“HR Analytics: Job Change of Data Scientists”

Juan Ignacio Ron

https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists

Índice

Introducción / Problema

Una compañía vinculada al mundo del Big Data y Data Science quiere ampliar su equipo de científicos de datos.

Para ello, busca candidatos entre aquellas personas que han asistido a los cursos de capacitación que la propia empresa brinda. Estos cursos son muy populares y tienen una gran cantidad de asistentes.

La compañía quiere saber **cuáles de sus alumnos estarían dispuestos a trabajar en ella** o bien **estarían interesados en cambiar de trabajo**. Esto ayudaría a reducir los tiempos y costos de contratación, así como también facilitaría la capacitación y la asignación de roles y tareas para los nuevos ingresantes.

El proceso de suscripción y admisión a las capacitaciones brinda información demográfica, educativa y laboral de los alumnos y potenciales candidatos.

¿Con qué información contamos?

Variables del Dataset*

```
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   enrollee_id             19158 non-null  int64
 1   city                    19158 non-null  object
 2   city_development_index  19158 non-null  float64
 3   gender                  14650 non-null  object
 4   relevent_experience     19158 non-null  object
 5   enrolled_university     18772 non-null  object
 6   education_level         18698 non-null  object
 7   major_discipline        16345 non-null  object
 8   experience              19093 non-null  object
 9   company_size            13220 non-null  object
 10  company_type            13018 non-null  object
 11  last_new_job            18735 non-null  object
 12  training_hours          19158 non-null  int64
 13  target                  19158 non-null  float64
dtypes: float64(2), int64(2), object(10)
```

- city → Código de la ciudad de origen del alumno.
- city_development_index → Índice de desarrollo de la ciudad (estandarizado de 0 a 1).
- enrolled_university → Tipo de cursada universitaria, en caso de estar inscripto (tiempo completo, parcial o ninguna).
- last_new_job → Diferencia en años entre el último empleo y el actual
- target → Variable objetivo:
  0 – No busca empleo.
  1 – En búsqueda de empleo

*__Nota__: para esta sección se trabajó con el set de entrenamiento

Variables del Dataset*

- Diez de las doce variables son categóricas, dejando de lado la variable target. En algunos casos, la cantidad de valores posibles es muy alta: ‘city’ tiene 127 valores posibles y ‘experience’ 22.

- ‘city development index’, una de las variables numéricas, está directamente relacionada a la variable ‘city’, dado que cada ciudad tiene asociado un índice determinado. En un futuro deberemos analizar si es necesario contar con las dos variables.

- Se observa una gran cantidad de **datos faltantes**. Este será un gran problema a resolver para un futuro análisis predictivo.

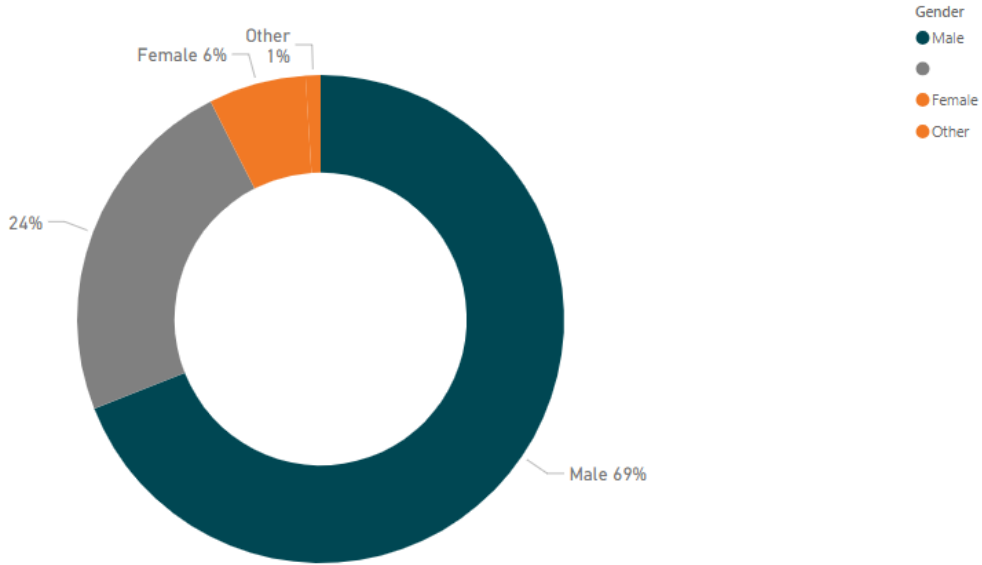
Porcentaje de nulos	
enrollee_id	0.00
city	0.00
city_development_index	0.00
gender	23.53
relevent_experience	0.00
enrolled_university	2.01
education_level	2.40
major_discipline	14.68
experience	0.34
company_size	30.99
company_type	32.05
last_new_job	2.21
training_hours	0.00
target	0.00

Entendiendo los datos

Trataremos de caracterizar el dataset a partir de un análisis exploratorio básico de las variables. Como un primer abordaje de la información que estamos tratando, las analizaremos de manera unidimensional.

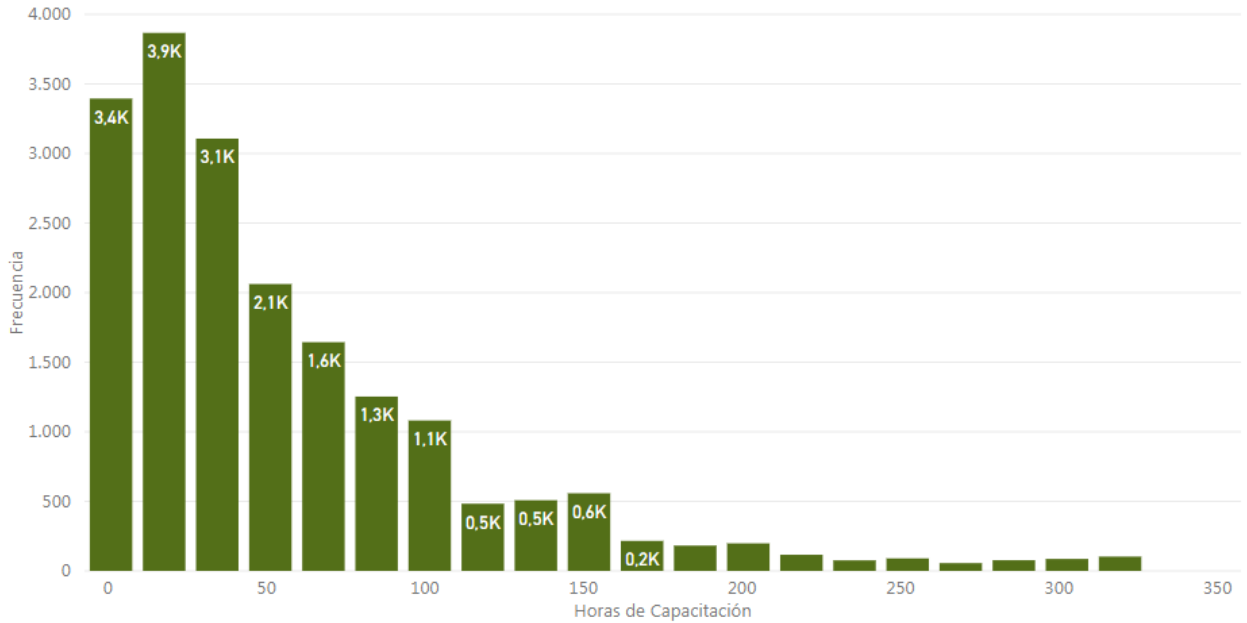
El dataset viene separado en set de entrenamiento y de testeo, para la aplicación de algoritmos de machine learning. De aquí en adelante (sólo hasta llegada dicha instancia) trabajaremos sobre los datos del set de entrenamiento.

Alumnos por Género



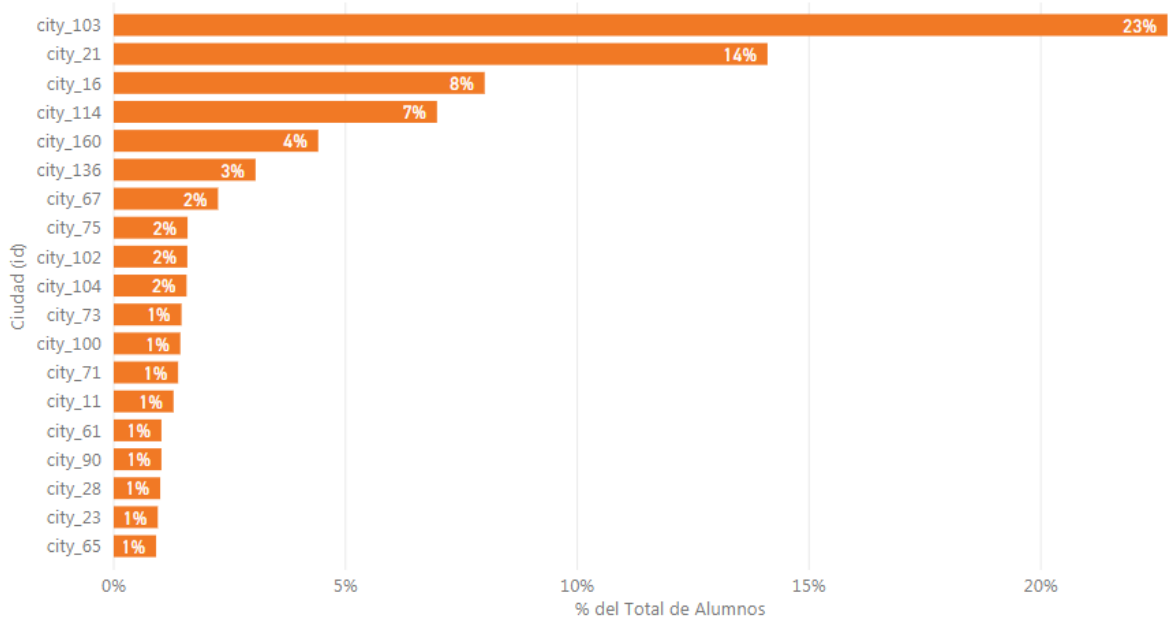
- Hay un total de 19.160 alumnos registrados, de los cuales un 69% son hombres.
- 24% de las datos de género están en blanco. La proporción de faltantes procede en mayor medida de las ciudades de menor índice de desarrollo.
- No se observan diferencias en la distribución de géneros en función de otras variables (demográficas, educativas o económicas).

Horas de capacitación



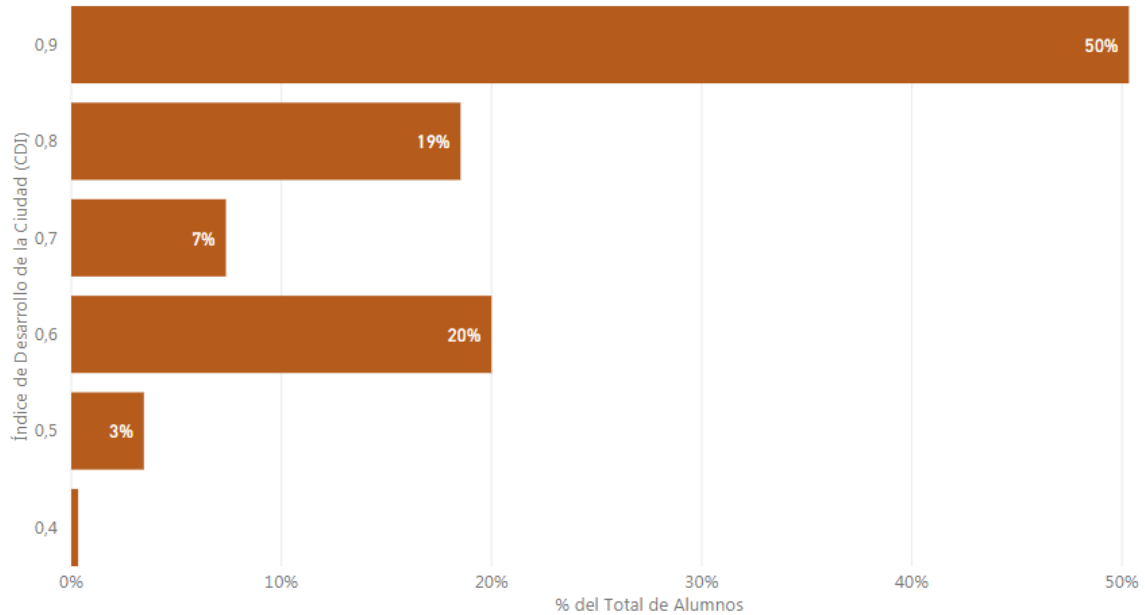
- Cada alumno se capacita 65,37 horas en promedio.
- La mediana es de 47 horas de capacitación y el Rango Intercuartílico (Q1-Q3) está entre 23 y 88 horas.
- La distribución de las horas no cambia cuando se mide en función otras las variables.

Alumnos por Ciudad de Origen



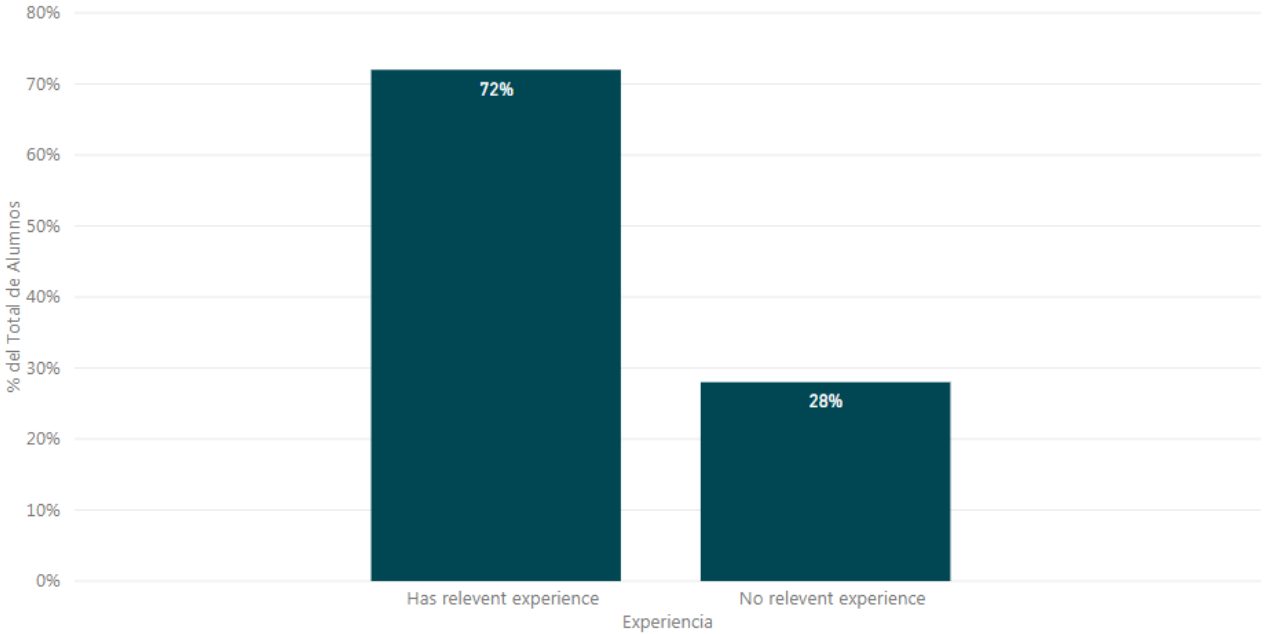
- Los alumnos provienen de 127 ciudades distintas.
- El 50% de los alumnos están concentrados en las cuatro primeras ciudades del ranking.

Alumnos por Indice de Desarrollo de su Ciudad



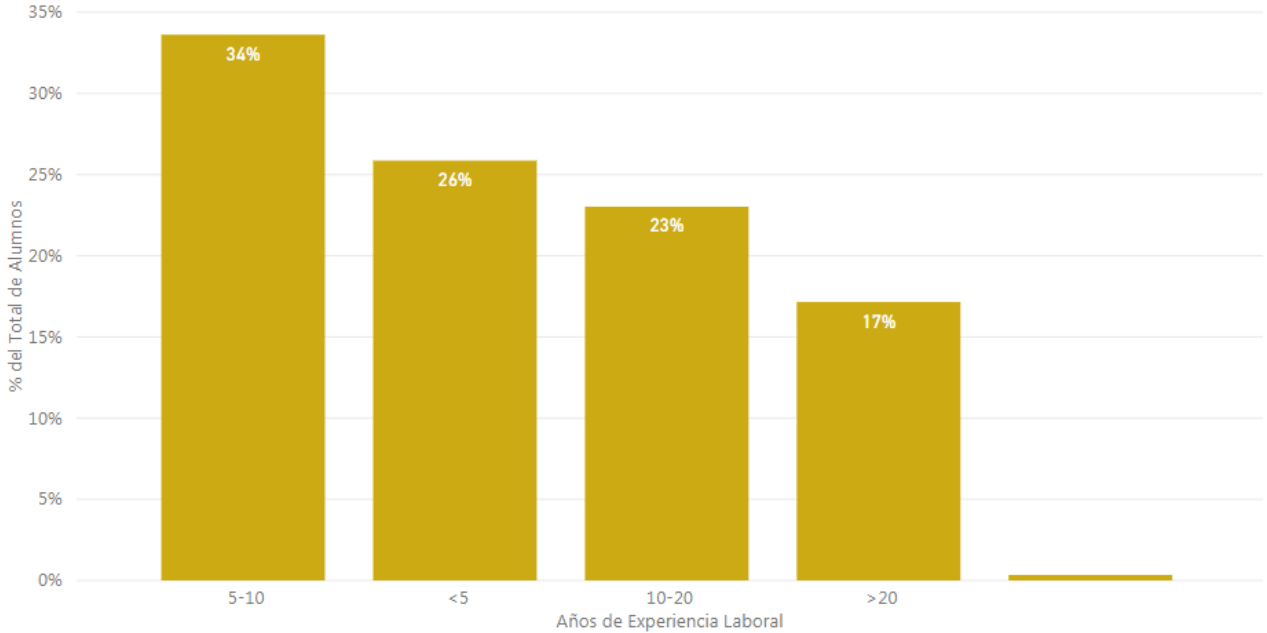
- El 50% de los alumnos provienen de ciudades de muy alto nivel de desarrollo (>=90%).
- Esto está estrechamente relacionado a la alta concentración de alumnos en ciertas ciudades.

Alumnos por Experiencia Relevante



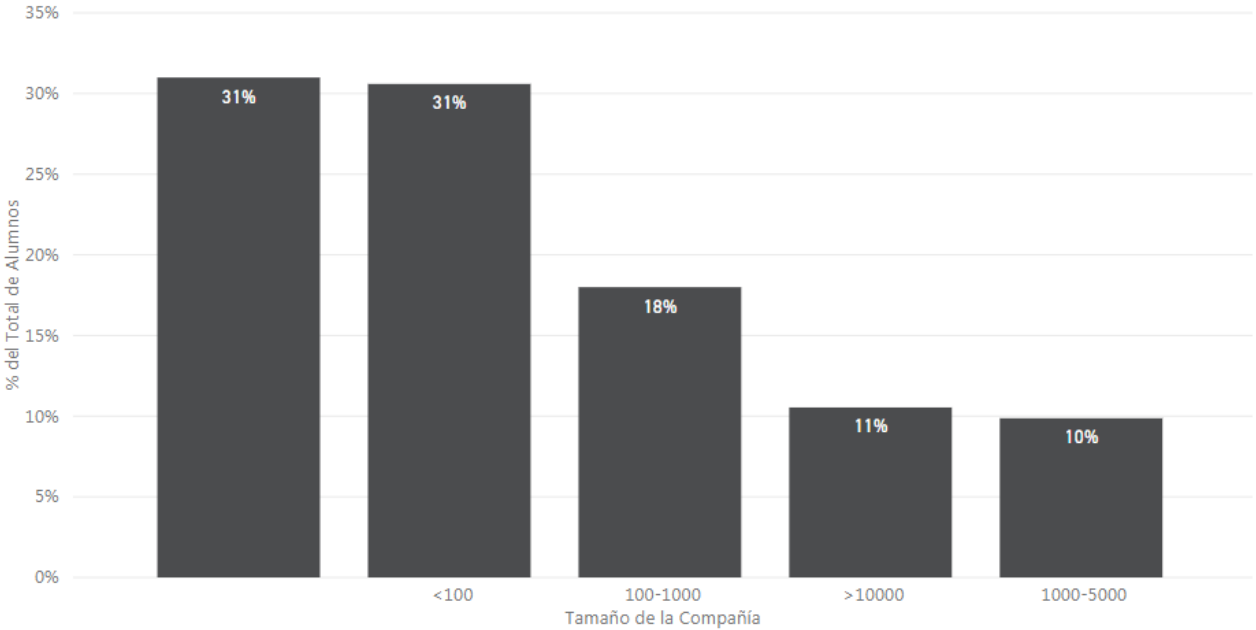
- El 72% de los alumnos tienen experiencia laboral relevante.
- La proporción de alumnos sin experiencia es mayor en aquellas ciudades con menor índice de desarrollo, entre los alumnos que trabajan en el sector público, con estudios primarios y secundarios y que nunca cambiaron de empleo.

Alumnos por Años de Experiencia Laboral



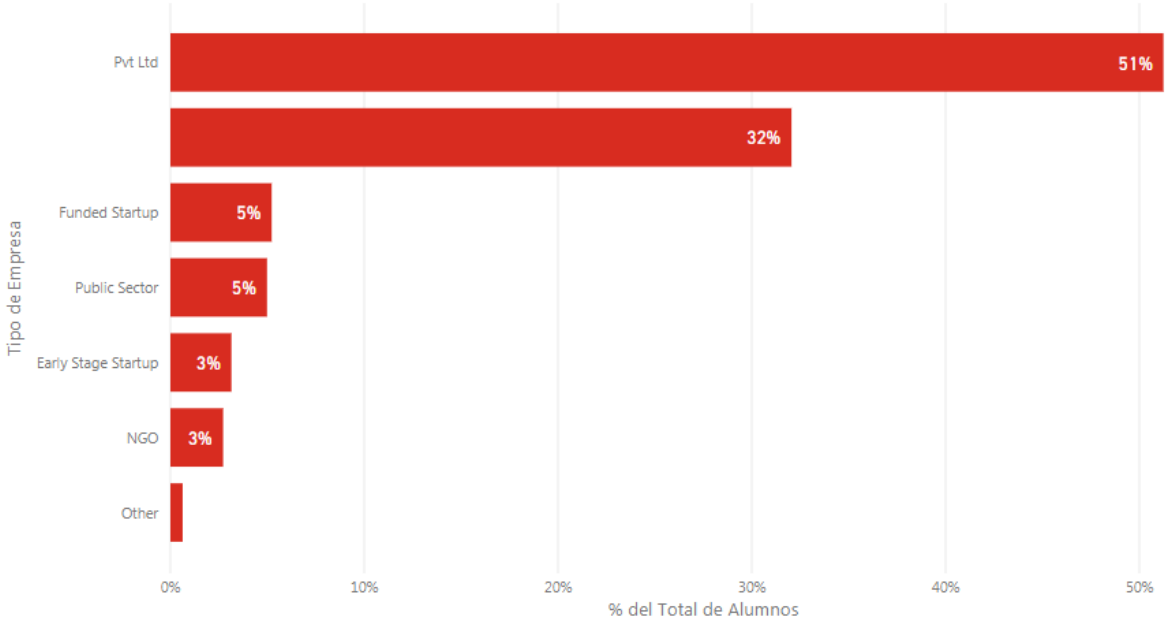
- El 60% de los alumnos tienen diez años o menos de experiencia laboral.
- Los de mayor experiencia (>20 años) tienden a proceder de ciudades de mayor índice de desarrollo.

Alumnos por Tamaño de la Empresa



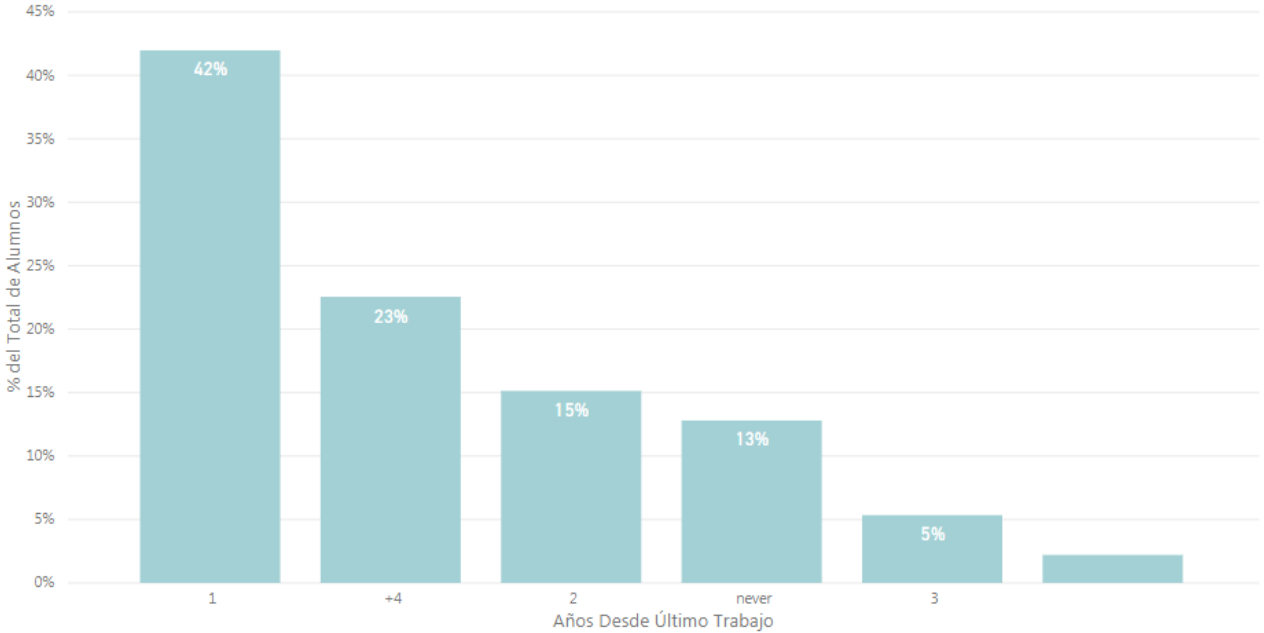
- El 31% de los registros está en blanco.
- El 57% de los alumnos restantes proviene de empresas medianas-chicas (< 1000 empleados).

Alumnos por Tipo de Empresa



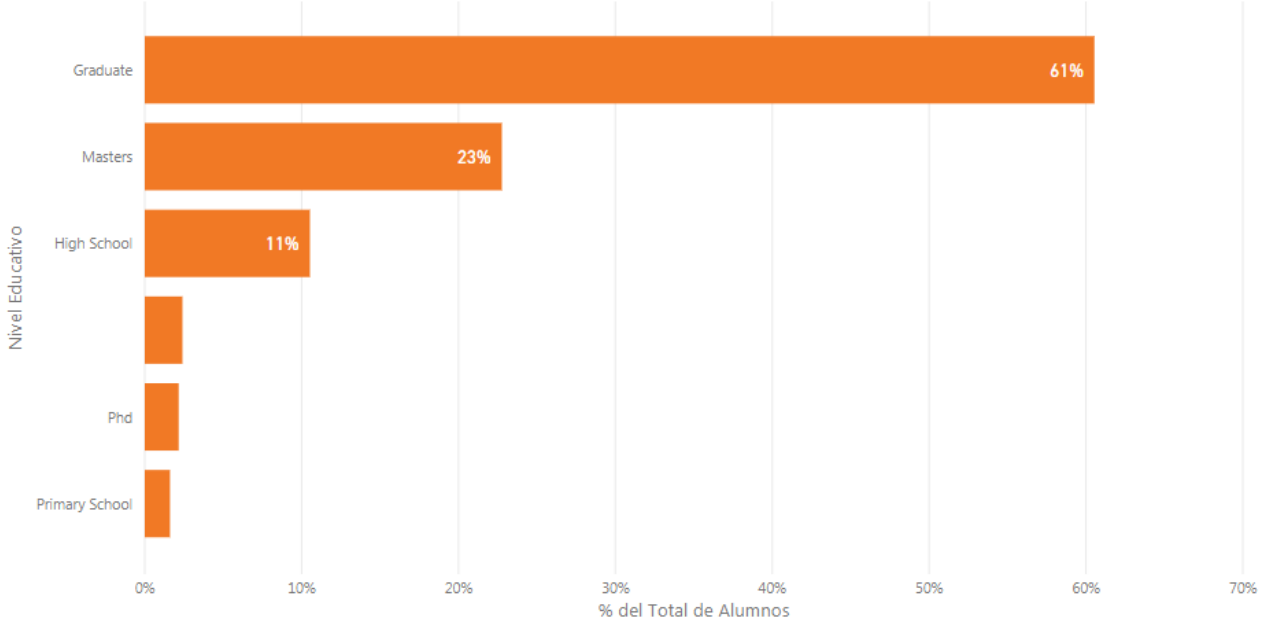
- El 32% de los datos están en blanco.
- El 75% de los alumnos restantes proviene de empresas privadas.

Alumnos por Años desde Último Trabajo



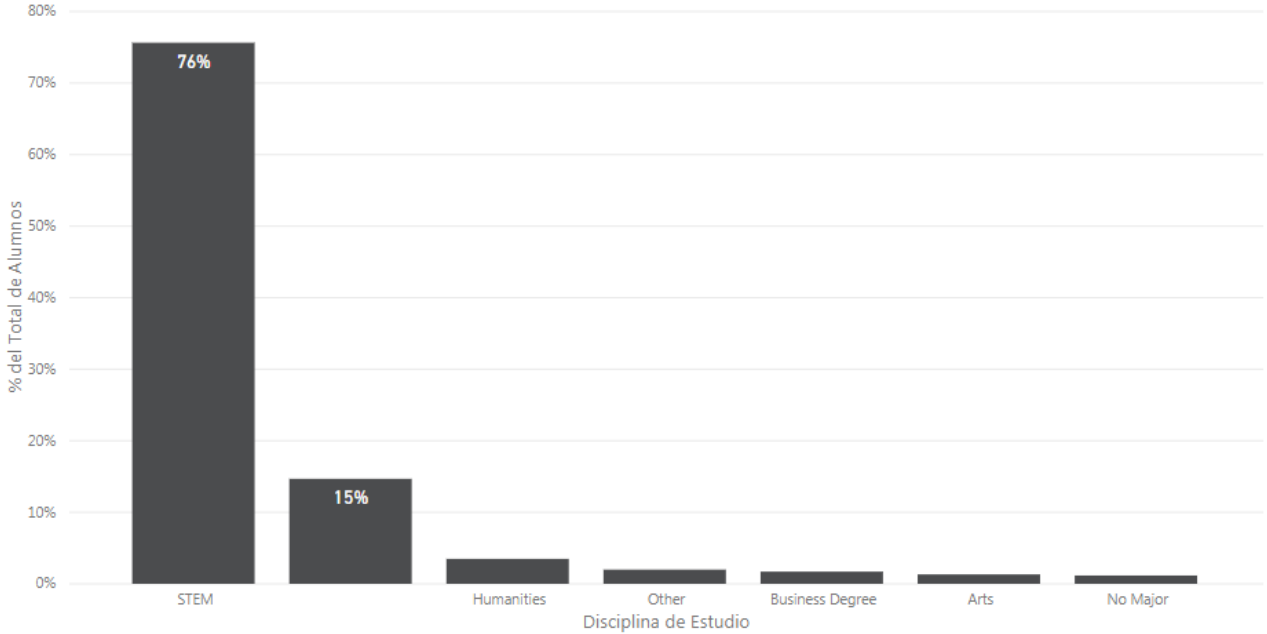
- El 42% de los alumnos han cambiado de trabajo en el último año.

Alumnos por Nivel Educativo



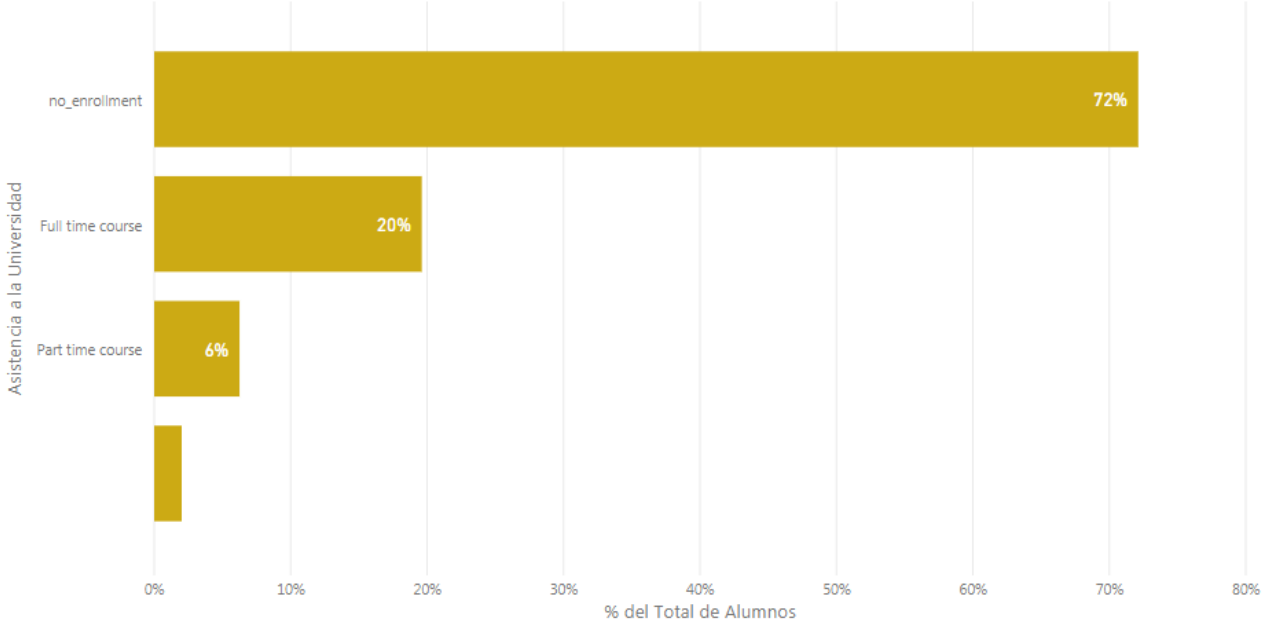
- El 84% de los alumnos tienen un título de grado o maestría.
- Los alumnos de menor nivel educativo tienden a un menor nivel de experiencia relevante.

Alumnos por Disciplina de Estudio



- El 76% de los alumnos tienen formación en Ciencias, Técnología, Ingeniería y Matemáticas.
- El 15% de los datos están en blanco.

Alumnos por Asistencia a la Universidad



- El 72% de los alumnos no se encuentra inscripto en la Universidad actualmente.
- Aquellos que no están inscriptos tienden a ser los alumnos de mayor antigüedad laboral, con mayor tiempo desde su último cambio de trabajo y con experiencia relevante.

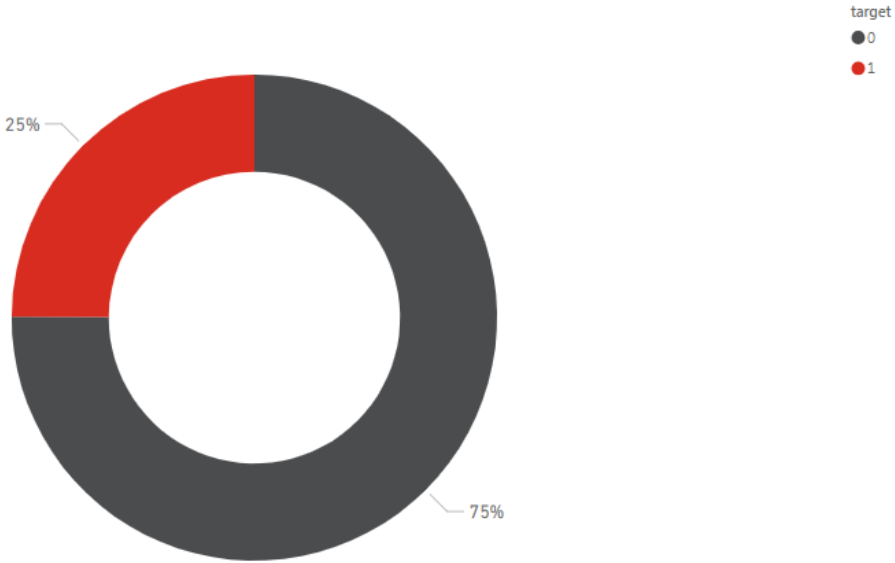
En resumen, el estudiante promedio:

- Es hombre.
- Proviene de una ciudad de alto grado de desarrollo.
- Tiene experiencia relacionada al Big Data / Data Science.
- Tiene una formación de grado o superior en Ciencias, Tecnología, Ingeniería o Matemáticas.
- Actualmente no está asistiendo a la Universidad
- Trabaja en una empresa privada.
- Tiene menos de diez años de experiencia laboral, y cambió de trabajo hace relativamente poco tiempo.
- Tomó un curso de entre 23 y 88 horas de duración.

¿Cómo se relacionan las distintas variables con respecto a la búsqueda laboral?

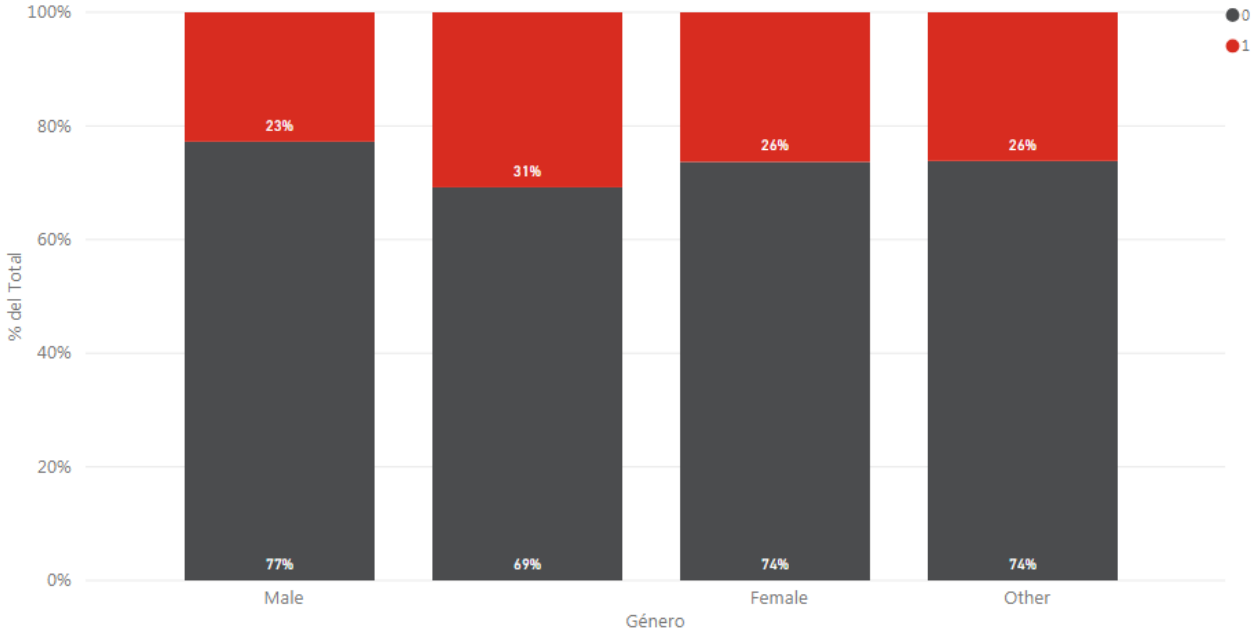
Realizaremos un análisis vinculado al objetivo del trabajo, relacionando nuestra **variable target** con las distintas variables disponibles, tratando de encontrar los primeros *insights* en esta exploración inicial de los datos.

Variable Target



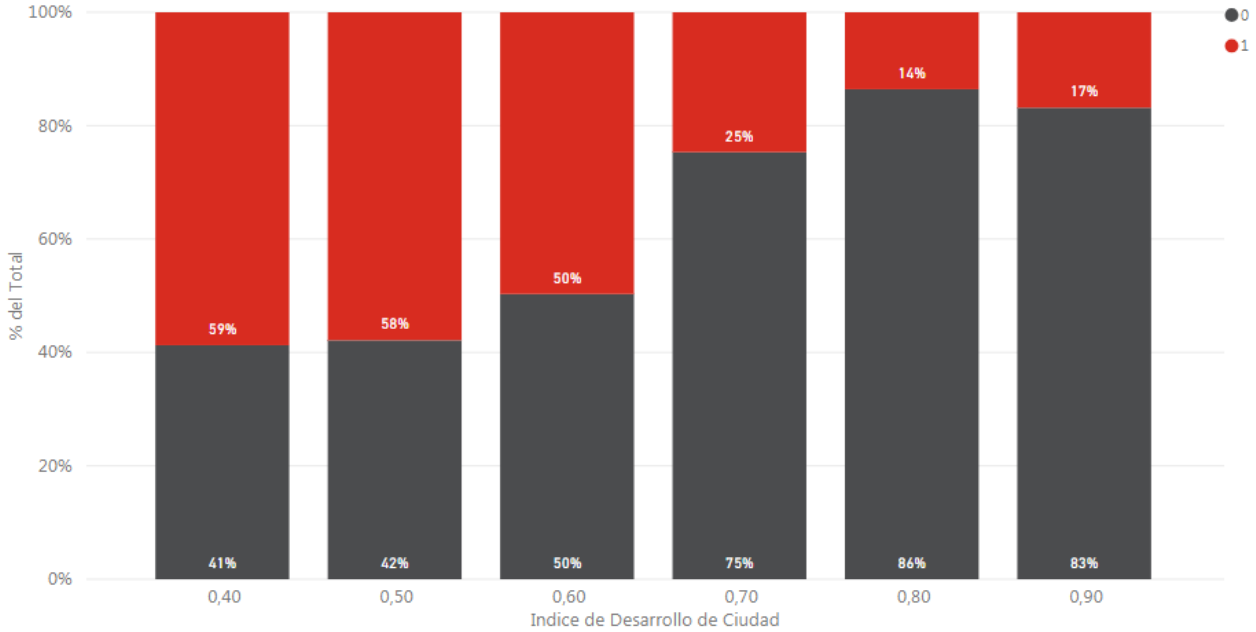
- El 25% de los alumnos busca cambiar de trabajo.

Target por Género



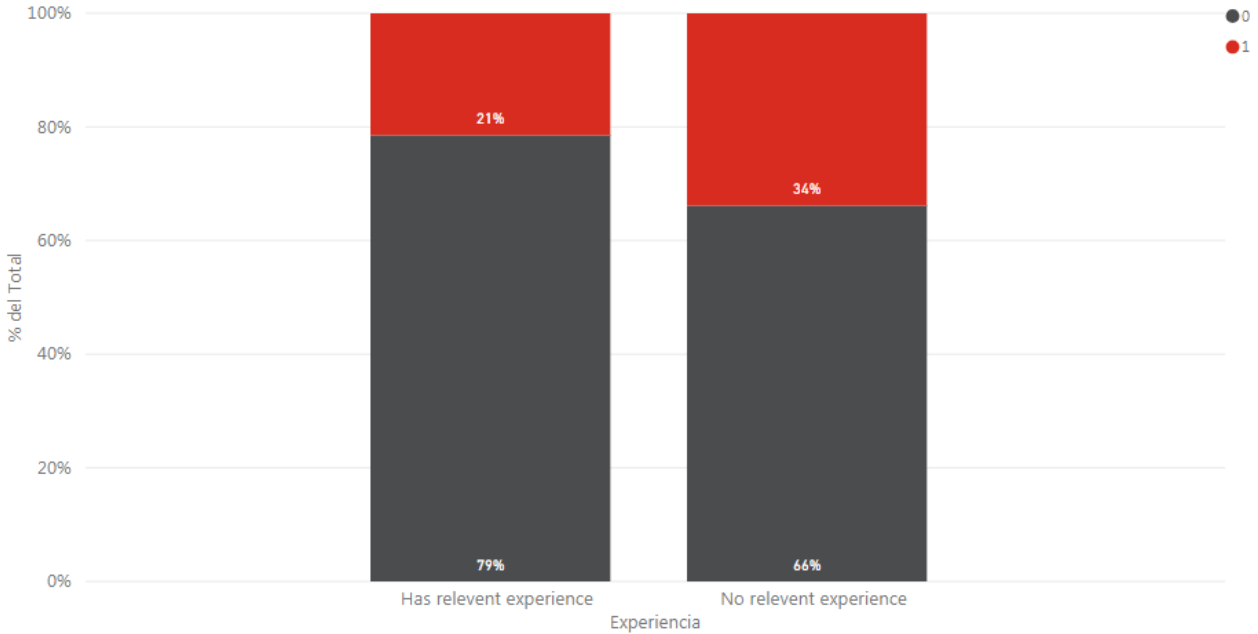
- En una primera instancia no se observan diferencias sustanciales en función del género. No obstante, dado el faltante del 24% de los datos, esta relación podría variar.

Target por Índice de Desarrollo de su Ciudad



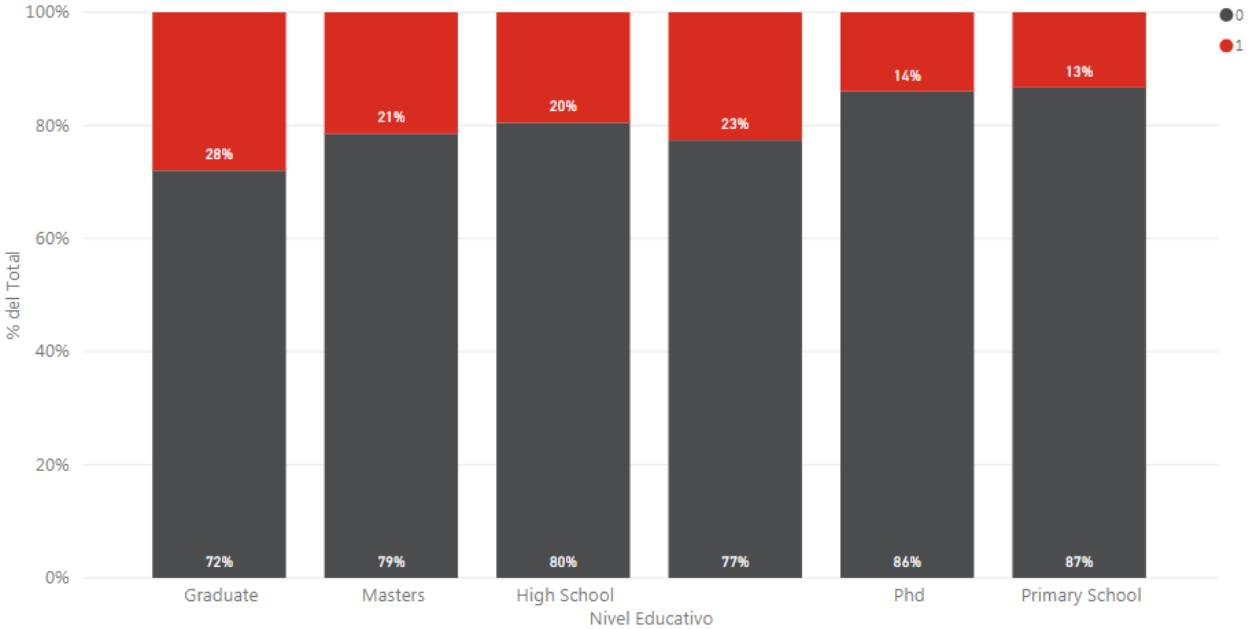
- Aquellos alumnos procedentes de ciudades de menor grado de desarrollo son más propensos a buscar un nuevo empleo.

Target por Experiencia Relevante



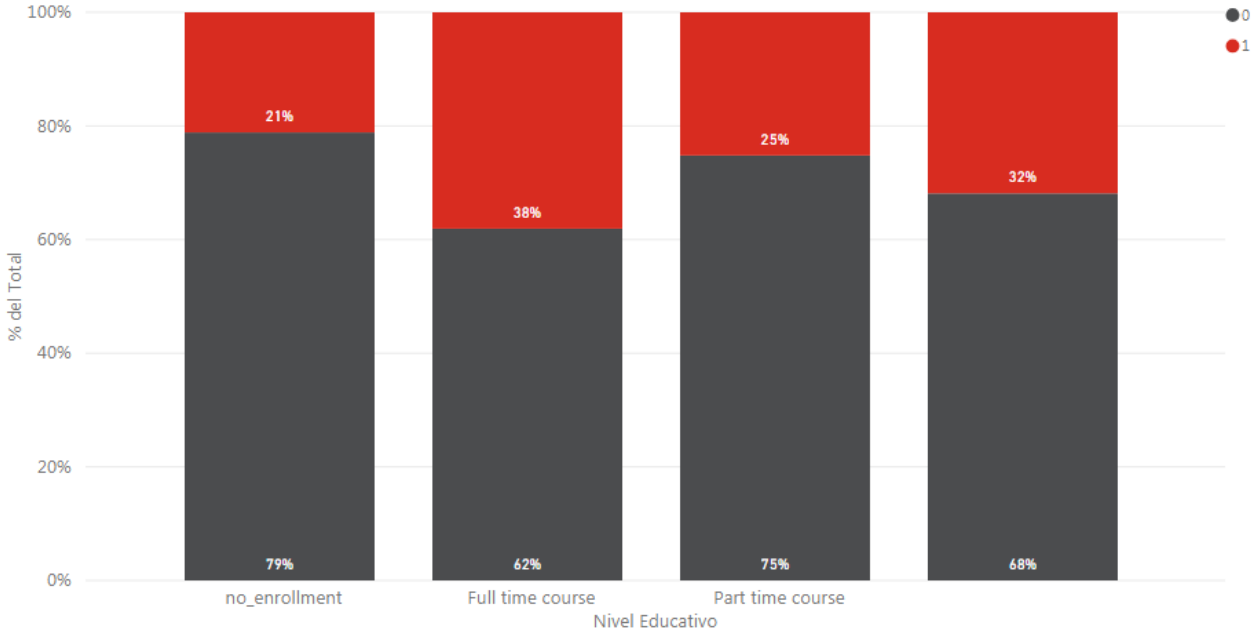
- Se observa una mayor tendencia a buscar empleo entre aquellos alumnos que no poseen experiencia previa en el tema.

Target por Nivel Educativo



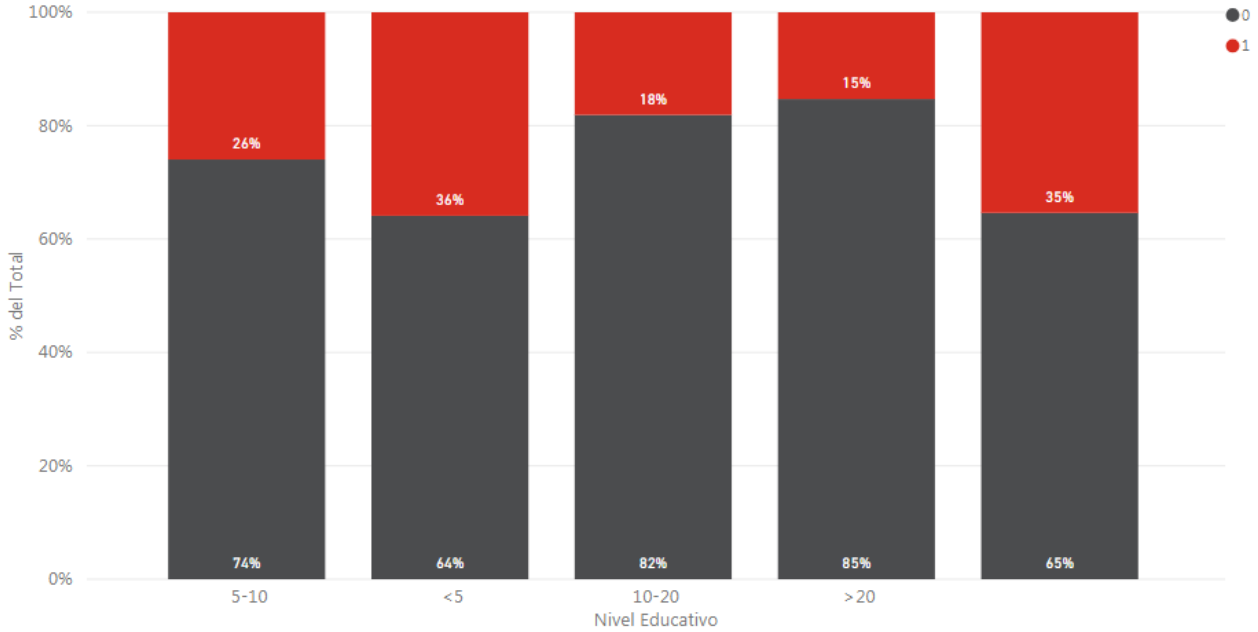
- Los alumnos de menor nivel educativo muestran una tendencia ligeramente mayor al cambio laboral. El faltante de datos (2,4%) podría alterar esta relación, pero es poco probable.

Target por Asistencia a la Universidad



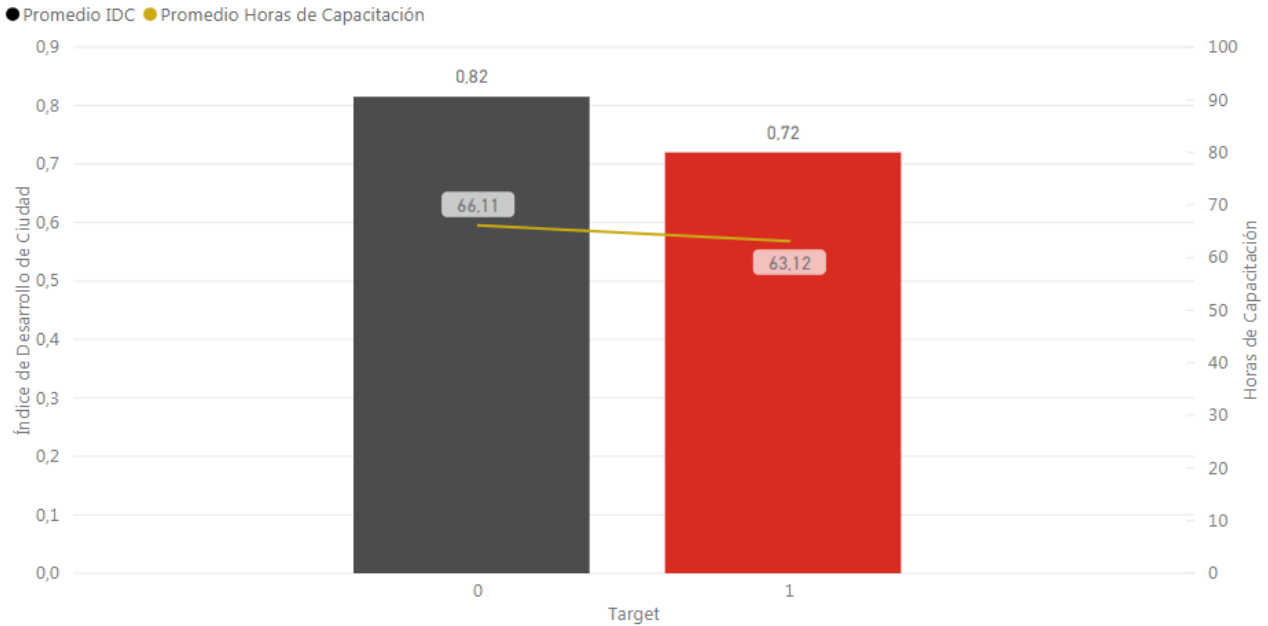
- Aquellos alumnos que actualmente se encuentran realizando un curso a tiempo completo son más propensos a buscar un empleo. Nuevamente, el faltante de datos (2,01%) podría alterar un poco estas relaciones.

Target por Años de Experiencia Laboral



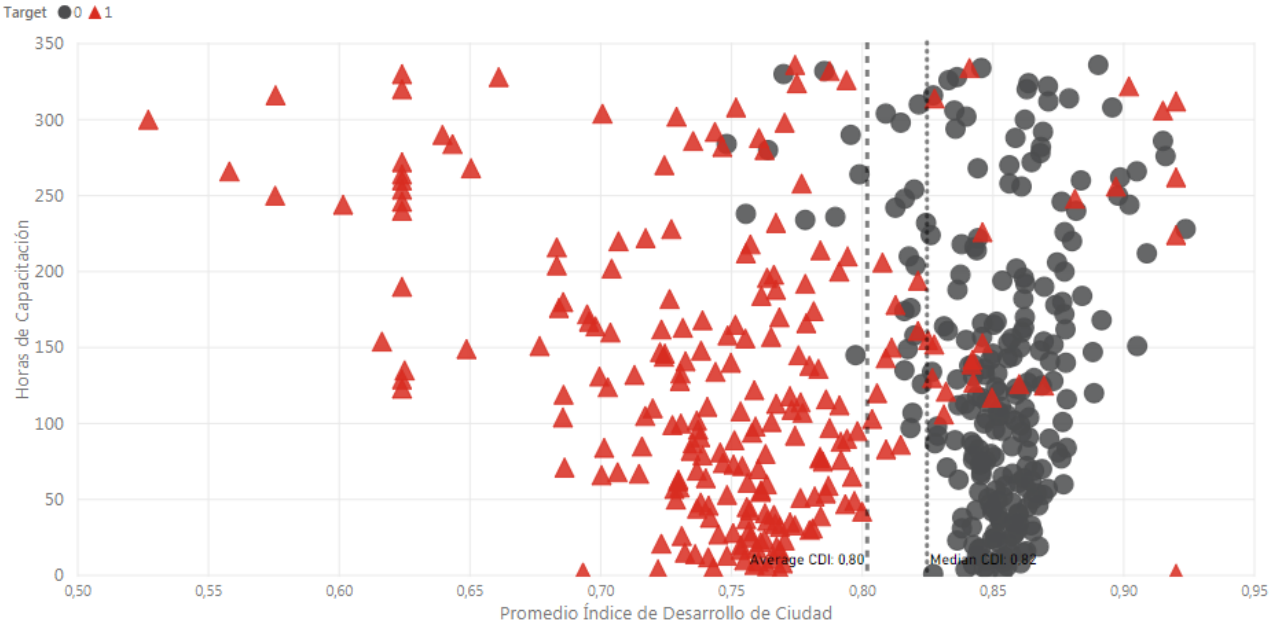
- Los alumnos con menor cantidad de años de experiencia laboral están más dispuestos a cambiar de trabajo.
- Esta relación puede deberse a cuestiones generacionales o estructurales, no necesariamente a la situación propia bajo análisis.

Target por Horas de Capacitación e Indice de Desarrollo



- Aquellas personas dispuestas a cambiar de trabajo, en promedio, provienen de ciudades de menor índice de desarrollo.
- La diferencia en horas de capacitación no es significativa.

Target por Horas e Indice de Desarrollo Promedio



- A un mismo nivel de horas de capacitación, la relación entre búsqueda de empleo e índice de desarrollo de ciudad se mantiene.
- La diferencia entre ambos grupos se vuelve un poco más difusa a mayores niveles de horas de capacitación (RIC: 23-88 horas).

Conclusiones

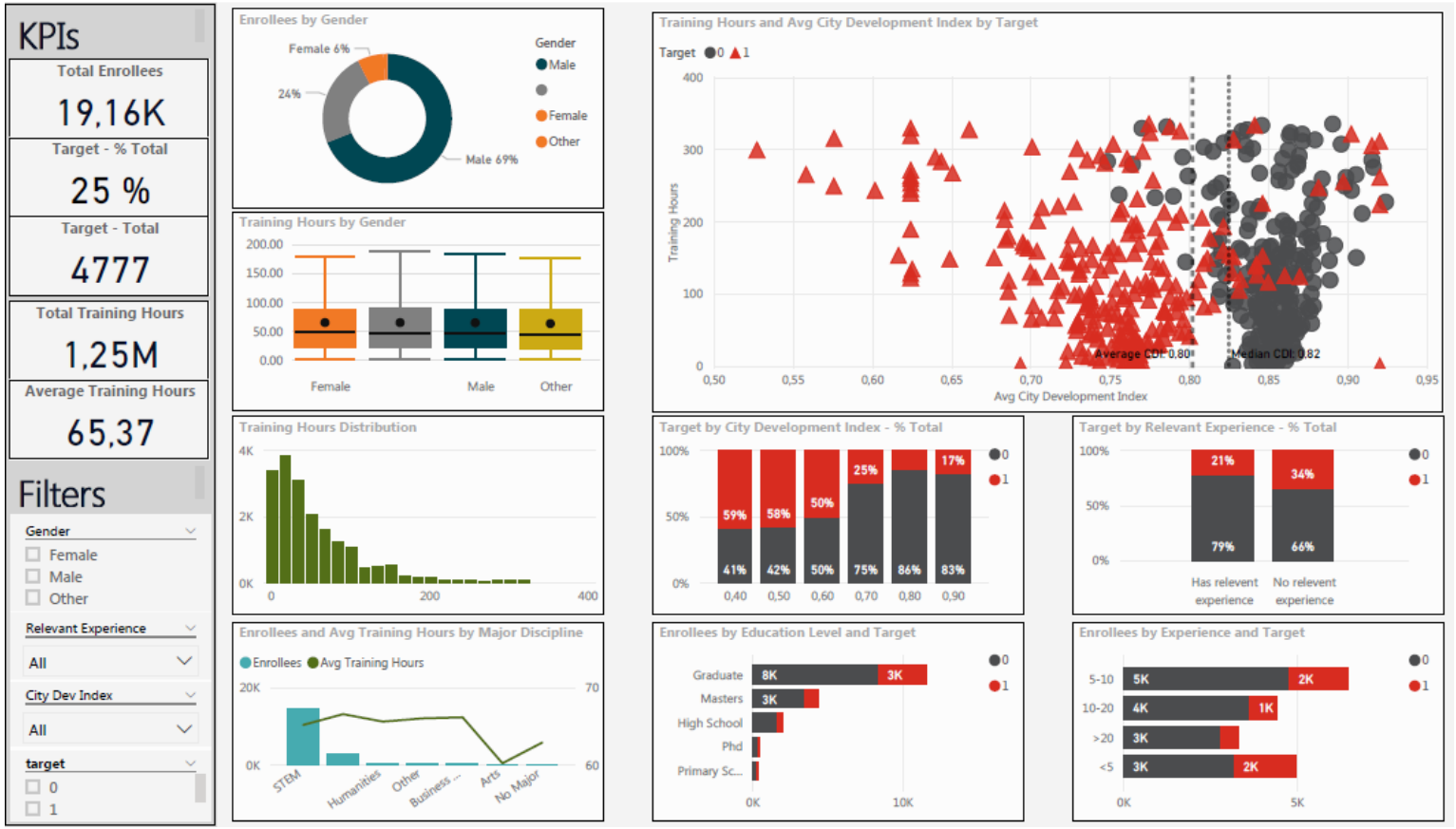
El estudiante que busca trabajo, en promedio:

- Proviene de una ciudad de nivel de desarrollo bajo-medio.
- No tiene experiencia previa relacionada al Big Data / Data Science.
- Posee título de grado.
- Está realizando un curso a tiempo completo en la Universidad.
- Tiene menos de diez años de experiencia laboral.
- No se observan diferencias sustanciales en relación al *Género*, *Disciplina de Estudio*, *Tamaño de la Compañía*.

Cabe destacar que estas conclusiones se derivan de un análisis exploratorio básico, a partir de la visualización y algunas medidas estadísticas. Los algoritmos de **machine learning** nos podrán brindar información mucho más concreta sobre la **importancia** de las distintas variables y cual es su incidencia en el target, pudiendo **predecir** cuál es la **probabilidad** de que una persona busque trabajo o no.

Dashboard: seguimiento contínuo de los alumnos y de posibles candidatos

A continuación proponemos un modelo de *dashboard* para el monitoreo contínuo de los alumnos, horas de capacitación, y la búsqueda de potenciales candidatos.


KPIs
Total Enrollees
19,16K
Target - % Total
25 %
Target - Total
4777
Total Training Hours
1,25M
Average Training Hours
65,37
Filters
Gender
Female
Male
Other
Relevant Experience
All
City Dev Index
All
target
0
1
Enrollees by Gender
Female 6%
24%
Male 69%
Gender
Male
Female
Other
Training Hours by Gender
200.00
150.00
100.00
50.00
0.00
Female
Male
Other
Training Hours Distribution
4K
2K
0K
0
200
400
Enrollees and Avg Training Hours by Major Discipline
Enrollees
Avg Training Hours
20K
0K
70
60
STEM
Humanities
Other
Business ...
Arts
No Major
Training Hours and Avg City Development Index by Target
Target 0 1
400
300
200
100
0
Training Hours
0,50
0,55
0,60
0,65
0,70
0,75
0,80
0,85
0,90
0,95
Avg City Development Index
Average CDI: 0.80
Median CDI: 0.82
Target by City Development Index - % Total
100%
50%
0%
59%
58%
50%
25%
17%
41%
42%
50%
75%
86%
83%
0,40
0,50
0,60
0,70
0,80
0,90
Target by Relevant Experience - % Total
100%
50%
0%
21%
34%
79%
66%
Has relevent experience
No relevent experience
Enrollees by Education Level and Target
Graduate
Masters
High School
Phd
Primary Sc...
8K
3K
3K
0K
10K
Enrollees by Experience and Target
5-10
10-20
>20
<5
5K
2K
4K
1K
3K
3K
2K
0K
5K

Información General

Características de los Alumnos en Relación a la Búsqueda Laboral

Indicadores

KPIs

Total Enrollees	Target - % Total	Target - Total	Total Training Hours	Average Training Hours
19,16K	25 %	4777	1,25M	65,37

Filtros

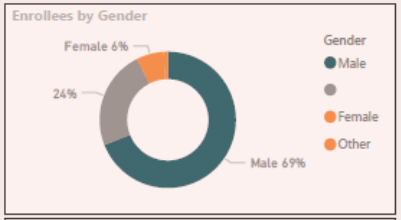
Filters

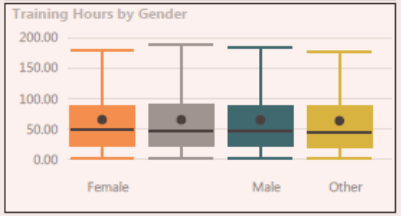
Gender
- Female
- Male
- Other

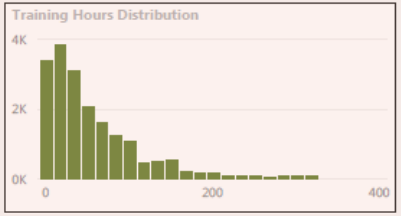
Relevant Experience: All

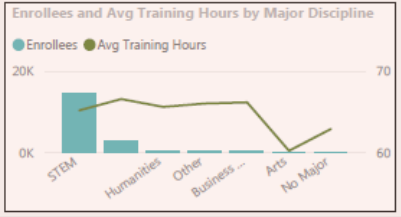
City Dev Index: All

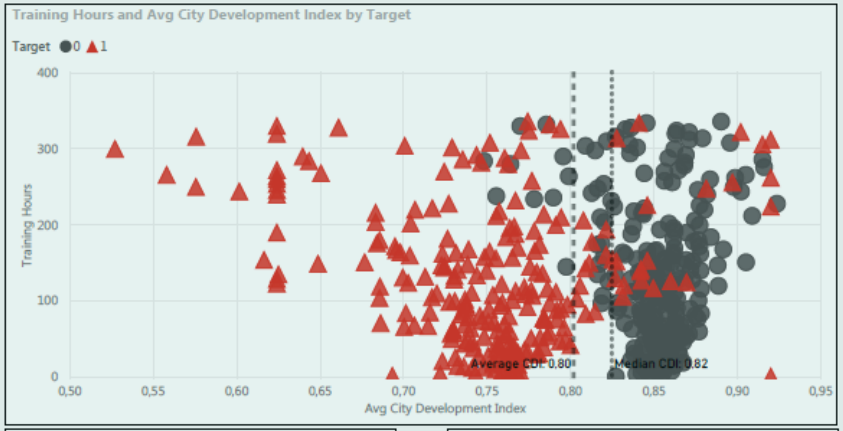
target
- 0
- 1

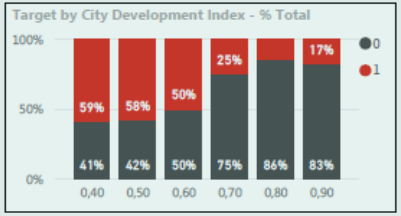


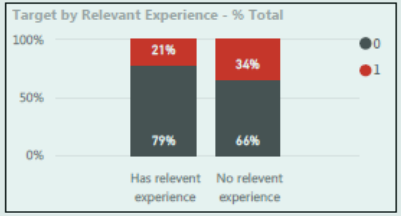


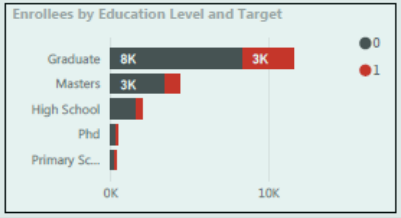


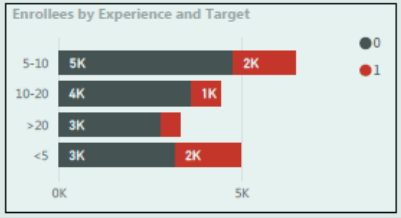












Próximos pasos

- Tratamiento de campos faltantes.
- Implementación de modelos Predictivos: Entrenamiento, Optimización y Validación.