

# I Entrega

Noelia Rojas Ramírez

## Propuesta 1 Agente

### Fuentes de datos analizadas

Datos obtenidos de la empresa en que trabajo, son dos dataset.

**Dataset1:** a través de un agente instalado en las computadoras de los colaboradores se puede obtener el registro de las aplicaciones y navegadores que el usuario Windows visita durante el día o periodos en que la computadora se encuentre encendida. En este set de datos se encuentran los usuarios de la empresa principal y de una subsidiaria, dado que se encuentran en el mismo dominio.

**Dataset2:** planilla de la empresa, se encuentra información como el nombre del empleado, usuario Windows, correo, cédula, dependencia. Esta hoja es importante para seleccionar solo los usuarios de la empresa de interés. Nota, en este trabajo se van a utilizar nombres genéricos para velar por la privacidad de los usuarios, dependencia y la compañía para la cual trabajo.

**Datsert3:** dependencias de la empresa principal, este set es necesario para ponerle nombre genérico a los departamentos que vienen en la planilla y de esta forma mantener la confidencialidad. Son más de 80 dependencias, aunque solo se van a trabajar con algunas.

### Descripción detallada de los datos

#### Dataset1

- Usuario: variable categórica, usuario Windows con los que la persona se loguea en su computadora, hay 185 en total.
- Aplicación: variable categórica, se refiere a la aplicación o navegador con la que interactúa el colaborador durante el uso de la aplicación, hay 263 aplicaciones en este set, al ser tantas se considera que la manera más efectiva de presentarlas visualmente es a través de un treemap, ver figura 1.

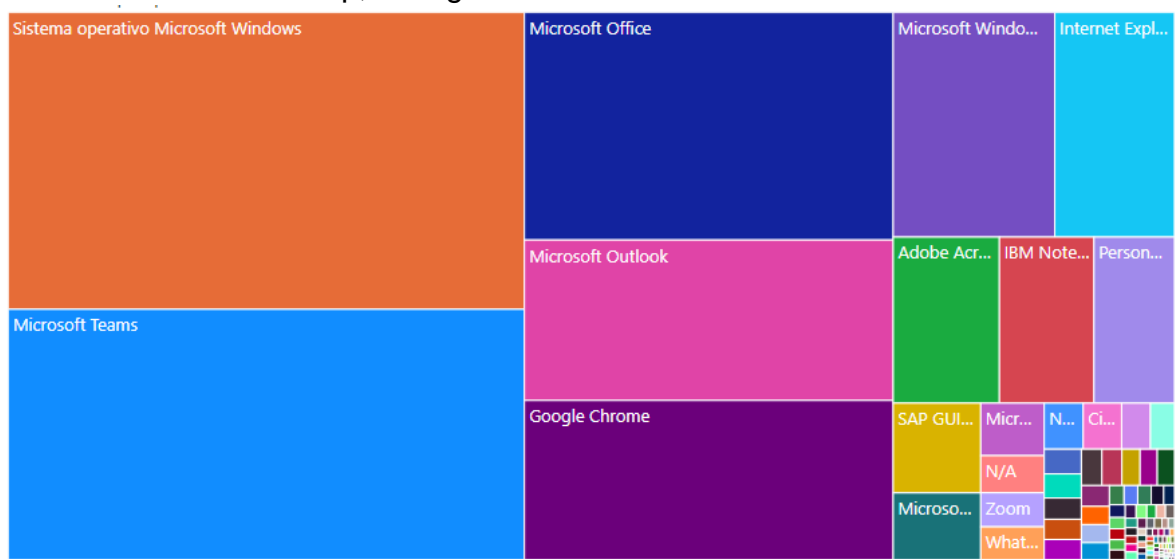


Figura 1. Duración en horas por aplicación

- Equipo: nombre del equipo/computadora que cuenta con el agente instalado por medio del cual el colaborador se loguea para hacer uso de esta. Esta variable no se considera de interés en el estudio.
- Detalle: variable categórica, es el nombre de la ventana en la que se encuentra el usuario, por ejemplo, si se encuentra en la aplicación de Google y está interactuando con la pág de Whatsapp Web, esta es el detalle.
- Detecta: indica si la pagina a la que ingresó el usuario es indebida, se obtiene de un análisis previo.
- Hora: variable numérica, contiene la fecha y hora exacta en que el usuario comienza a interactuar con la computadora, se muestra un gráfico de calor que muestra los usuarios activos según la hora y el día de la semana, el azul más oscuro es donde se concentra la mayor cantidad de usuarios y el azul claro la menor cantidad. Ver figura 2.



Figura 2. Conexión de usuarios por día de la semana.

- Hora fin: es igual a la anterior, pero representa la hora en la que el usuario deja de interactuar con la aplicación-ventana. Esta variable no se considera de interés, dado que se usará la hora de inicio y la duración en cada aplicación.
- Duración segundos: variable numérica, indica la cantidad de tiempo en segundos que tardó la persona en cada aplicación/ventana. A continuación, se presenta un histograma de la variable. Esta variable se puede pasar a minutos u horas, según convenga para una mayor comprensión.

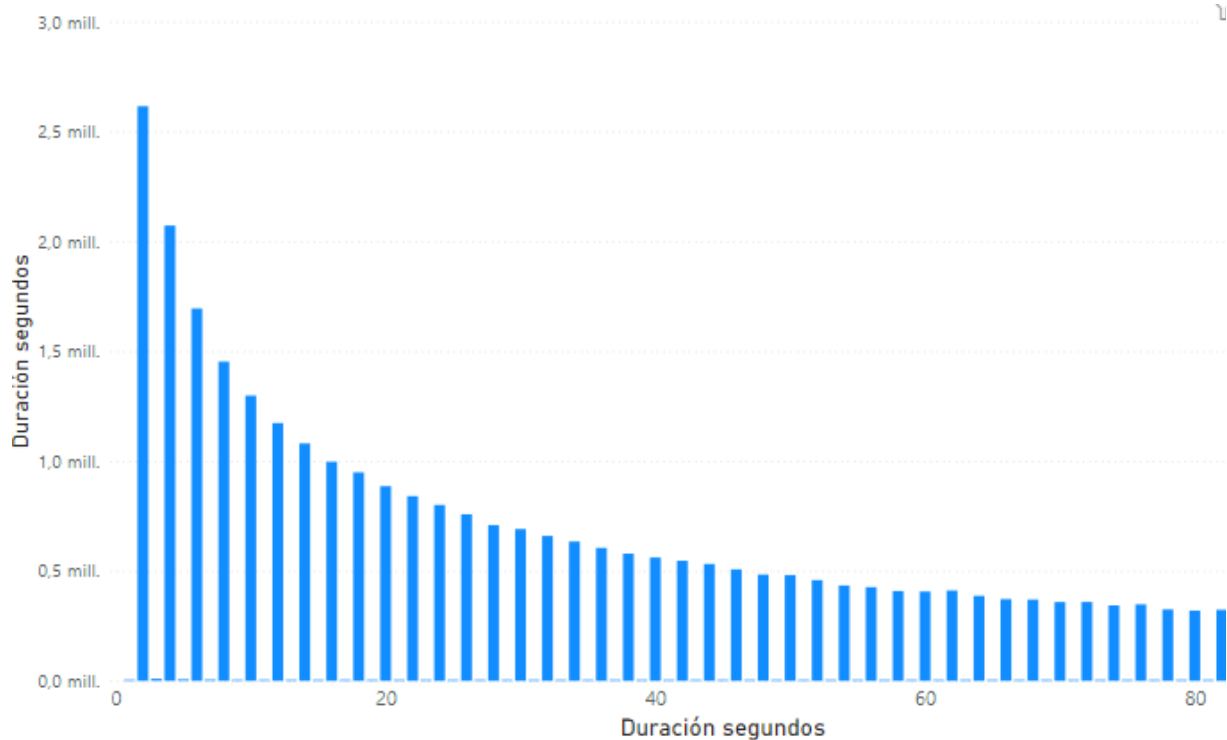


Figura 3. Histograma de duración en segundos.

- Estado: variable categórica, indica si el equipo está activo o inactivo, por ejemplo, si el colaborador se va de su sitio de trabajo y se le suspenda la computadora o se le ponga el protector de pantalla el estado es inactivo. Ver figura 4.

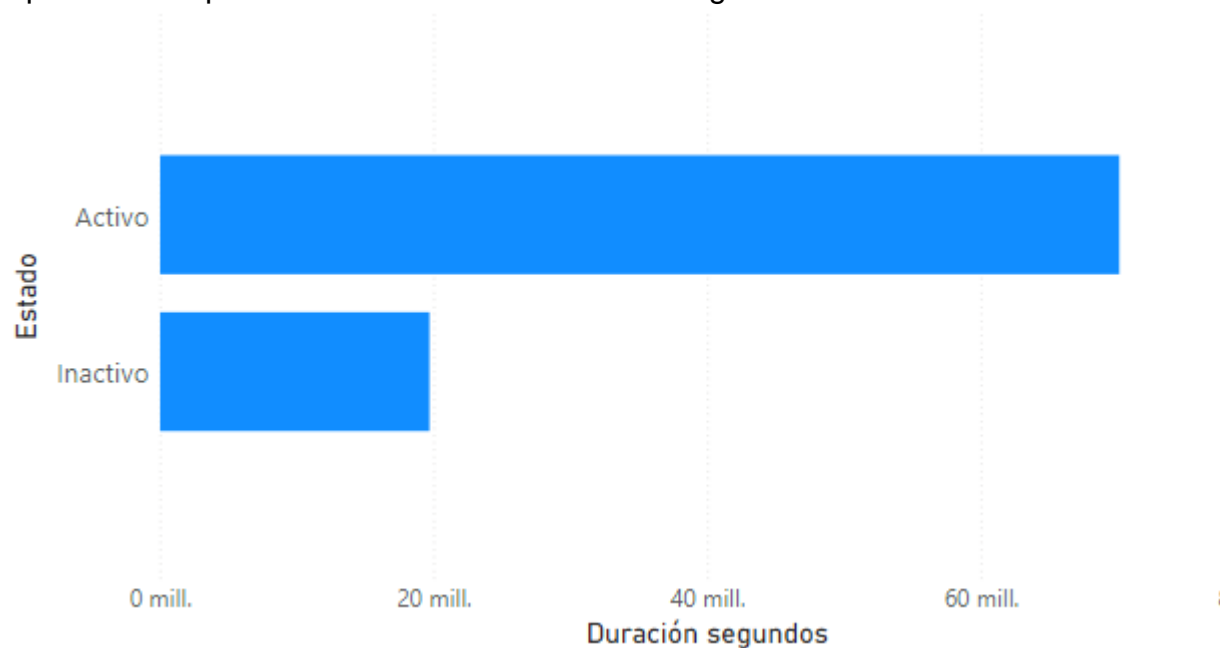


Figura 4. Duración en segundo por estado

## Dataset2

La planilla de la empresa está conformada por las siguientes variables

- Nombre: indica el nombre de la persona. Para este trabajo no se considera relevante.
- Numero de empleado: indica el número de emplado. Para este trabajo no se considera relevante.
- Usuario Windows: variable categórica, usuario Windows con los que la persona se loguea en su computadora. Con esta variable se realizará el join.
- Email: correo electrónico del colaborador, no se considera relevante para el estudio.
- Departamento: es la dependencia a la que pertenece cada colaborador. Ver figura 5.

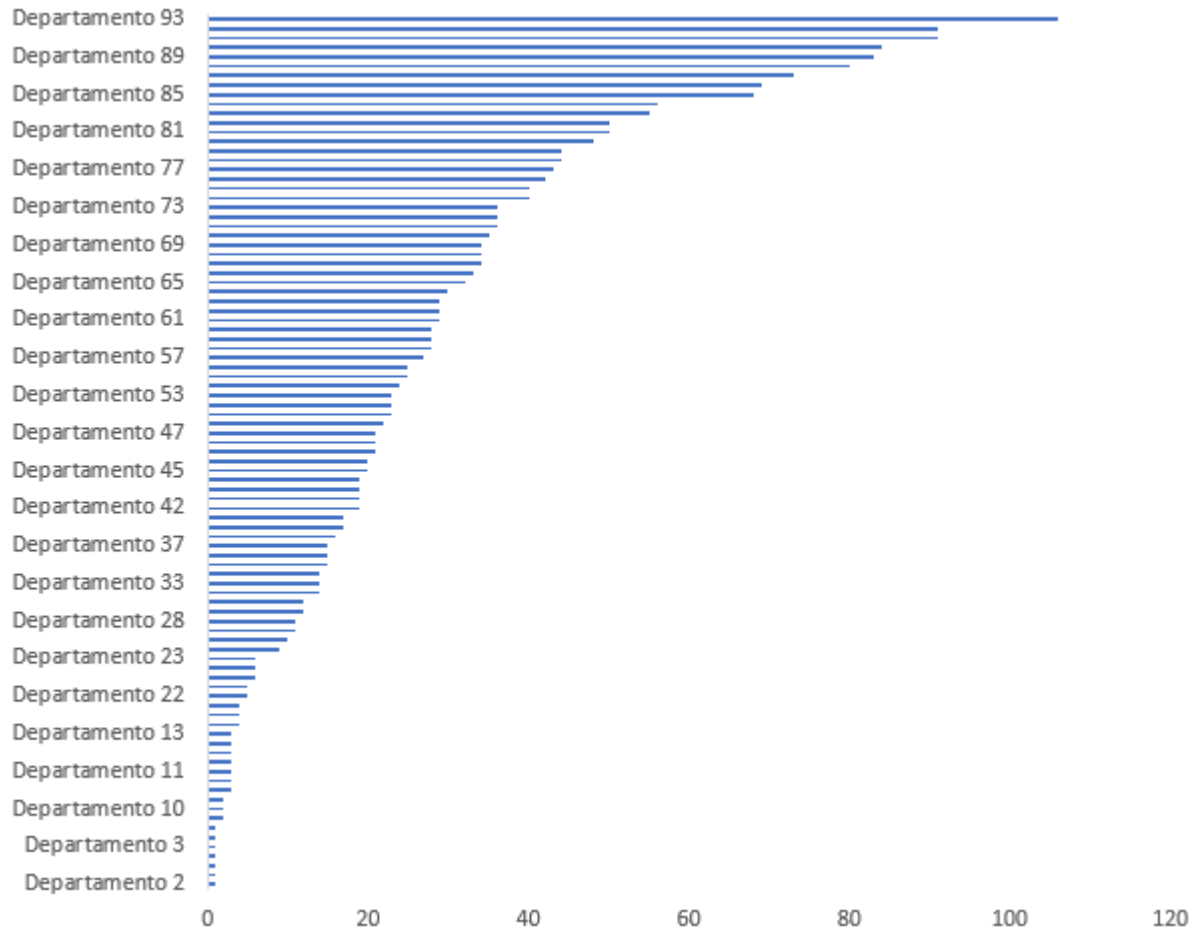


Figura 5. Cantidad de colaboradores según la dependencia a la que pertenecen

- Cargo: variable categórica, puesto de trabajo en el que se encuentra cada colaborador de la empresa, ver figura 6.

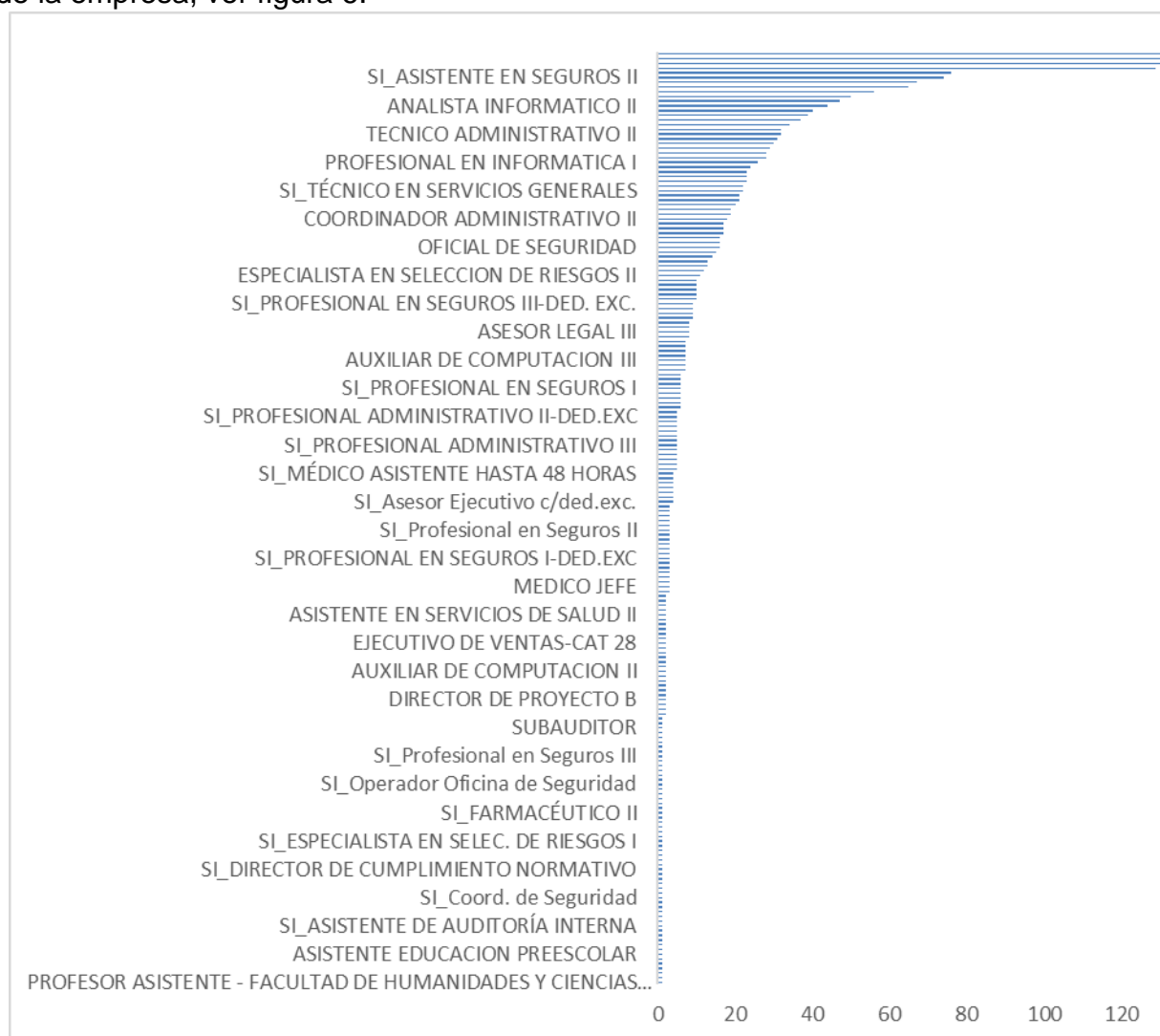


Figura 6. Cantidad de colaboradores por cargo

- Correo del jefe: esta variable no es de interés para el estudio.
- Genero: variable categórica, indica si el colaborador es femenino, masculino u otro. Ver figura 7.

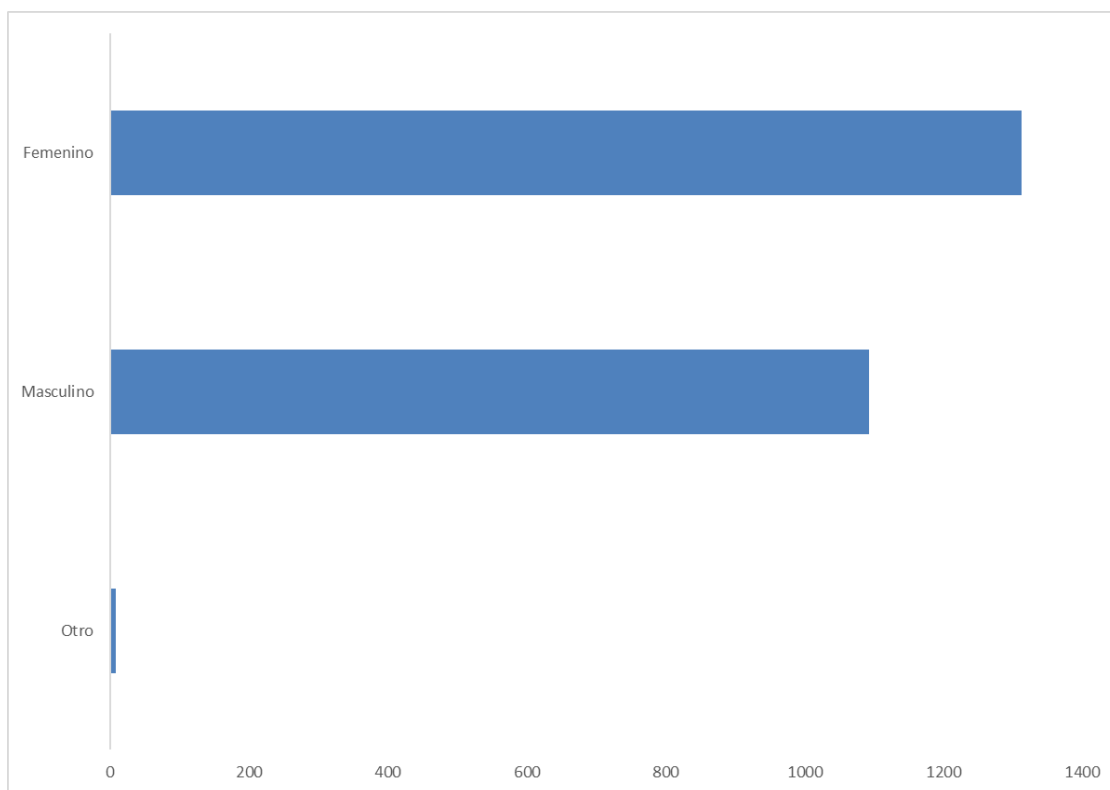


Figura 7. Cantidad de colaboradores según el género

- Fecha de ingreso: la fecha en que ingresó a trabajar en la empresa, de acá se puede calcular los años que lleva laborando.
- Rol: variable categórica, indica si la persona es colaborador, encargado o jefatura. Ver figura 8.

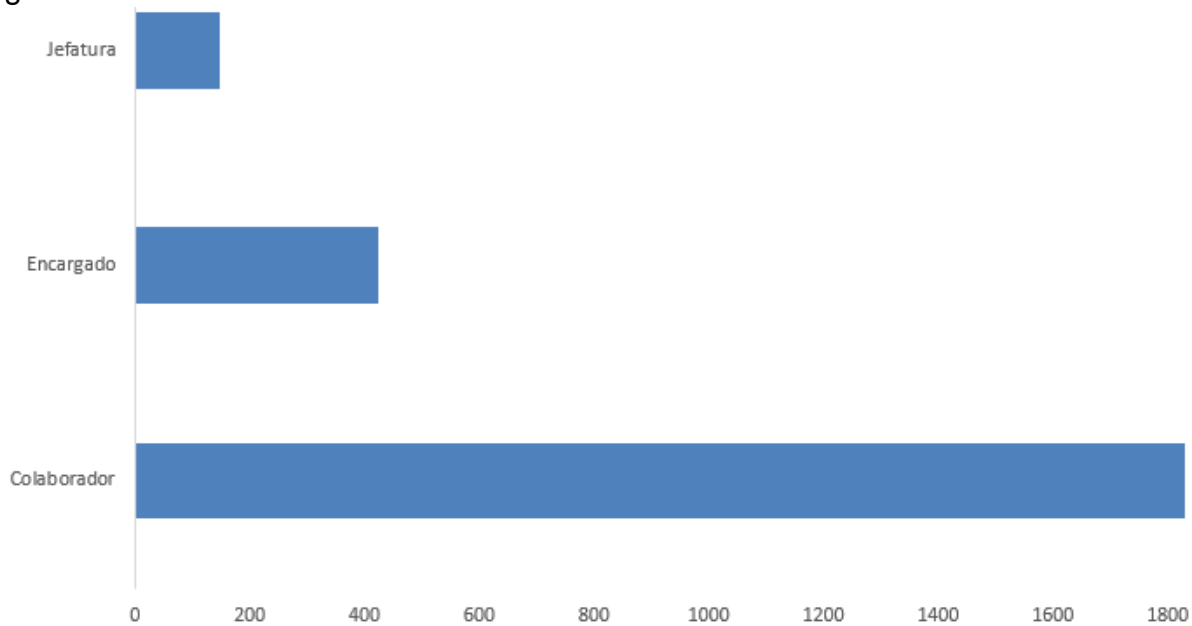


Figura 8. Cantidad de personas por rol.

Objetivo predictivo

Predecir la probabilidad de que el usuario se encuentre en una pagina web indebida o regular para un mes. Se realiza de manera mensual porque es de interés de la empresa hacer la revisión de esta manera.

Nota: Los datos se deben de balancear, dado que se espera que hayan menos visitas a páginas indebidas que en las regulares. Esto se va a tomar en cuenta en el entrenamiento de los datos, durante la validación cruzada. Se consideran páginas indebidas aquellas que no son de uso laboral, por ejemplo, que el usuario pase tiempo en Pinterest, tik tok, entre otras.

Propuesta 2 Matriculas

Fuentes de datos analizadas

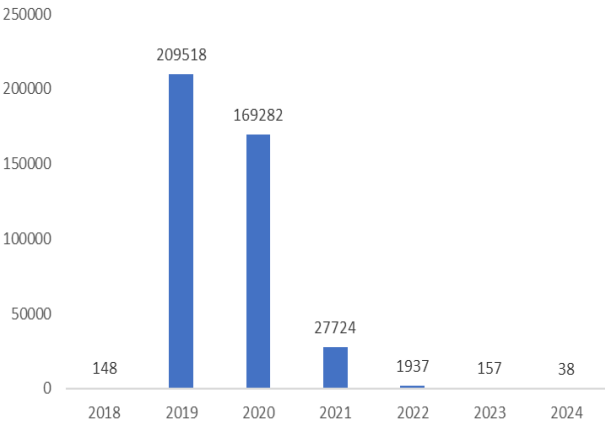
Aprobación de cursos de un instituto (**confidencial**)

**Dataset1:** matriculas de los diferentes programas realizadas en el instituto desde el 2018 al 2020. Cada persona se inscribe a un programa el cual se compone de uno o más cursos.

**Dataset2:** es el Índice de Desarrollo Humano, este se obtuvo del PNUD.

Descripción detallada de los datos

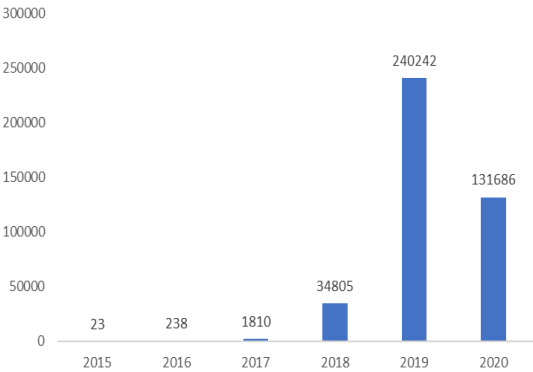
Dataset1

Variables	Descripción	Grafica	Nivel de medición
id_persona	Número de identificación en el dataset No es una variable de interes	NA	Escala
ANO_FIN	Año de finalización del programa		Escala

ANO\_INI

Año de inicio del programa

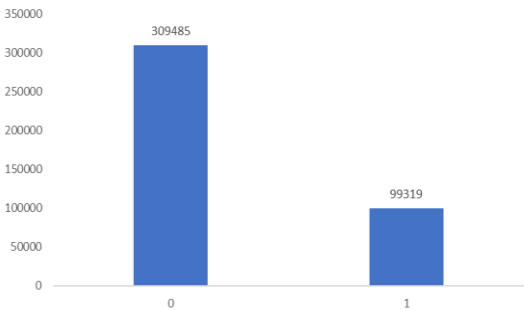
Nominal



becado

Si la persona recibió beca para matricular el curso, 0: no y 1 : sí

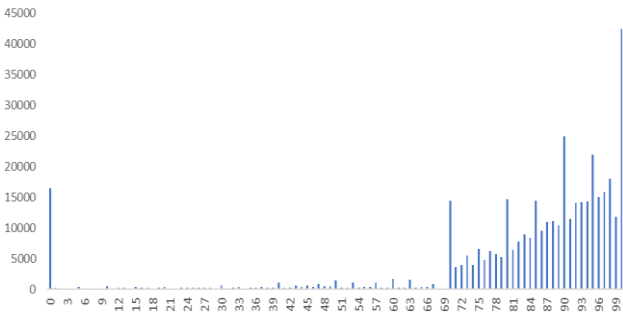
Nominal



CALIFICA

Calificación del estudiante

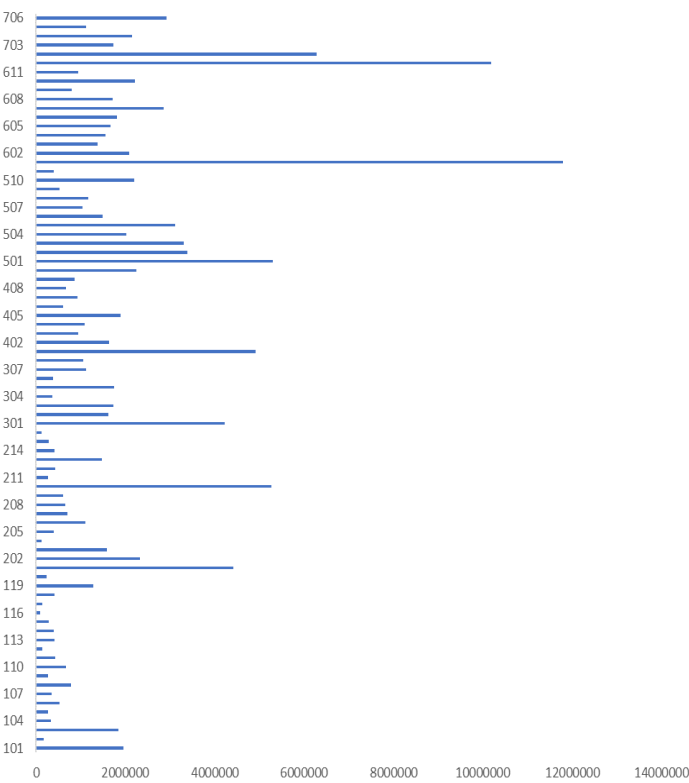
Escala





cantestu

Canton  
101San José,  
102Escazú,103Desampara  
dos,104Puriscal,105Tarraz  
ú,106Aserri,107Mora,108  
Goicoechea,109Santa  
Ana,110Alajuelita,111Vázq  
uez de  
Coronado,112Acosta,113T  
ibás,114Moravia,115Mont  
es de  
Oca,116Turubares,117Do  
ta,  
118Curridabat,119Pérez  
Zeledón,120León  
Cortes,201Alajuela,202San  
Ramón,203Grecia,  
204San  
Mateo,205Atenas,206Nar  
anjo,207Palmares,208Poá  
s,209Orotina,210San  
Carlos,211Zarcelero,212Valv  
erde Vega,  
213Upala,214Los  
Chiles,215Guatuso,216Rio  
Cuarto,301Cartago,302Par  
aíso,303La Unión,  
304Jiménez,305Turrialba,3  
06Alvarado,307Oreamuno  
,308El  
Guarco,401Heredia,402Ba  
rva,403Santo Domingo,  
404Santa Bárbara,405San  
Rafael,406San  
Isidro,407Belén,408Flores,  
409San  
Pablo,410Sarapiquí,  
501Liberia,502Nicoya,503  
Santa  
Cruz,504Bagaces,505Carril  
lo,506Cañas,507Abangare  
s,508Tilarán,  
509Nandayure,510La  
Cruz,511Hojancha,601Pun  
tarenas,602Esparza,603Bu  
enos Aires,604Montes de  
Oro,  
605Osa,606Quepos,607Go  
lfito,608Coto  
Brus,609Parrita,610Corred  
ores,611Garabito,701Limó  
n,702Pococí,



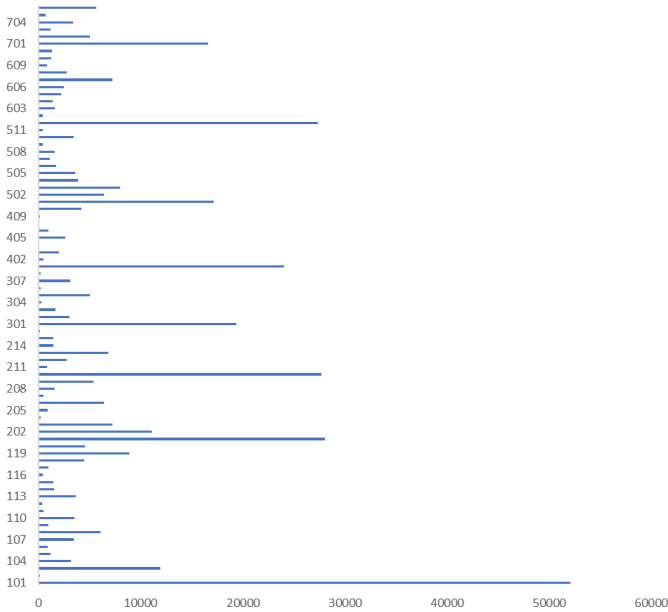
Escala

703Siquirres,704Talamanc  
a,705Matina,706Guácimo

cantonac

Mismo codigo que el anterior

Escala



COD\_MODULO

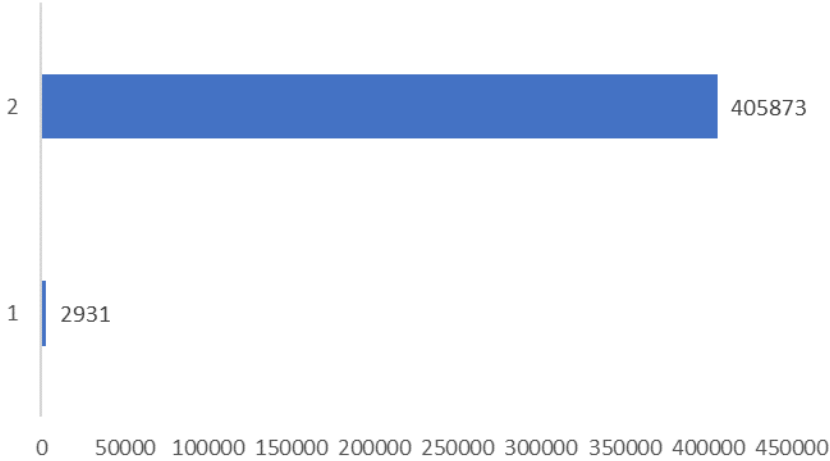
Codigo del modulo, no es relevante

Nominal

DISCAPACID

Si el estudiante cuenta con alguna discapacidad, 1: si y 2: no

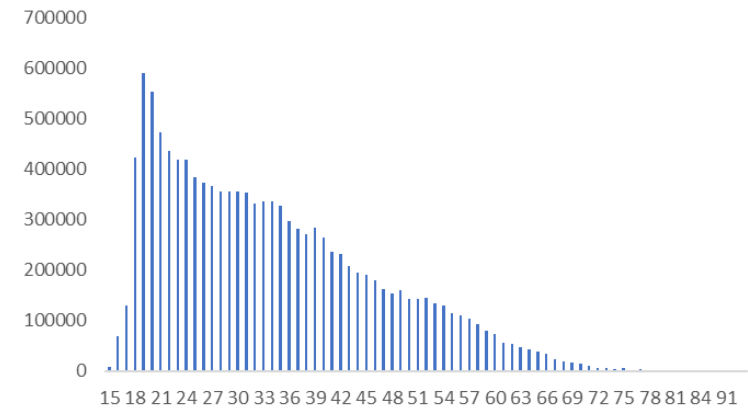
Nominal



EDAD

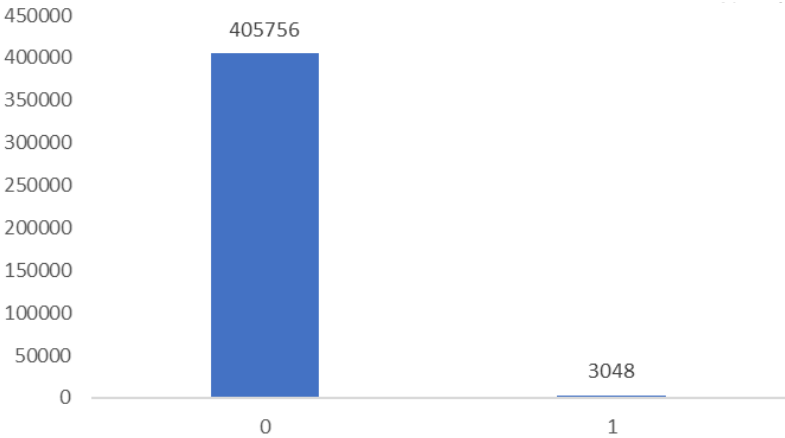
Edad del estudiante

Escala



emprendedor

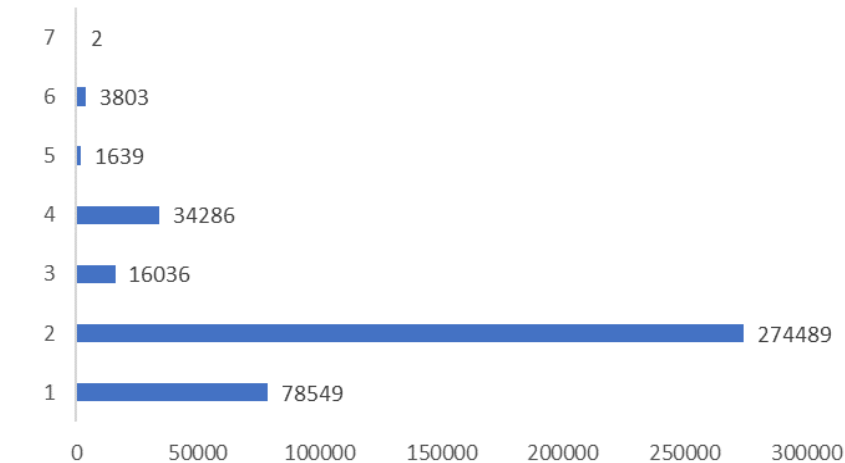
Si la persona tiene algun  
negocio de  
emprendimiento, 0: no y  
1: si



ESTA\_CIVIL

Estado civil de la persona  
1 Casado  
2 Soltero  
3 Divorciado  
4 Unión Libre  
5 Viudo  
6 Separado  
7 No Especificado

Nominal

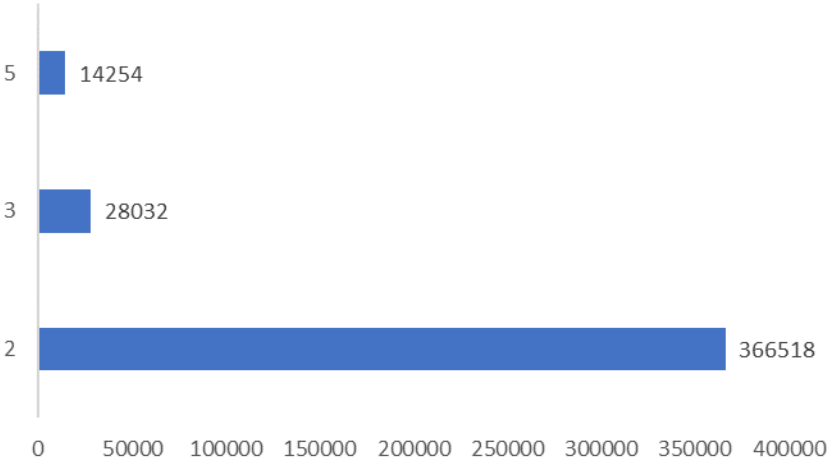


ESTA\_CURSO

Estado del Curso

Nominal

- 1 Matriculado
- 2 Aprobación
- 3 Reprobación
- 4 Equiparado
- 5 Deserción
- 6 NSP
- 7 RPA
- 8 Situación Especial

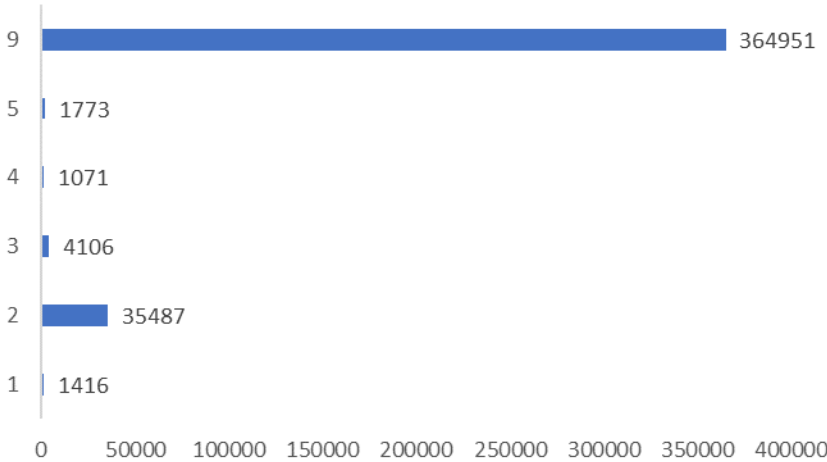


ESTU\_ACT

Estudia actualmente el alumno

Nominal

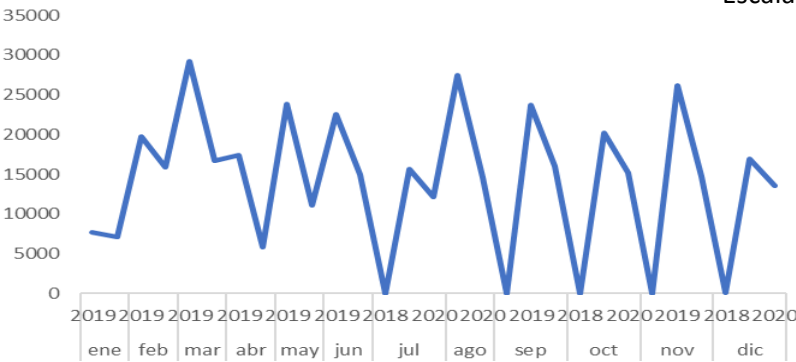
- 1 Primaria
- 2 Secundaria
- 3 Universidad
- 4 Colegio Técnico
- 5 Otro
- 9 No Estudia



FECHA\_FIN

Fecha en la que finaliza el curso

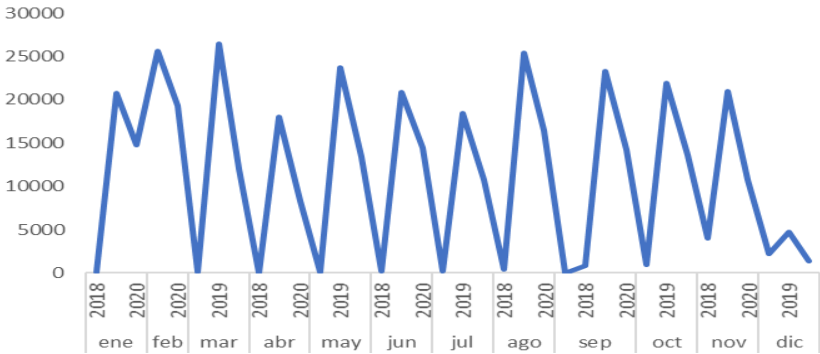
Escala

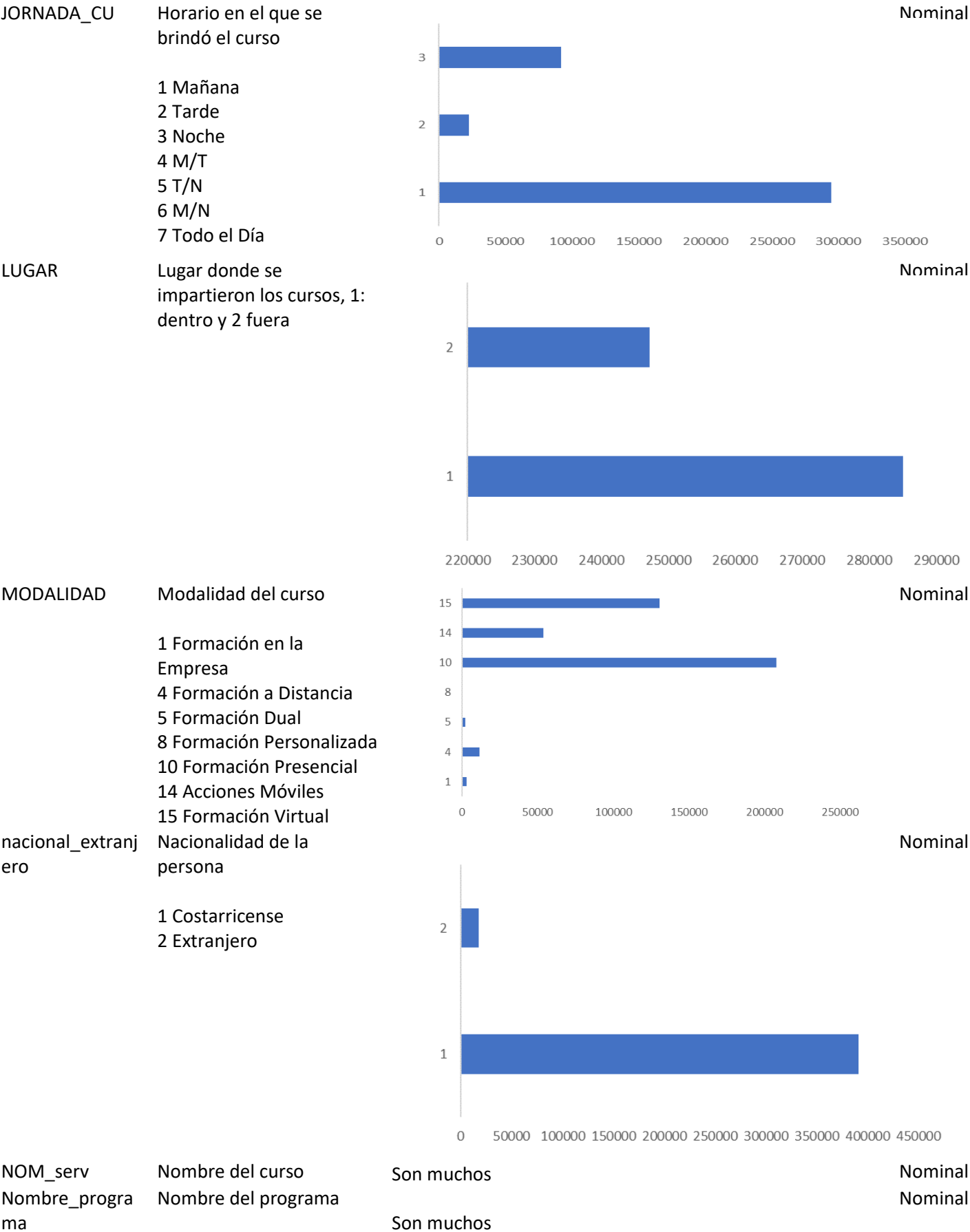


FECHA\_INI

Fecha en la que inicia el curso

Escala





TOTAL\_HRS

Total de horas dedicadas

2500000

Escala

2000000

1500000

1000000

500000

0

9 21 33 40 47 54 62 69 78 85 94 102 112 120 130 141 152 162 171 184 198 210 229 260 272 310 345 430 521 690

aprobo

Si aprobó el curso

1

0

0 50000 100000 150000 200000 250000 300000 350000 400000

## Dataset2

Cantón: trae todos los cantones de CR. Con esta variable se hace el join

Año: desde el 2018 al 2020.

Total: viene el total de empresas según la provincia, cantón y distrito del 2018 al 2020. Esto es importante porque es donde las personas del instituto se puede colocar o las empresas solicitar capacitaciones para el personal.

## Objetivo predictivo

Predecir si los estudiantes van a aprobar o no aprobar los cursos.