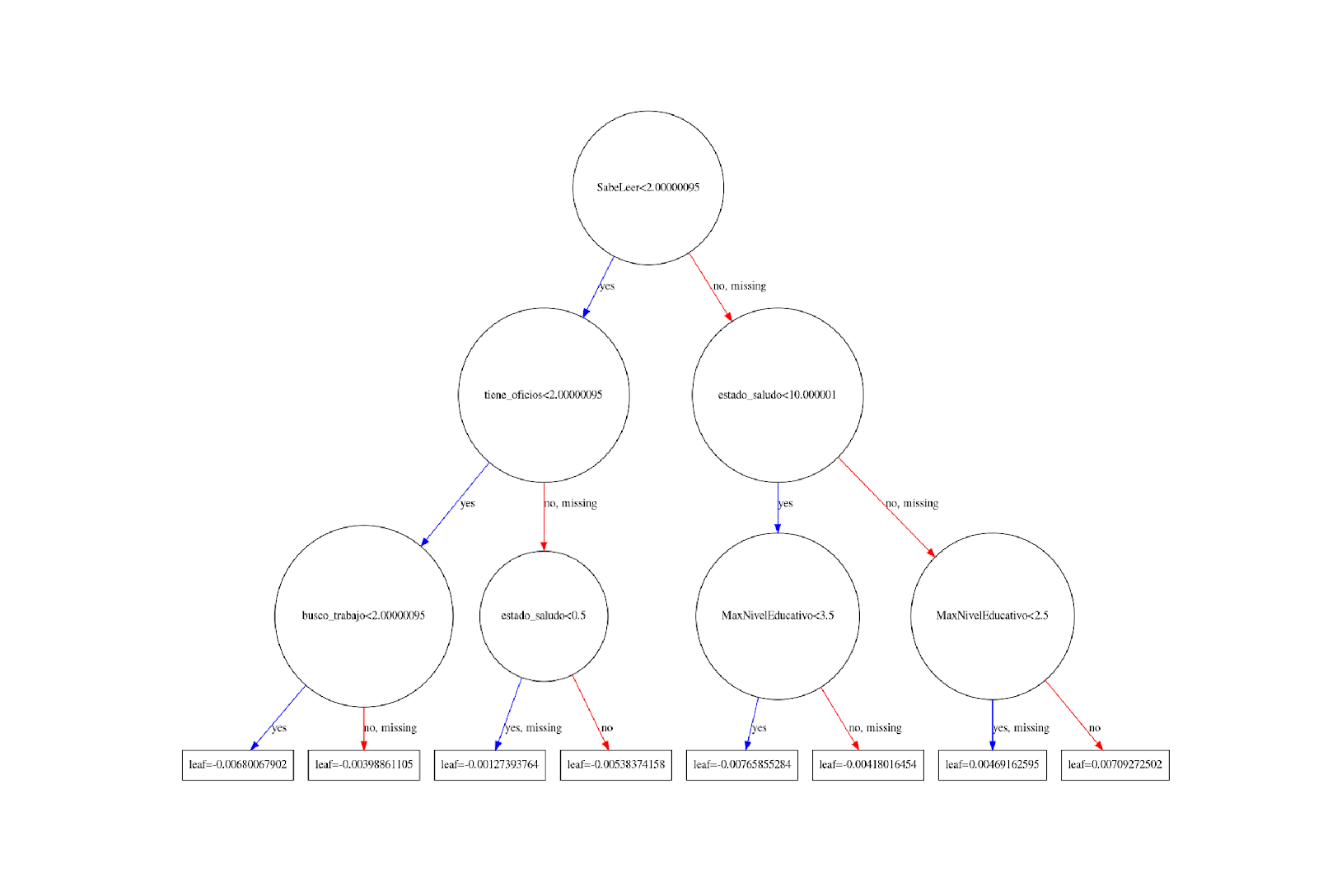


*Figura 11: Comparación entre la distribución de la edad de los titulares que tienen y que no tienen empleo, según sexo*

## Prototipo 3: Modelos predictivos

Este prototipo consiste en la implementación de modelos de Aprendizaje Automático, que apunten a inferir (*predecir*) si un titular está empleado, a partir de los atributos del titular (columnas del dataset). Para esta tarea en particular se requieren modelos de clasificación supervisada, las cuales aprenden de un conjunto de datos (datos de entrenamiento) relaciones funcionales entre los atributos (columnas del dataset) y la variable a predecir. Para este prototipo se usará *XGBoost* [1], que es un modelo de *ensembles* de árboles de decisión muy usado en la literatura.

A modo de ejemplo, en la Figura 12 mostramos un esquema de uno de los árboles de decisión aprendido por el modelo XGBoost. En el gráfico se muestra como se suceden las cadenas de decisiones dicotómicas a la que es sometido cada titular para finalmente asignarle un score (*leaf* en la figura). El XGBoost entrena *n* árboles (*n* =300 en nuestro modelo) por lo que el *score* final de cada titular está dado por la suma de todos los *scores* de los *n* árboles. Titulares con *scores* positivos son clasificados como *con empleo*, y coso contrario son clasificados como *sin empleo.* Es importante notar que las decisiones incluyen la existencia de registros sin datos. Por ejemplo la primera decisión separa en dos grupos de titulares según el valor del atributo *Sabe leer*: (1) valores menores a 2 (2) valores mayores a 2 + valores sin datos (*missing*). Dado que el registro de si sabe leer o no es representado con un *1* o un *0* respectivamente, esta operación está separando a los titulares que tienen el registro vacío de los que no.



*Figura 12: Esquema de uno de los árboles de decisión entrenados con XGBoost* Metodología

Para este prototipo se seleccionó el siguiente conjunto de atributos: {sexo, educacion\_especial - MaxNivelEducativo - Asiste - SabeLeer - barrio\_popular - busco\_trabajo - tiene\_oficios - durante\_el\_ano\_2017\_alguna\_capacitacion - posibilidades\_de\_conseguir\_un\_empleo\_formal - edad - anos\_en\_el\_programa - estado\_saludo - LineaAnterior - oficio - tematicas\_continuacion}

Los atributos categóricos como *LineaAnterior*, *oficio* y *tematicas\_continuacion* requieren ser introducidos en el modelo como atributos *dummy* (booleanos) para ser incorporados en los modelos predictivo. Por ejemplo, el atributo *LineaAnterior* que tiene como posibles valores *AT, DB* y *EH,* es descompuesto en los atributos booleanos *LineaAnterior\_AT*, *LineaAnterior\_DB* y *LineaAnterior\_EH* que toman valores de *0* o *1* según el valor el valor del atributo original *LineaAnterior.*

Luego, el dataset es separados en dos partes: el dataset de *entrenamiento* (80% de los datos) y el de *testeo* (20% de los datos)*.*

Sobre el set de entrenamiento se realizó un random-search en el espacio de hiper parámetros para seleccionar la combinación de los mismos que produce un mejor performance, usando un esquema de *5-fold cross validation*. Como método de scoring se utilizó el *f-score* y se implementó un weight scaling de los casos positivos para abordar el desbalance de los datos (hay solo un %3,85 de titulares con empleo)

Luego de seleccionar el conjunto de parámetros óptimos, se re-entrena el modelo con todo el set de entrenamiento y se clasifican los titulares del set de testeo.

### Resultados

Los resultados del clasificador pueden observarse en la matriz de confusión de la Tabla 1.

|  |  |  |
| --- | --- | --- |
|  | clasificado: con empleo | clasificado: sin empleo |
| con empleo | 1438 (True Positive) | 697 (False Negative) |
| sin empleo | 8192 (False Positive) | 47176 (True Negative) |

*Tabla 1: Matriz de confusión*

Esta tabla nos muestra que de los 2135 titulares con empleo en el set de testeo, 1438 fueron correctamente clasificados como *con empleo* (*True Positive*)*,* mientras que 697 fueron incorrectamente clasificados como *sin empleo* (*False Negative*). Por otro lado, de los 55368 titulares sin empleo en el set de testeo, 47176 fueron correctamente clasificados como *sin empleo* (*True Negative*), mientras que 8192 fueron *incorrectamente clasificados* como *con empleo* (False Positive).

A partir de estos valores se pueden computar el *recall*, el *precision* y el *f-score*:

* El *recall* computa la fracción de los titulares con empleo, correctamente identificados

recall = *True PositiveTrue Postitive* +*False Negative* =14381438+697=0,673

* El *precision* computa la fracción de los clasificados como *con empleo* que efectivamente fueron clasificados correctamente

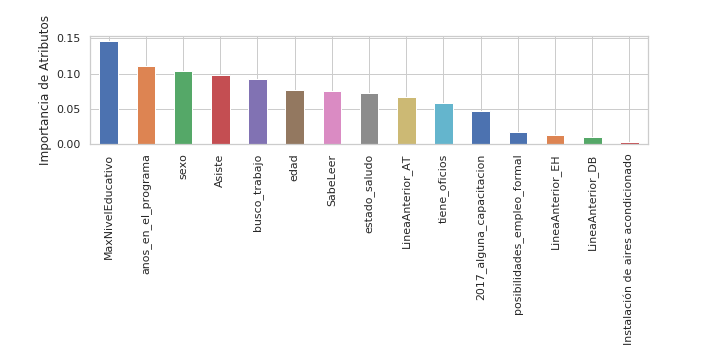
precision = True PositiveTrue Postitive +False Positive =14381438+8192=0,149

* El f-score es el promedio armónico entre el recall y el precision. El promedio armónico es un tipo de promedio que tiende a darle más peso a los valores más pequeños.

f-score =2 precision x recallprecision+recall =20,149 x 0,6730,149+0,673=0,244

Con el fin de entender la capacidad predictiva de este modelo, se armó un clasificador *dummy* que predice *con/sin* *empleo* de manera azarosa, respetando la distribución de las clases del set de entrenamiento. Este modelo obtuvo un recall, precision y f-score de aproximadamente 0.04. Este resultado contextualiza el precision obtenido por nuestro modelo, que a pesar de ser un valor bajo, resulta más de 3 veces mayor que lo esperado por azar.

Finalmente, utilizamos el método de estimación de la importancia de los atributos que tiene implementado nativamente el XGBoost. En la Figura 13 mostramos los 15 atributos que le resultaron más relevantes al XGBoost para el entrenamiento del modelo. El índice de importancia de atributos es calculado nativamente por el XGBoost a partir de la cantidad de veces que cada feature es seleccionado como nodo de decisión.



*Figura 13: Importancia de atributos del modelo XGBoost*

Este resultado sirve a su vez como guía para seguir investigando los factores asociados al empleo de los titulares.

## Conclusiones

En el prototipo 3, entrenamos el algoritmo XGBoost para la clasificación automática de titulares empleados. Aquí mostramos que el dataset cuenta con atributos predictivos del empleo. En este prototipo utilizamos solo un subconjunto de los datos existentes, 16 de los más de 100 atributos presentes en el dataset, y probamos un único algoritmo de clasificación. En este contexto, creemos que hay mucho espacio para mejorar el modelo, especialmente por la inclusión de atributos, tanto de los existentes en el dataset como de nuevos atributos socioeconómicos y de producción que puedan extraerse fuentes externas. Este tipo de modelos al ser alimentado con series temporales pueden ser utilizados para predecir escenarios futuros, pudiendo utilizarse para guiar políticas públicas. Más aún, estos modelos pueden utilizarse como simuladores de escenarios futuros, computando el impacto en el empleo al introducirse distintos valores en el modelo.

Por lo expuesto, recomendamos enfocar los esfuerzos de la primera etapa en lae recolección, sistematización y explotación de datos en un Sistema experto de análisis de oferta y demanda laboral basado en tableros de comando para luego evolucionar a una segunda etapa que utilice modelos predictivos a partir de los datos disponibles.

## Bibliografía

[1] Chen, T., & Guestrin, C. (2016) Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.