INGENIERÍA EN COMPUTACIÓN

MINERÍA DE DATOS

-PROYECTO FINAL: ENCUESTA DE SELECCIÓN DE CARRERA-

DÁVILA ORTEGA JESÚS EDUARDO

FERNÁNDEZ ROSALES SEBASTIAN

HERNÁNDEZ JAIMES ROGELIO YAEL

LÓPEZ BECERRA RICARDO

FECHA DE ENTREGA: 29 DE NOVIEMBRE DEL 2023

INTRODUCCIÓN

La minería de datos ha emergido como una disciplina fundamental en la era digital, brindando a profesionales e ingenieros la capacidad de extraer conocimientos valiosos a partir de grandes conjuntos de datos. En este proyecto, se explora la aplicación práctica de la minería de datos con un enfoque específico en la elección de carrera entre 310 estudiantes. La importancia de esta disciplina radica en su capacidad para revelar patrones, tendencias y relaciones ocultas, proporcionando una base sólida para la toma de decisiones informada. Esto partiendo de todos los algoritmos utilizados durante todas las prácticas y el conocimiento teórico sobre algoritmos de minería de datos, partiendo del funcionamiento de 2 principales softwares: Orange y Power Bi.

Orange es una plataforma de código abierto que combina herramientas de visualización y análisis de datos con potentes algoritmos de minería de datos. Su interfaz gráfica e intuitiva permite a los usuarios explorar y entender la estructura de los datos sin requerir habilidades de programación avanzadas. Además, abarca desde el preprocesamiento de datos hasta la construcción de modelos predictivos. Ofrece una amplia variedad de widgets (bloques de construcción visuales) que facilitan la implementación de algoritmos, la evaluación de modelos y la visualización de resultados.

Power BI, desarrollado por Microsoft, es una herramienta de análisis empresarial que permite la visualización de datos y la creación de informes interactivos. Se integra fácilmente con diversas fuentes de datos, facilitando la transformación de datos en información significativa.

Dominar estos softwares y las técnicas de minería de datos asociadas se ha vuelto esencial para los ingenieros en el entorno empresarial actual. La capacidad para analizar grandes cantidades de datos y derivar información valiosa no solo optimiza la toma de decisiones, sino que también impulsa la innovación. La minería de datos capacita a los ingenieros para abordar problemas complejos, identificar patrones en datos aparentemente caóticos y mejorar continuamente sus procesos y productos. Este proyecto ejemplifica cómo la minería de datos puede aplicarse en el contexto de la selección de carreras, pero su alcance se extiende a diversas disciplinas, contribuyendo significativamente al avance tecnológico y científico, dando una posible solución a esta situación.

PLAN DE TRABAJO

El objetivo planteado para este trabajo es lograr proponer un modelo el cual ayude a la orientación para elegir la carrera que un alumno pueda llegar a escoger para estudiar, esta guía es con el fin de evitar la deserción escolar y la insatisfacción con la carrera que pueda llegar a tener el alumno por tomar una decisión sin una guía clara.

Para lograr este objetivo se utilizarán softwares como power BI para poder realizar de una manera más cómoda la visualización de los datos, así mismo se utilizará el software Orange para realizar el proceso de limpieza, así como la realización de los distintos modelos que puedan ser de ayuda para encontrar las problemáticas planteadas.

Para trabajar se contará con un equipo de 4 personas las cuales trabajaremos en simultaneo para poder tener un desarrollo constante a lo largo de la duración del proyecto. Para esto mismo se realizó el siguiente diagrama de Gantt con el desarrollo esperado del proyecto.

Actividad	Inicio	Fin	14-ago-23	22-sep-23	23-sep-23	07-oct-23	08-oct-23	11-oct-23	12-oct-23	28-oct-23	29-oct-23	30-oct-23	31-oct-23	11-nov-23	12-nov-23	23-nov-23	24-nov-23	27-nov-23	28-nov-23	29-nov-23
Selección de colección de datos	14-ago-23	22-sep-23																		
Exploración y caracterización de datos	23-sep-23	07-oct-23																		
Definición de catálogos de Materias y Habilidades	08-oct-23	11-oct-23																		
Homogenización de datos (Materias, Habilidades, Carreras, etc.)	12-oct-23	28-oct-23																		
Avance de Proyecto (Presentación)	29-oct-23	30-oct-23																		
Estadística de la información (Información transformada)	31-oct-23	11-nov-23																		
Modelos Descriptivos	12-nov-23	23-nov-23																		
Modelos Predictivos	12-nov-23	23-nov-23																		
Documentación escrita (reporte estadístico y reporte de resultados)	24-nov-23	27-nov-23																		
Presentación Final	28-nov-23	29-nov-23																		

Pero este flujo de trabajo puede verse afectado por distintas situaciones en nuestro caso el proyecto corre riesgo de retraso debido a que puede aumentar la carga de trabajo de la carrera y puede llegar a afectar los tiempos que se establecieron en el anterior diagrama de Gantt.

PROBLEMA DE NEGOCIO

Esta encuesta sobre "Selección de Carrera" sirve para recopilar información relevante acerca de las preferencias e intereses de 310 estudiantes de la facultad en relación con la elección de sus carreras profesionales. Se busca comprender los factores que influyen en la toma de decisiones respecto a la elección de una carrera universitaria y su sentimiento con la misma.

Datos relevantes:

- Preferencias personales y profesionales.
- Intereses y áreas de especialización de interés.
- Factores que influencian la elección de carrera (económicos, personales, sociales, etc.).
- Expectativas y metas profesionales a largo plazo.
- Pasión por la carrera.
- Materias que consideren más importantes de la carrera.
- Habilidades que consideren más importantes para la carrera

El proyecto busca proporcionar información valiosa que pueda ser utilizada para asesorar a personas en la toma de decisiones respecto a su futuro académico y profesional.

OBSERVACIONES COLECCIÓN DE DATOS

La colección de datos contiene un total de 310 registros. Cada registro posee un máximo de 36 atributos. La cantidad de atributos puede variar según el registro según la pregunta. Por ejemplo, la pregunta "¿Qué materias del bachillerato consideras relevantes en la elección de tu carrera profesional?" se limita a un máximo de 5 materias, sin embargo, no es obligatorio dar el máximo. En este mismo sentido, cada atributo es resultado de preguntas previamente definidas a las que cada alumno entrevistado dio respuesta.

En el conjunto de preguntas (18 en total) encontramos que se definen 14 preguntas de valor nominal, para las cuales se da un conjunto finito posible de valores. Las siguientes preguntas poseen un valor nominal numérico dado en el intervalo de [1-5]:

- ¿Tu carrera profesional es lo que realmente querías?
- ¿Te apasiona tu carrera profesional?
- ¿Investigaste las oportunidades laborales de tu carrera profesional?
- ¿Visitaste alguna feria o expo de Carreras Profesionales para documentarte?
- ¿Consideraste los recursos financieros y materiales que requiere tu carrera profesional para elegirla?
- ¿Consideraste el Pase automático reglamentado como un criterio de decisión de carrera profesional?
- ¿Consideraste el promedio de calificación mínimo para elegir tu carrera profesional?
- ¿Consideraste la distancia a la universidad como criterio de selección de tu carrera profesional?
- ¿El prestigio del título profesional fue factor para elegir tu carrera profesional?
- ¿El salario promedio para profesionistas de tu carrera fue criterio de decisión?

Los valores numéricos representan valores nominales definidos: 1: No, 2:Parcial, 3:NS/NC, 4:En cierto modo, y 5:Si. Del resto de las preguntas se tiene un total de 4 con respuesta nominal:

- ¿En qué tipos de espacios se ejerce tu carrera profesional? Valores: Oficinas, Campo, Extranjero, Otros.
- La elección de tu carrera profesional, ¿fue una decisión? Valores: Personal, Familiares, Amigos, Otras.
- ¿Qué influencia mediática tuviste para elegir tu carrera profesional? Valores: TV/Medios Impresos, Web/Redes Soc, Carrera de Moda, Carrera Nueva, Ninguno.
- ¿La carrera profesional de tus padres es? Valores: Misma, Similar, Diferente, S/Carrera.

En las preguntas con opción "Otros" u "Otras" se provee una segunda columna donde se especifica un valor propio en el registro. Este valor, pese a no ser definido, provee únicamente la posibilidad de ingresar un único valor, por lo que el manejo de valores fuera de los preestablecidos es posible. El resto de las preguntas son de carácter abierto, limitado a un máximo de 5 respuestas por pregunta. De estas preguntas se tiene una segunda pregunta par, donde se piden valores relacionados con las

respuestas anteriores. Por ejemplo, de la pregunta "¿Que materias del bachillerato consideras relevantes en la elección de tu carrera profesional?" se pueden contestar hasta 5 materias. Por su parte, la pregunta "¿Qué calificación obtuviste en las materias mencionadas? (valores de 0 a 10)" cada respuesta está ligada con cada una de las materias respondidas para la pregunta anterior.

En una exploración superficial de la colección de datos vemos cómo se distribuyen los atributos. Las preguntas nominales de valores predefinidos se define una única columna como atributo correspondiente a esta. Por otra parte, aquellas que ofrecen la opción "Otros" se tiene una segunda columna para describir el atributo, aunque no todos los registros proveen valor a esta opción. Por último, aquellas preguntas de carácter abierto (donde el entrevistado provee respuestas a su criterio) se define una columna para cada una de las cinco posibles respuestas. Así, preguntas como "¿Que materias del bachillerato consideras relevantes en la elección de tu carrera profesional?" poseen cinco columnas asignadas, aunque el registro no necesariamente provea valores para cada una de estas.

ESTADÍSTICA DESCRIPTIVA: APROXIMACIÓN INICIAL

Previa a la realización de modelos para la Minería de Datos, realizamos una exploración de los datos por medios estadísticos con Power BI. Primeramente, podemos comprobar lo descrito en la sección anterior sobre los valores posibles de cada columna:

Catalogo

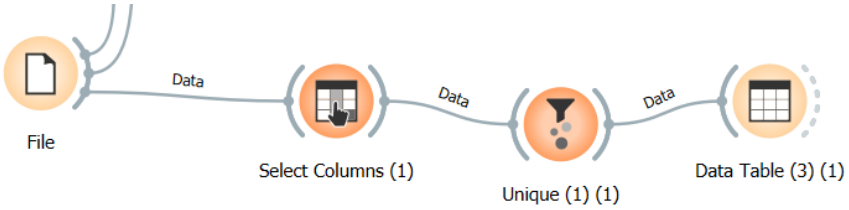
Column1	Column2	Column3	Column4	Column5	Column6	Column7
1	Genero	Oficinas	1	Personal	Misma	TV/Medios impresos
2	Hombre	Campo	2	Familiares	Similar	Web / Redes Soc
3	Mujer	Extranjero	3	Amigos	Diferente	Carrera de Moda
4	null	Otros	4	Otros	S/carrera	Carrera Nueva
5	null	null	5	null	null	Ninguno
null	null	null	6	null	null	null
null	null	null	7	null	null	null
null	null	null	8	null	null	null
null	null	null	9	null	null	null
null	null	null	10	null	null	null

Entre los registros, se encontraron valores que pese a ser similares (o representar teóricamente lo mismo) se tomaban como valores distintos dada la forma en que se ingresaron los valores. Por ejemplo, en el atributo "Carrera" se encontraron respuestas de "Ingeniería en Computación", sin embargo, se encontraron variaciones de esta al omitir acentos, o por el uso de mayúsculas:

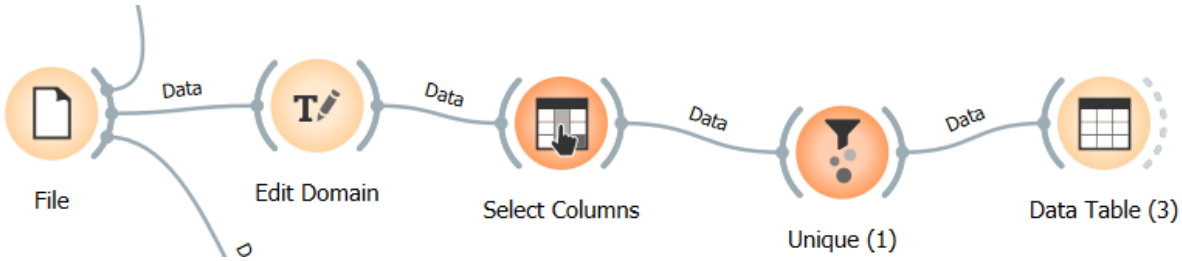
	CARRERA	AD_
1	ingenieria computacion	18
2	Ingeniería en Computación	21
3	Administracion	24
4	Ingeniería de Software	30
5	Informatica	30
6	Medicina veterinaria	6

Estas inconsistencias generarían un sesgo en nuestros modelos al evaluar como valores distintos a aquellos que representan lo mismo, pero fueron ingresados con variaciones. Teniendo esto en cuenta hacemos un preproceso para homogenizar aquellas respuestas cuyo carácter lo permita. Por ejemplo, el atributo género mostraba cuatro valores posibles. Empleando los widgets Unique y Select Columns pudimos extraer estos valores:

	GENERO
1	Hombre
2	Mujer
3	hombre
4	mujer



Con esto podemos emplear el widget Edit Domain para editar estos valores y homogenizar las respuestas a “Hombre” y “Mujer”, por lo que las variaciones en minúsculas son reemplazadas con su respectivo equivalente:



	GENERO
1	Hombre
2	Mujer

LIMPIEZA DE DATOS

Homogenización

La homogenización de datos desempeña un papel fundamental en proyectos de minería de datos, especialmente cuando se enfrenta a conjuntos de datos que presentan inconsistencias y variabilidades en su estructura. Este proceso, aunque a menudo puede requerir esfuerzos manuales,

es esencial para garantizar la calidad y confiabilidad de los resultados obtenidos a través del análisis de datos.

En el transcurso de este proyecto, nos encontramos con un conjunto de datos proveniente de la encuesta que exhibía notables inconsistencias, las cuales no podían ser corregidas mediante los recursos automatizados proporcionados por los widgets de Orange. Ante esta circunstancia, optamos por realizar una exhaustiva revisión manual con el objetivo de identificar y abordar dichas inconsistencias.

Durante el proceso de revisión manual, llevamos a cabo una clasificación meticulosa de las irregularidades detectadas (ej. Empatía, empatía). Estas irregularidades se corrigieron a modo de clasificar en una misma clase todas estas variaciones. Además, durante estas correcciones, se siguieron parámetros predefinidos por nuestro equipo: todas las clases empiezan por mayúsculas y se omiten acentos. Esto último con el fin de tener una mejor respuesta de la computadora al analizar los valores dados.

Adicionalmente, en el caso de las materias, decidimos agrupar todas las materias relacionadas con las matemáticas bajo la categoría "Matemáticas". Este enfoque no solo implicó la consolidación de información altamente relacionada, sino también la inclusión de aquellas instancias que presentaban faltas ortográficas en la misma categorización.

Estas correcciones no solo permitieron homogeneizar la presentación de datos, sino que también facilitó la posterior aplicación de técnicas y algoritmos en el proceso de minería de datos. La inclusión de categorías específicas, basadas en criterios definidos, no solo contribuyó a la claridad y consistencia de los datos, sino que también allanó el camino para análisis más profundos y significativos.

B	C	D	E	F	G	H	I	J	K	L
Habilidad(valor final)	**Posibles valores**									
Adaptabilidad	Adaptabilidad, Adaptacion, rápida adaptación									
Administracion	Administración									
Algoritmica	Algoritmica									
Analitico	Analisis,Análisis, Análisis Crítico, Pensamiento analítico, Razonamiento analítico, analisis, analista, analizar, análisis, análisis de sistemas, observacion									
Autoaprendizaje	Autoaprendizaje, autodidacta, habilidad autodidacta									
Pensamiento Abstracto	Capacidad de Abstracción, Pensamiento abstracto, abstraccion, imaginacion espacial, manejo de conceptos abstractos									
Trabajo en Equipo	Compañerismo, Trabajo en equipo, Trabajo en quipo, compañerismo, trabajo en equipo									
Comprension	Comprension, Entendimiento, comprensión									
Comprometido	Aplicación, Comprometido, Dedicacion, Dedicación, compromiso									
Comunicacion	Comunicación,Expresion oral y escrita, Expresión, Habilidades comunicativas, comunicación, habilidades comunicativas, hablar en público									
Confianza	Confianza									
Conocimiento Contextual	Conocimiento del contexto ambiental									
Constancia	Constancia, constancia									
Creatividad	Crear, Creatividad, Creativo, Cretividad, Imaginación, Ingenioso, creatividad, factibilidad de ideas									
Pensamiento Critico	Criterio, Pensamiento critico, Pensamiento crtitico, Pensamiento crítico, Toma de decisiones, criterio, pensamiento crítico, toma de decisiones									
Disciplina	Disciplina, disciplina, disciplinado									
Diseño	Diseño, diseño									
Empatia	Empatia, Empatía, empatia, empatía									
Emprendimiento	Emprendimiento									
Enseñanza	Enseñar									
Planeacion	Estrategia, Planificacion, Planificación, creación de proyectos									
Estructurar	Estructurar									
Habilidad Matematica	Analisis matematico, Habildiad Matemática, Habildiad matemática, Hambilidad numérica, Matemática, Matemáticas, Pensamiento, matematico, Pensamiento matemático, Razonameinto Matemático, habilidad para									
Hidraulica	Hidráulica									

Hidraulica	Hidráulica
Lengua Extranjera	Idiomas, Inglés, idiomas, ingles
Innovacion	Innovación, inovacion
Inteligencia	Inteligencia, Inteligencias
Investigacion	Investigación, saber buscar informacion
Lectura	Lectura
Liderazgo	Liderar
Logica	Lógica, Logica, Lógico, Pensamiento lógico, Razonamiento Lógico, logica, lógica, lógicas
Manejo de equipo	Manejo de equipo, Manejo de sistemas operativos
Objetividad	Objetivo
Organizacion	Organización, Programación de actividades, admon del tiempo, organización
Paciencia	Paciencia, paciencia
Pasion	Pasión, apasionado, pasión
Perseverancia	Perseverante, perseverancia, persistencia
Proactividad	Proactivo
Productividad	Productivo
Programacion	Programación, programacion, programacion básica
Destreza	Rapidez, destreza
Razonamiento	Razonamiento, razonamiento
Resolucion de Problemas	Resolución de Problemas, Resolución de problemas, Resolver problemas, para resolver problemas, resolución de problema, resolución de problemas, resolver problemas
Respeto	Respeto
Responsabilidad	Responsabilidad, responsable
Memorizacion	Retención, memoria
Sociable	Sociable, sociable, social
Trabajo	Trabajo
Ventas	Ventas

Trabajo	Trabajo
Ventas	Ventas
Vision	Vision, visión
Altruismo	altruismo
Sintesis	capacidad de síntesis, sintesis
Conceptualizacion	conceptualización
Contabilidad	contabilidad
Dominio Conocimientos	dominio conocimientos
Fisica	fisica
Pensamiento Prospectivo	pensamiento prospectivo, prevención
Pragmatica	pragmática
Relaciones Interpersona	relaciones interpersonales
Tolerancia	tolerancia, tolerante
Control Emocional	tolerancia a la frustración, trabajar bajo presión, trabajo bajo presión
Pensamiento Flexible	pensamiento flexible

La importancia de la homogenización radica en varios aspectos clave, ya que la homogenización garantiza que los datos estén presentados de manera coherente y sigan un formato uniforme. Esto es esencial para realizar análisis comparativos y obtener conclusiones significativas. La inconsistencia en la representación de datos puede llevar a interpretaciones erróneas y conclusiones no fiables. Como se mencionó en la estadística descriptiva, puede haber casos donde la carrera “Ingeniería en Computación” podría presentarse en diferentes formatos (ej. ingeniería en computación, Ing. computación) y todas estas refieren al mismo valor. Sin embargo, para los algoritmos que puedan aplicarse se considerarían propiamente como dos clases distintas.

Homogenizar datos implica la estandarización de variables, lo que simplifica la comparación y el análisis. Por ejemplo, la inclusión de todas las materias de cálculo en la categoría "Matemáticas" permite consolidar información y facilita la identificación de patrones y tendencias específicas a través de las diferentes instancias de las variables.

Trabajar con datos homogéneos simplifica los procesos de limpieza y preprocesamiento. En lugar de abordar inconsistencias individualmente, el analista puede enfocarse en patrones y características relevantes para el análisis, acelerando así el flujo de trabajo. Así como la clasificación manual de inconsistencias, como la inclusión de materias con faltas de ortografía en una categoría específica, mejora la consistencia y facilita la posterior aplicación de técnicas de minería de datos, como la construcción de modelos predictivos o la identificación de tendencias.

La homogenización contribuye a la interpretabilidad de los resultados. Al presentar datos de manera uniforme, se simplifica la comunicación de hallazgos y conclusiones a partes interesadas, facilitando la comprensión y toma de decisiones informadas.

De esta manera pasamos de tener una colección de datos con poco más de 30 columnas, a una colección en donde obtuvimos más de 180 columnas, esto es debido a la codificación One-Hot que realizamos con nuestros datos. La codificación One-Hot fue utilizada en el procesamiento de datos categóricos para convertir variables categóricas en una forma que puede ser proporcionada a algoritmos de aprendizaje automático, en este caso los algoritmos que implementaremos de minería de datos.

Lo que hicimos, en términos sencillos, en lugar de representar las categorías como valores numéricos directos, se crea una columna binaria (0 o 1) para cada categoría posible, y solo una de estas columnas tiene un valor de 1 para indicar la presencia de la categoría, mientras que las demás tienen un valor de 0. Además de que para los valores donde se necesitaba una calificación, hicimos lo mismo, solo que aquí si se asignaba la calificación que se otorgó a la materia.

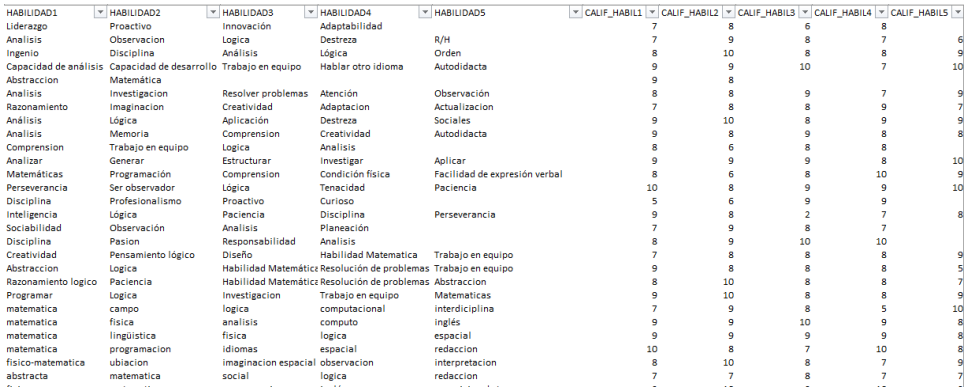
Esto lo hicimos ocupando con un Script de Python, donde nos transformaba estos valores a una columna distinta para un archivo de Excel.

Siguiendo la misma lógica se cambiaron las habilidades y materias, esto con ayuda de un script realizado en Python.

Script para transformar las habilidades:

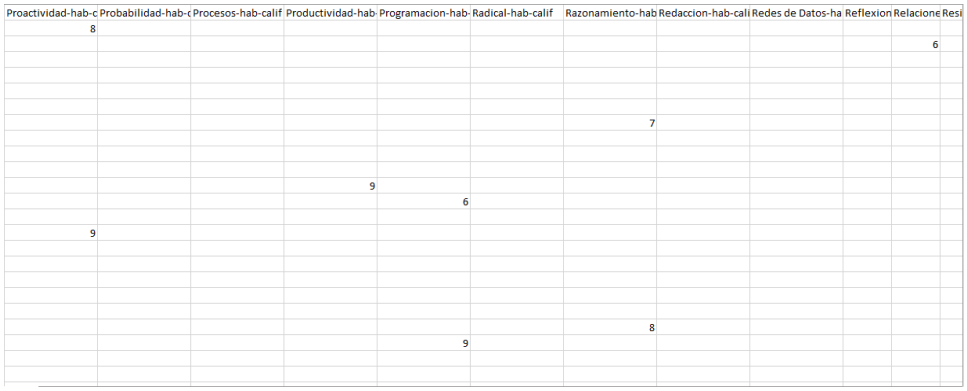
```
import pandas as pd
from Orange.data import Table, Domain, pandas_compat


df = pandas_compat.table_to_frame(in_data)
df1 = df[['HABILIDAD1', 'HABILIDAD2', 'HABILIDAD3', 'HABILIDAD4', 'HABILIDAD5']].copy()
df1 = df1.stack().str.get_dummies().groupby(level=0).sum()
df1.values[df1 != 0] = 1
#print(df1)
df = df.reset_index()
for num_habilidad in range(1, 6):
    for index, row in df.iterrows():
        if row[f'CALIF_HABIL{num_habilidad}']:
            df1.loc[index, row[f'HABILIDAD{num_habilidad}']] = row[f'CALIF_HABIL{num_habilidad}']
df1 = df1.add_suffix('-hab')
out_data = pandas_compat.table_from_frame(df1)
```

Forma original de la colección de datos:



HABILIDAD1	HABILIDAD2	HABILIDAD3	HABILIDAD4	HABILIDAD5	CALIF_HABIL1	CALIF_HABIL2	CALIF_HABIL3	CALIF_HABIL4	CALIF_HABIL5
Liderazgo	Proactivo	Innovación	Adaptabilidad		7	8	6	8	
Analisis	Observacion	Logica	Destreza	R/H	7	9	8	7	6
Ingenio	Disciplina	Análisis	Lógica	Orden	8	10	8	8	9
Capacidad de análisis	Capacidad de desarrollo	Trabajo en equipo	Hablar otro idioma	Autodidacta	9	9	10	7	10
Abstraccion	Matemática				9	8			
Analisis	Investigacion	Resolver problemas	Atención	Observación	8	8	9	7	9
Razonamiento	Imaginacion	Creatividad	Adaptacion	Actualizacion	7	8	8	9	7
Análisis	Lógica	Aplicación	Destreza	Sociales	9	10	8	9	9
Analisis	Memoria	Comprension	Creatividad	Autodidacta	9	8	9	8	8
Comprension	Trabajo en equipo	Logica	Analisis		8	6	8	8	
Analizar	Generar	Estructurar	Investigar	Aplicar	9	9	9	8	10
Matemáticas	Programación	Comprension	Condición física	Facilidad de expresión verbal	8	6	8	10	9
Perseverancia	Ser observador	Lógica	Tenacidad	Paciencia	10	8	9	9	10
Disciplina	Profesionalismo	Proactivo	Curioso		5	6	9	9	
Inteligencia	Lógica	Paciencia	Disciplina	Perseverancia	9	8	2	7	8
Sociabilidad	Observación	Analisis	Planeación		7	9	8	7	
Disciplina	Pasion	Responsabilidad	Analisis		8	9	10	10	
Creatividad	Pensamiento lógico	Diseño	Habilidad Matematica	Trabajo en equipo	7	8	8	8	9
Abstraccion	Logica	Habilidad Matemática	Resolución de problemas	Trabajo en equipo	9	8	8	8	5
Razonamiento logico	Paciencia	Habilidad Matemática	Resolución de problemas	Abstraccion	8	10	8	8	7
Programar	Logica	Investigacion	Trabajo en equipo	Matematicas	9	10	8	8	9
matematica	campo	logica	computacional	interdiciplina	7	9	8	5	10
matematica	fisica	analisis	computo	inglés	9	9	10	9	8
matematica	lingüistica	fisica	logica	espacial	9	9	9	9	8
matematica	programacion	idiomas	espacial	redaccion	10	8	7	10	8
fisico-matematica	ubiacion	imaginacion espacial	observacion	interpretacion	8	10	8	7	9
abstracta	matematica	social	logica	redaccion	7	7	8	7	7

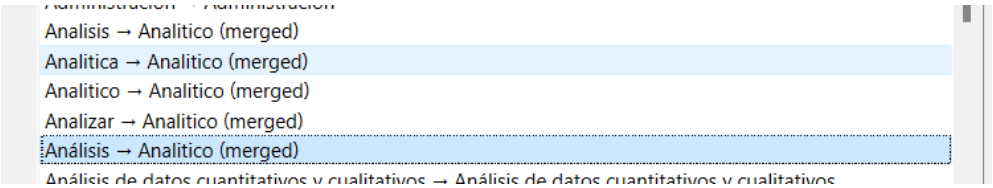
Con la programación One-Hot:



Proactividad-hab-c	Probabilidad-hab-c	Procesos-hab-calif	Productividad-hab	Programacion-hab	Radical-hab-calif	Razonamiento-hab	Redaccion-hab-cali	Redes de Datos-ha	Reflexion	Relacione	Resi
8											
										6	
						7					
			9								
					6						
9											
						8					
					9						

A pesar de la importancia de la homogenización, es crucial reconocer que puede ser un proceso que consume tiempo, especialmente en proyectos donde los datos presentan múltiples irregularidades. Sin embargo, la inversión de tiempo y esfuerzo en este paso inicial puede traducirse en un análisis de datos más preciso y confiable, con resultados que brindan una base sólida para la toma de decisiones y la identificación de patrones significativos. En última instancia, la homogenización representa una etapa esencial para transformar datos crudos en información valiosa y utilizable.

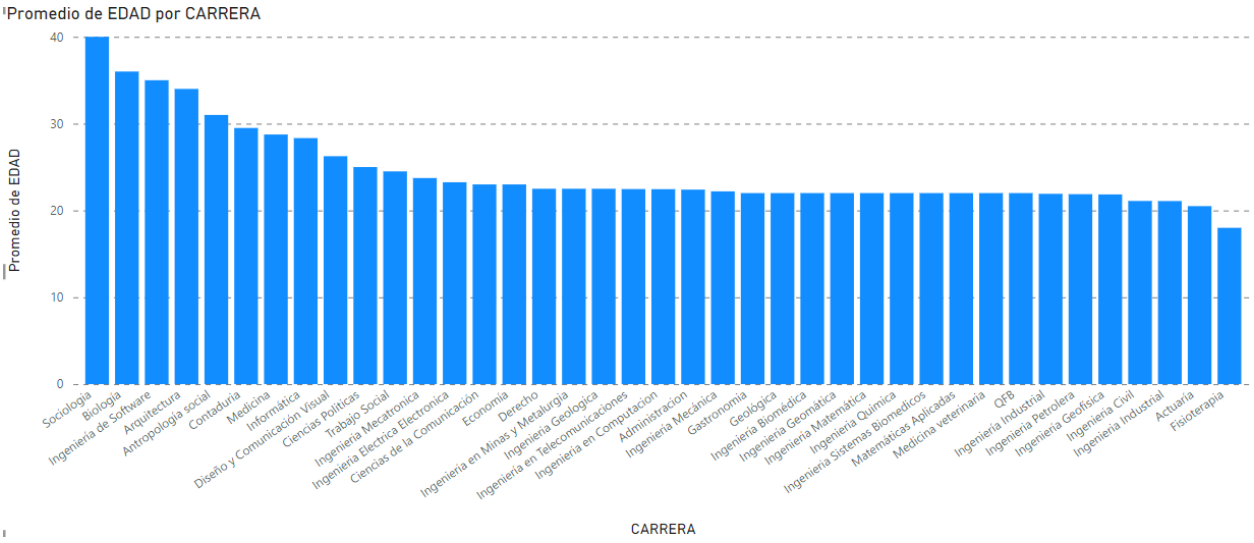
Para nuestros análisis, se decidió aplicar la homogenización a todos los atributos cuya respuesta de tipo textual, pues incluso con la existencia de clases predefinidas, los registros podían tener variaciones en la escritura de estas. Particularmente, se hizo énfasis en la homogenización de los atributos de Materias y Habilidades, las cuales presentaban una mayor tendencia a variaciones en las respuestas. Se pretende hacer uso de estas como variables predictoras para un modelo predictivo que emplea árboles de decisión, tal y como se verá posteriormente.



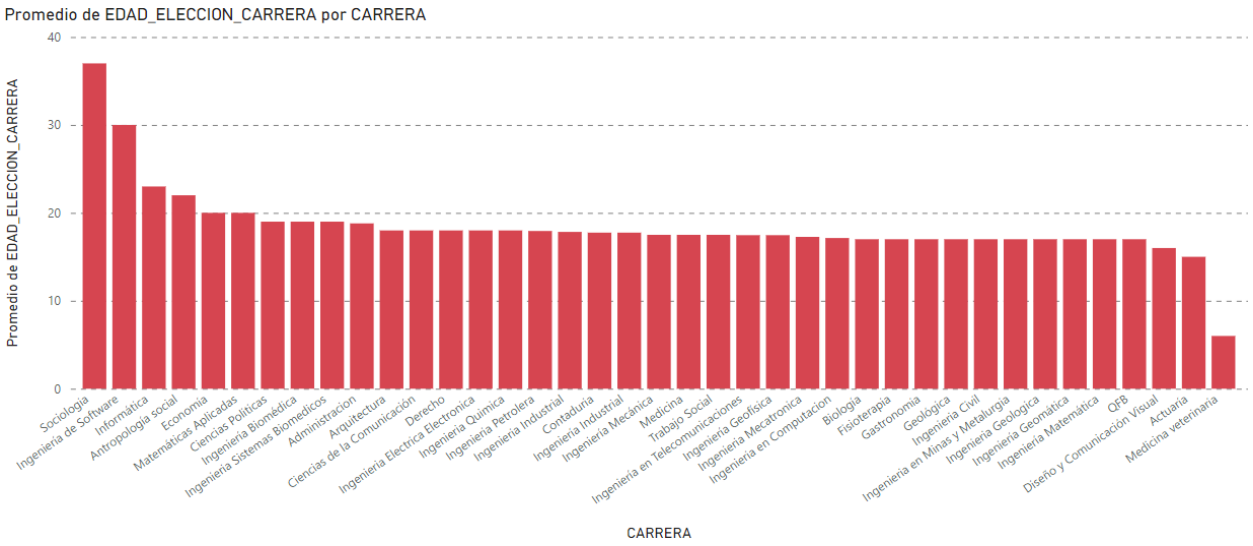
REPORTE DE ESTADÍSTICAS DE **INFORMACIÓN QUE FUE TRANSFORMADA**

Estadística de la información (información preprocesada)

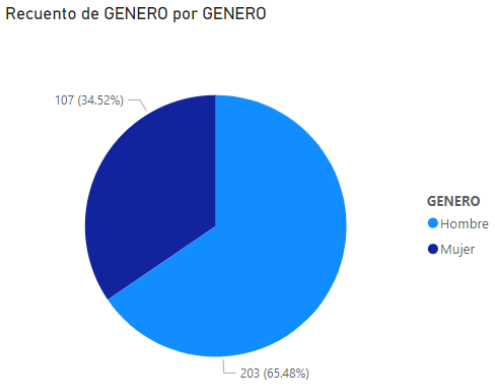
Ya con las correcciones mencionadas con anterioridad, se procede a hacer una exploración estadística de la información. Se pretende visualizar la información contenida en la colección y tratar de comprender los distintos parámetros estadísticos que nos provee la información. En esta sección buscamos primeramente comprender las características de nuestros datos, refiriéndonos al contexto sobre el cual fueron obtenidos. De este modo, primero se aborda un estudio sencillo de los datos. Por ejemplo, usando Power Bi, se observa que la edad promedio de los encuestados por carreras ronda los 23 años, siendo que en las carreras como Sociología poseen una edad promedio de 40 años. Este último caso (y algunos similares) se deben a que dichas carreras contienen pocos o un único registro (en el caso de Sociología) donde la edad de los encuestados rondaba esta edad.



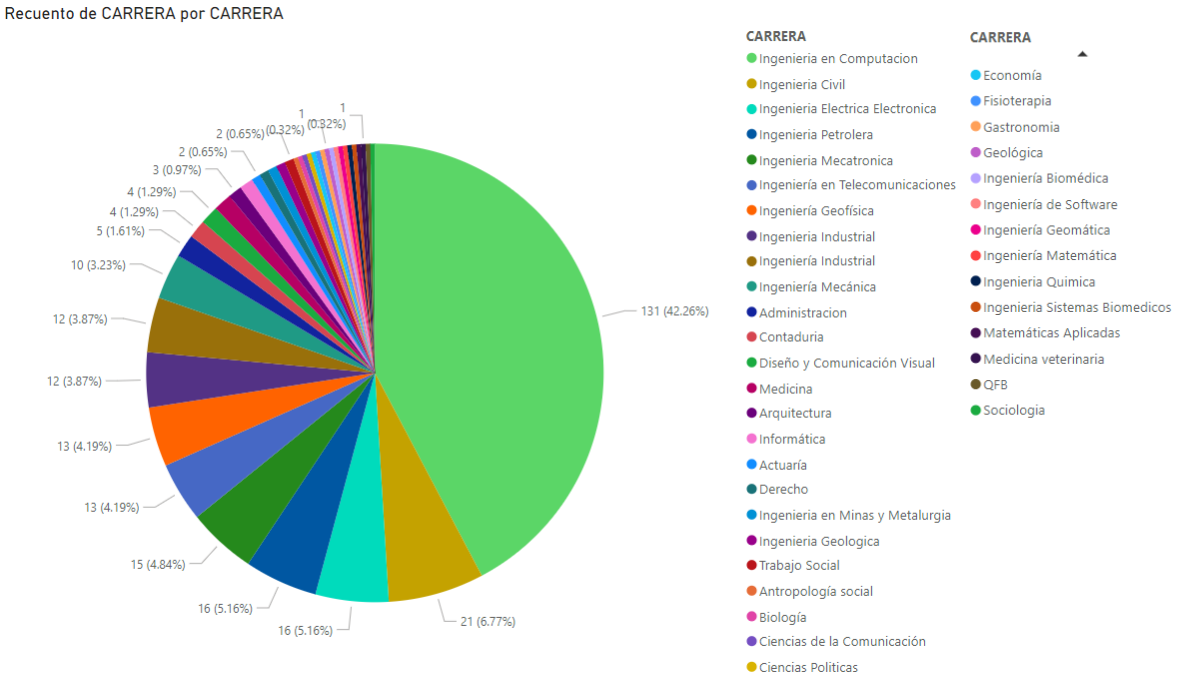
Por otra parte, la edad promedio de elección de carrera donde en torno a los 18 años, nuevamente con las variaciones de algunas carreras como Sociología (37 años) o medicina veterinaria (6 años). Podemos concluir que la mayoría de encuestados tenían 18 años al momento de realizar la elección de carrera. Este factor resulta importante, pues como se verá más adelante, el grado de satisfacción en este rango de edad es más variante. Si, por ejemplo, tomáramos los registros de mayor edad, encontraríamos que estos son los que tienden a tener un mayor grado de satisfacción y pasión por su carrera. De ahí que el uso de un posible modelo predictivo esté orientado a los jóvenes entre 17 y 18 años para ayudarles a tener una mejor aproximación a la carrera que puedan elegir.



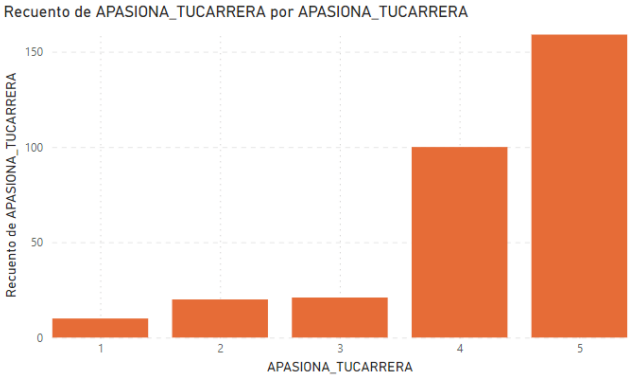
Además, de esto, encontramos que un 65.48% corresponden a hombres, mientras que el restante 34.52% corresponde a mujeres.



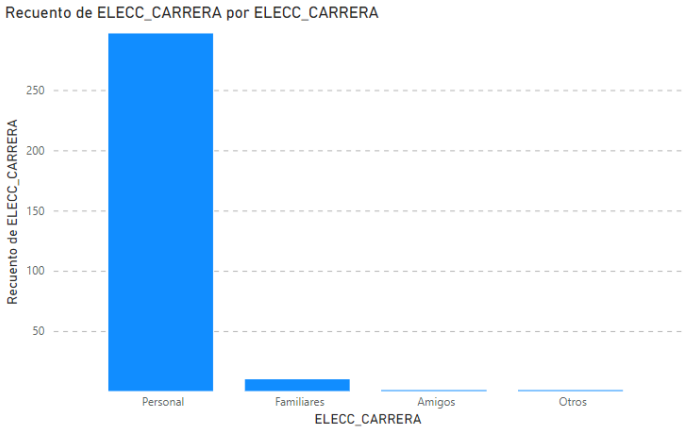
Por otra parte, podemos observar la distribución de las carreras en los encuestados. Notamos una gran presencia de Ingeniería en Computación (42.26%), una cantidad a tener en cuenta en el desarrollo de modelos y para poder sacar conclusiones de los resultados de estos. Tras esta ingeniería, la ingeniería civil posee un 6.77% del total de entrevistados. Carreras como Biología, Sociología QFB e Ingeniería Química cuentan con un solo registro cada una.



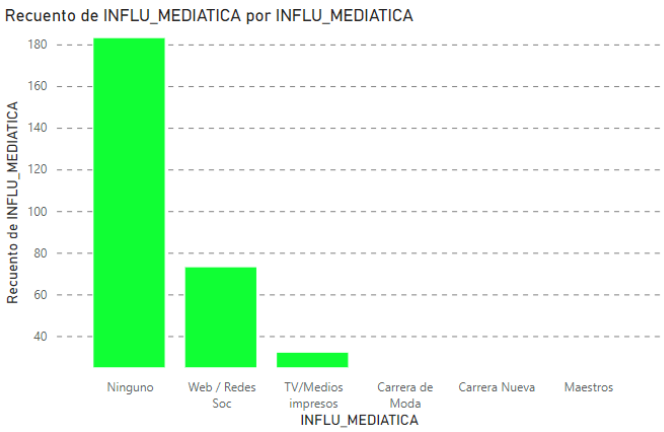
Otro factor a revisar es el valor que los encuestados indican que les apasiona su carrera. Este factor es importante puesto que nos permite comprender más si la elección de la carrera, a perspectiva de cada encuestado es adecuado de acuerdo a qué tanto le apasiona su carrera. En este caso, observamos que una buena parte de los encuestados (159) tienen el más alto nivel de pasión por su carrera. Sin embargo, aproximadamente un sexto de la muestra (51) se encuentran en nivel medio o bajo (clase 1, 2 y 3). Este factor nos permite ver que posiblemente la elección de carrera de estos sujetos no fue la mejor guiada o no tuviese las mejores referencias originalmente. Además, el estado 4, si bien es por encima del "regular" podría tenerse en cuenta pues son personas que tienen un interés en su carrera, pero posiblemente tuviesen algunos problemas menores que podrían haberse evitado. Estos valores son los que nos hacen optar por la generación de un modelo predictivo que ayude a las personas (más orientado a grupos jóvenes como los encuestados) tengan una primera sugerencia como carrera posible. No buscamos instruir a la persona en escoger la carrera resultante, sino sugerir una primera opción que podría considerar investigar y que podría interesarle.



Otro factor que podemos observar el tipo de elección hecha. En este campo, la gran mayoría indicó que fue una elección personal. La gran cantidad de respuestas de este tipo, nos hace comprender que la elección de carrera tiene un carácter casi por completo personal. Por esto, si se quisiese hacer propaganda de alguna carrera para incitar sus inscripciones, el uso de publicidad dirigido a familiares y amigos podría no tener efecto significativo en el resultado.



Por último, observamos la influencia de los medios para la propaganda de las distintas carreras vistas en la encuesta. Esto, si bien no será empleado directamente en los algoritmos aquí utilizados, podría tener relevancia para futuros análisis. Por ejemplo, si una determinada carrera no está recibiendo suficiente número de inscripciones, podría verse los medios más efectivos para hacer promoción de la misma, captando mayor atención de nuevos estudiantes.



REPORTE DE INTERPRETACIÓN DEL RESULTADO DE CADA ALGORITMO

Para realizar la implementación de cada algoritmo partimos de la creacion de los modelos descriptivos y predictivos. Los Modelos Predictivos y Descriptivos son dos pilares fundamentales en el ámbito de la minería de datos. Cada uno cumple con un propósito específico:

Los Modelos Descriptivos se centran en comprender y resumir la estructura y características fundamentales de los datos, identificando patrones y relaciones entre variables. Funcionan como un mapa que guía al científico de datos a través del conjunto de datos desconocido. En cambio, los

Modelos Predictivos buscan predecir valores futuros o clasificar datos en categorías específicas. Utilizan los patrones identificados por los modelos descriptivos para hacer inferencias sobre datos no vistos, siendo esenciales en la toma de decisiones basadas en datos y herramientas poderosas para anticipar tendencias y comportamientos futuros.

Tanto para un Científico de Datos como para un Ingeniero, la Generación de Modelos Predictivos y Descriptivos representa un proceso esencial por varias razones:

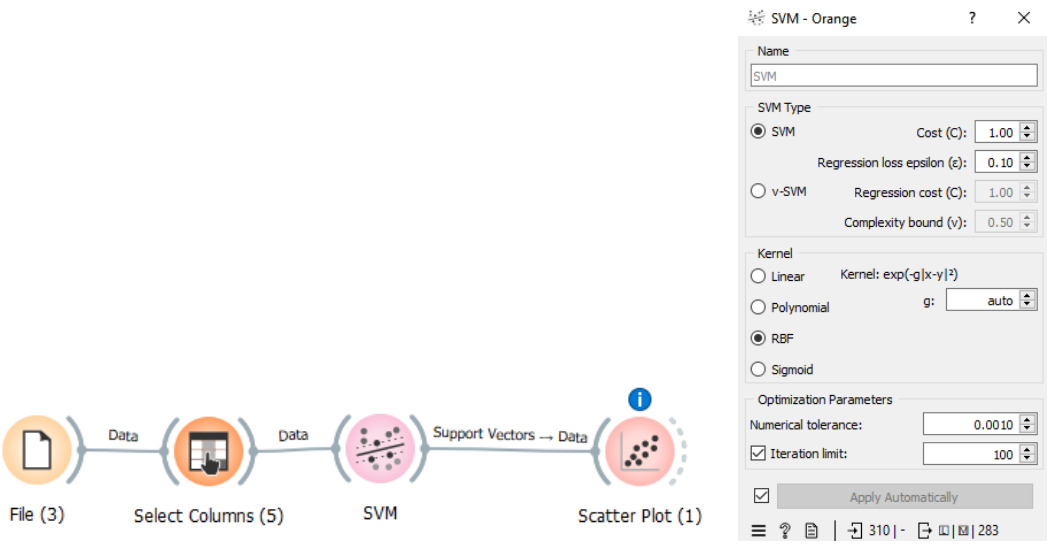
- Guía en la Toma de Decisiones: Proporcionan una base sólida para tomar decisiones informadas. Los modelos descriptivos permiten entender a fondo el comportamiento de los datos, mientras que los modelos predictivos brindan la capacidad de anticipar escenarios futuros.
- Optimización de Procesos: Al comprender las relaciones entre las variables, se pueden identificar áreas de mejora y optimizar procesos para aumentar la eficiencia y reducir costos.
- Mejora en la Precisión: Los modelos predictivos permiten predecir con mayor precisión los resultados futuros, lo que puede traducirse en un aumento de la eficacia operativa y una ventaja competitiva.
- Innovación y Desarrollo: Estos modelos son fundamentales para la innovación, ya que proporcionan información valiosa sobre cómo mejorar productos o servicios existentes, o incluso para crear soluciones completamente nuevas.

Modelado Descriptivo

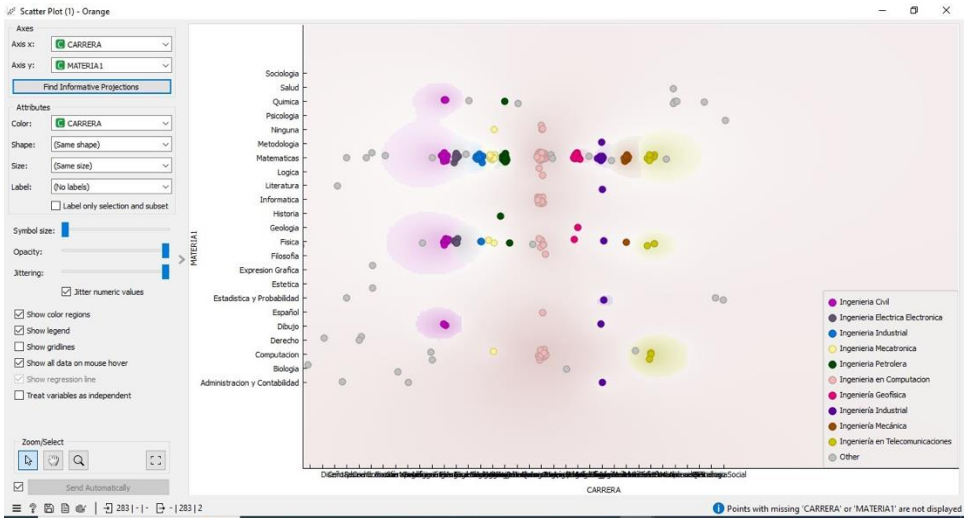
Clasificación SVM

El Support Vector Machine (SVM) es una técnica de aprendizaje supervisado que destaca por su capacidad para clasificar datos y prever valores numéricos. En el contexto de Orange, este algoritmo es instrumental para definir hiperplanos óptimos que separan clases y facilitan la identificación de patrones en conjuntos de datos, permitiendo así realizar clasificaciones más precisas.

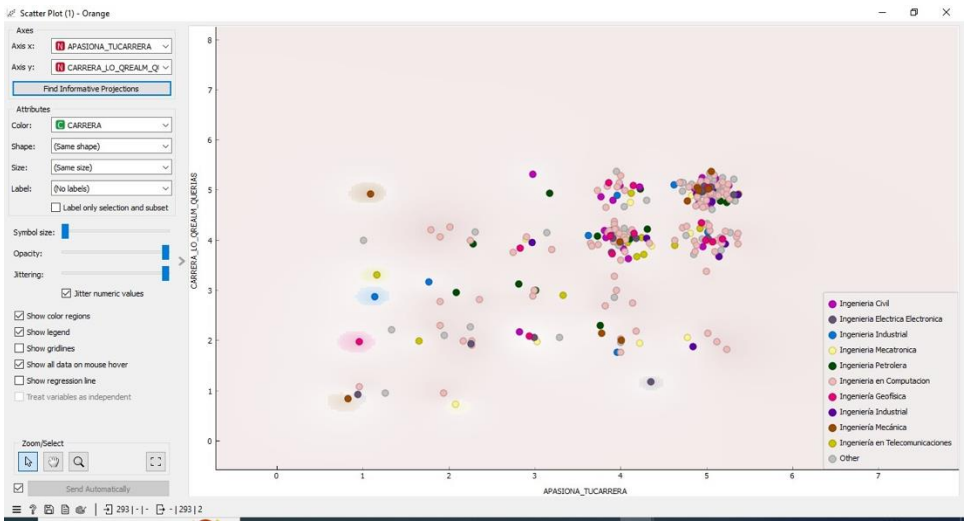
El Scatter Plot, por otro lado, es una herramienta visual poderosa que representa datos en un espacio bidimensional, proporcionando una visión gráfica de la distribución y relaciones entre variables. En Orange, el Scatter Plot se utiliza para explorar patrones, agrupaciones y correlaciones entre datos. La relación entre SVM y Scatter Plot radica en la capacidad del segundo para ofrecer una visión inicial de los datos, inspirando la aplicación del primero para una clasificación más detallada y la predicción de valores específicos basados en patrones identificados visualmente. Esta combinación brinda una aproximación integral y efectiva para explorar y modelar datos de manera precisa.



En el primer modelo, al analizar las materias consideradas más importantes en relación con las carreras, se obtuvo una visión detallada de las preferencias académicas. Por ejemplo, se observó que en ingeniería civil existe una gran valoración de la materia de dibujo, mientras que en computación esta percepción es considerablemente menor. Esta información estratégica proporcionada por el SVM sobre la importancia de las materias para cada carrera sienta las bases para futuros modelos predictivos, permitiendo una orientación más precisa y personalizada para los estudiantes en función de sus preferencias académicas y profesionales.



En el primer modelo, al comparar la pasión por la carrera con la percepción de si es lo que realmente querían, se identificó que la mayoría de los encuestados experimenta una conexión positiva con sus carreras, reflejando satisfacción y alineación con sus expectativas. Sin embargo, se destacó un grupo significativo que presenta menor grado de apasionamiento o no encuentra su carrera como lo esperaba. La aplicación de SVM permitió identificar las carreras más propensas a esta situación, brindando así un enfoque específico para abordar inquietudes y mejorar la experiencia de este segmento de estudiantes.

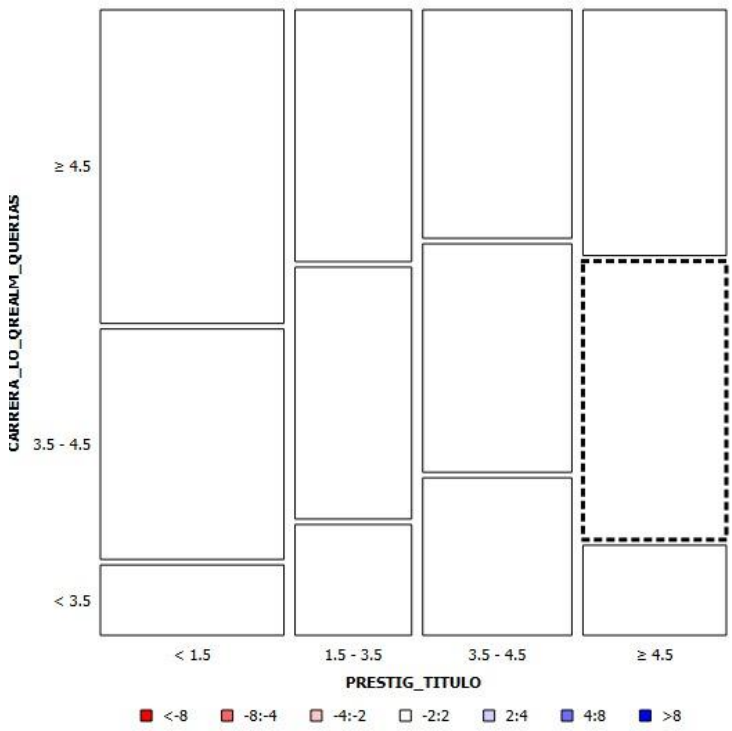


Mosaicos

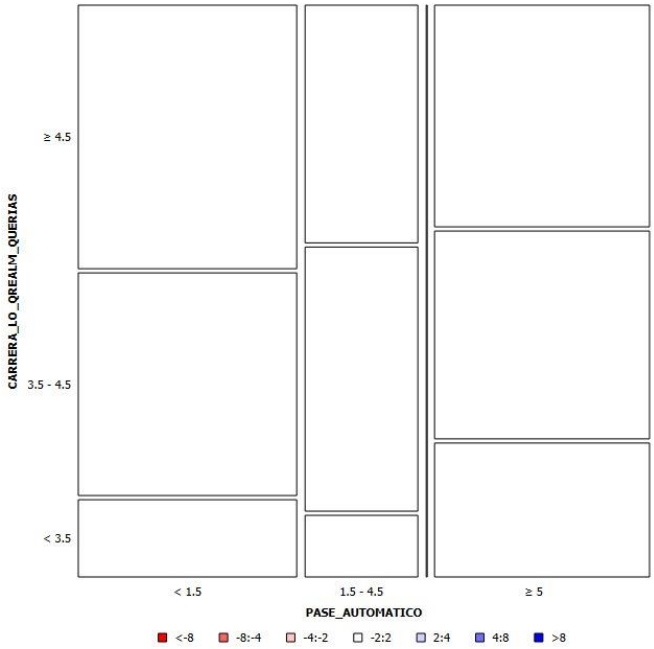
En este contexto, la elección del modelo de mosaicos (mosaicos regresión) puede ser más apropiada. Este método, también conocido como regresión de mosaicos, divide el espacio de entrada en regiones más pequeñas y ajusta un modelo lineal o polinómico a cada región. Esto permite capturar relaciones más específicas en áreas donde hay datos limitados, mejorando la capacidad del modelo para adaptarse a la variabilidad en las calificaciones de habilidades.

Para nuestro caso específico, los mosaicos resultaron ser una herramienta muy útil para encontrar relaciones entre los atributos de los registros. Con este modelo logramos encotrar características de los estudiantes interesantes.

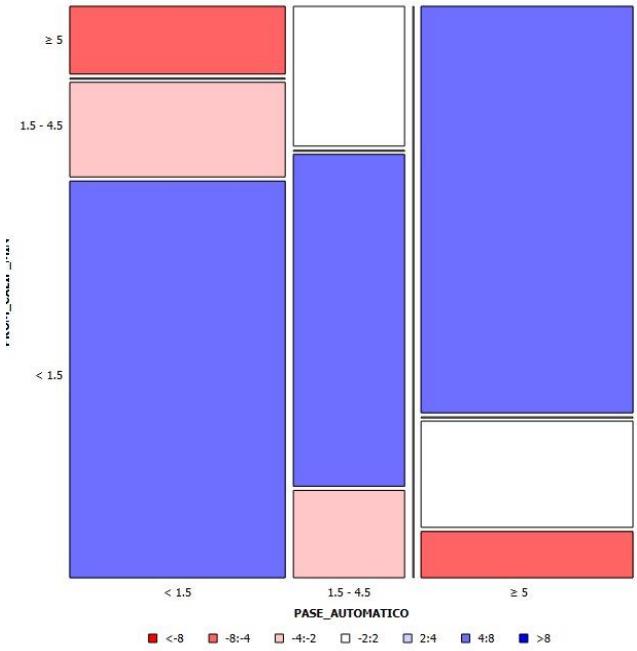
Entre más prestigioso el titulo, es menos probable que el alumno este a gusto con la carrera. En la siguienete gráfica en el eje horizontal se encuentra la importancia que le dieron los estudiantes a el pretigio del titulo obtenido con la carrera. En el eje vertical se representa la satisfaccipon de los estudiantes con su carrera. En la gráfica se puede apreciar como al aumentar la importancia del titulo en la elección de carrera, la proporción de alumnos a gusto con la misma disminuye.



Entre màs importo el pase reglamentado en el proceso de elección de carrera menos a gusto suelen estar los alumnos. De manera similar al caso anterior, cuando aumentamos la importancia del pase reglamentado la satisfacción de los alumnos es proporcionalmente menor. Una vez más, en el eje vertical se encuentra la satisfacción de los alumnos y en el horizontal la importancia del pase reglamentado.

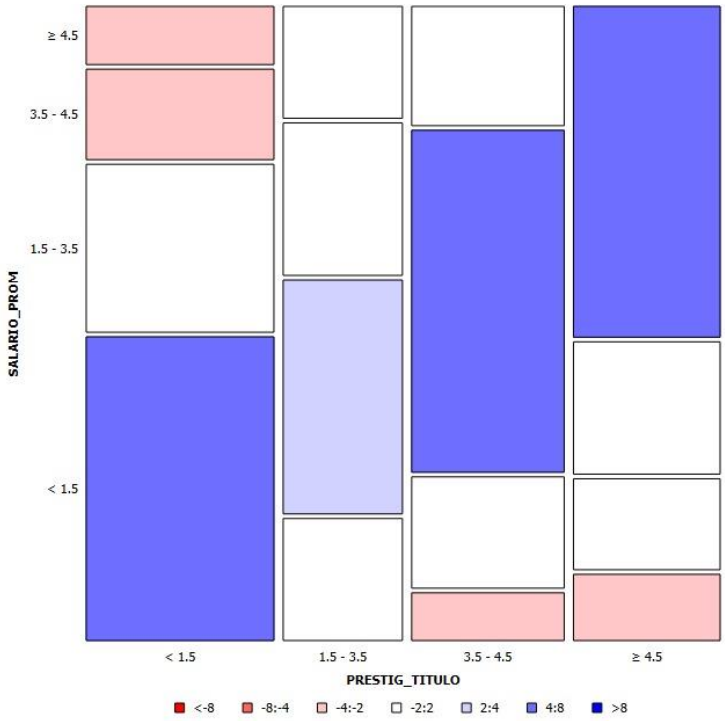


Entre más le importa el pase reglamentado a los estudiantes más les importa el promedio. Esta es una deducción facil de hacer, pero con los datos de la encuesta podemos comprobarlo. En el eje horizontal se encuentra la importancia del pase reglamentado y en el eje horizontal se encuentra la la importancia que le dieron los alumnos al promedio de calificación mìnimo requerido para entrar en la licenciatura.



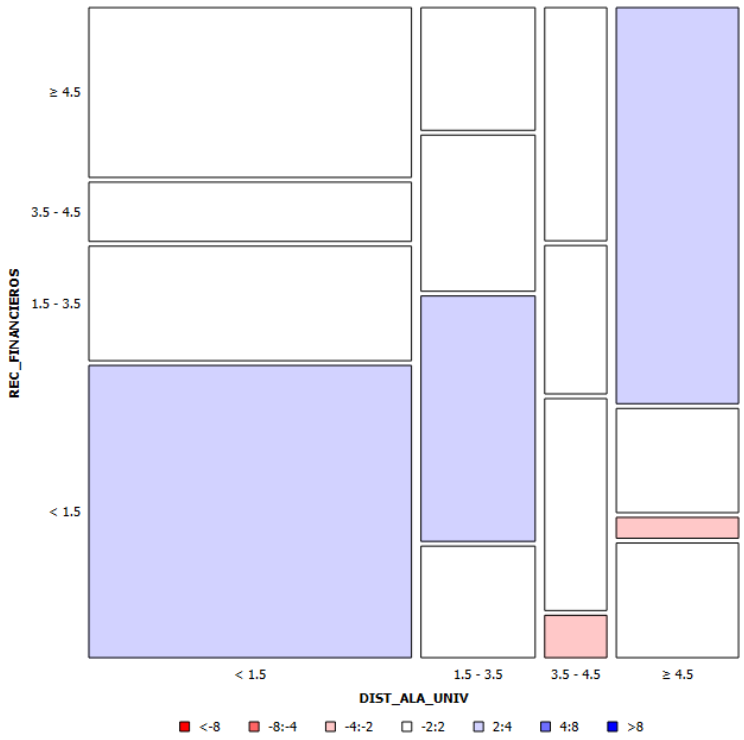
Darle más importancia al prestigio del título está correlacionado con darle más importancia al salario promedio de la carrera. Cuando los alumnos le dieron gran importancia al presitigo

de la carrera tambien le dieron gran importancia al salario promedio de un egresado de la carrera. En la siguiente gráfica, en el eje horizontal tenemos el prestigio del titulo y en el eje horizontal tenemos el salario promedio de un egresado. Como se puede ver en la diagonal principal, hay una relación directa entre el salario del egresado y el prestigio del titulo.

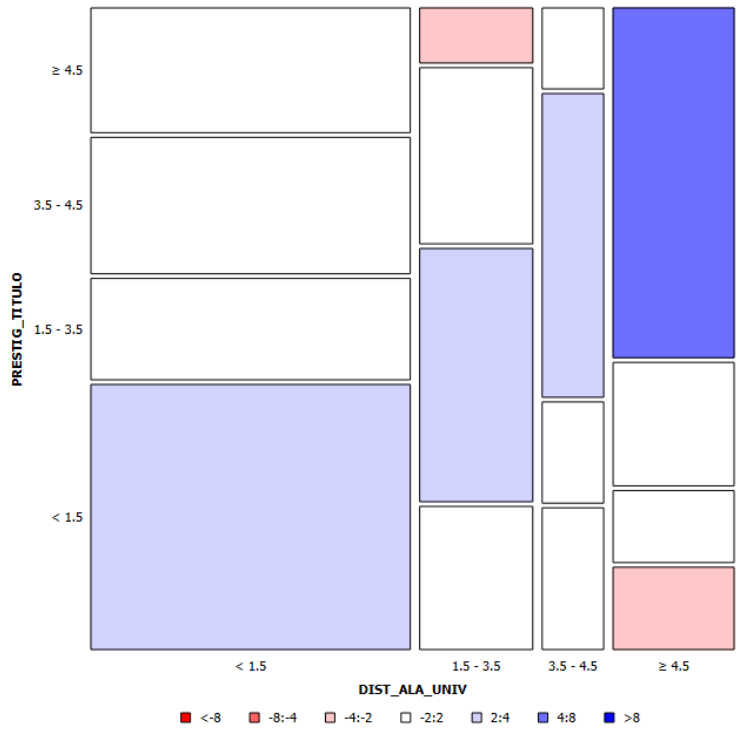


La distancia a la universidad hace que se le den mayor importancia a otros factores los alumnos, pero eso no implica que tomen una mejor decisión. Al analizar el atributo de dsitancia a la escuela nos percatamos de que esta variable esta relacionada con otras varaibles. Lo que encontramos es que los alumnos que consideraron a la distancia como un factor importante consideraron otros factores con mayor importancia que los alumnos que consideraron menos la distancia. Los factores relacionados encontrados fueron el prestigo de la licenciatura, los recuros económos y el pase reglamentado, aunque no descartamos que este relacionado con más. En las siguientes gráficas se muestra la relación cone estas tres variables. En el eje horizontal permanece la distancia a la escuela y en el eje vertical cambia la variable evaluada.

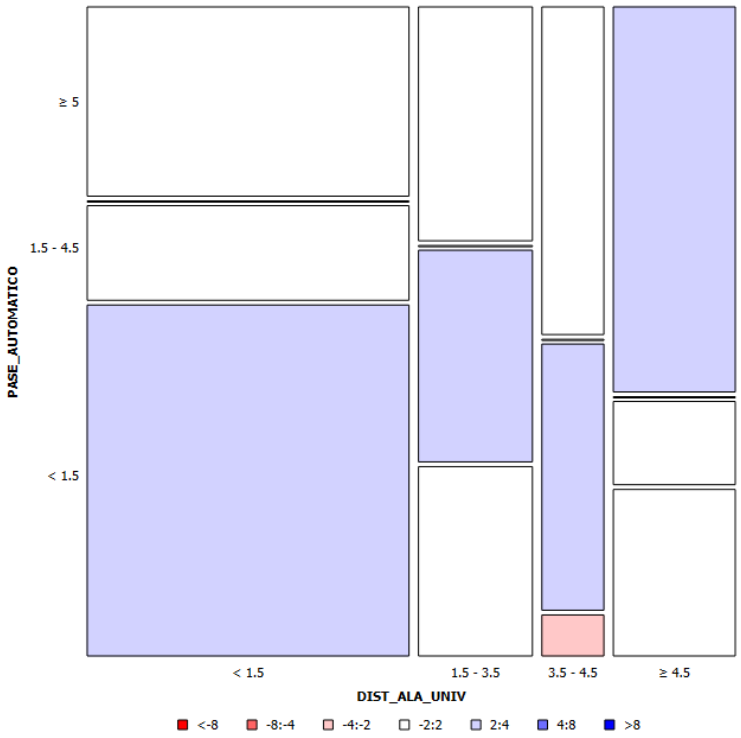
- Distancia y recursos financieros: En esta gráfica se puede apreciar como la proporcion de alumnos que tomaron más en cuenta los recursos financieros aumenta con respecto a la distancia a la escuela.



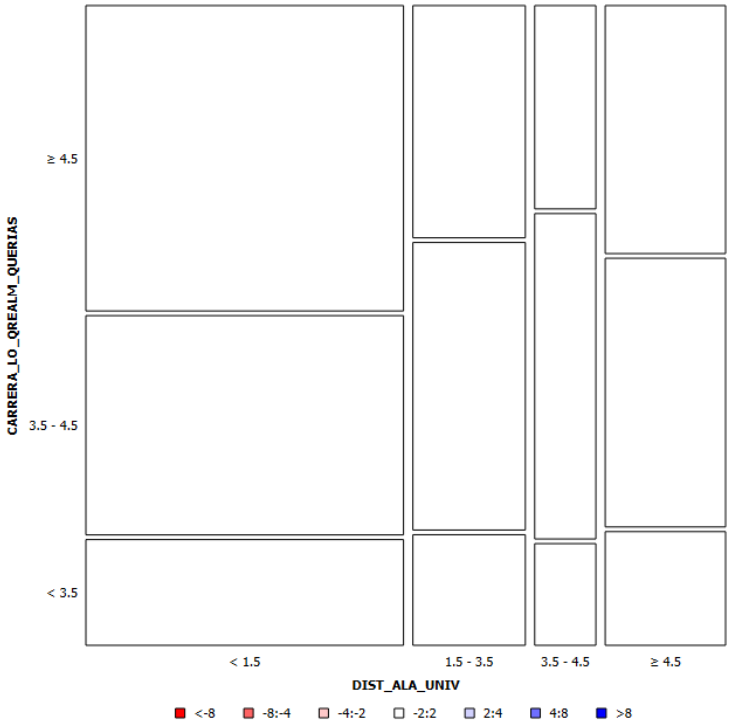
- Distancia y prestigo de la licenciatura: Al aumentar la importancia de la distancia a la universidad, aumenta la proporción de alumnos que le dan importancia al prestigio de la licenciatura. Esto se aprecia en el aumento del tamaño de los recuadros superiores.



- Distancia y pase reglamentado: De manera similar, el pase reglamentado importa más a los alumnos cuando aumenta la importancia que se le da a la distancia a la universidad.



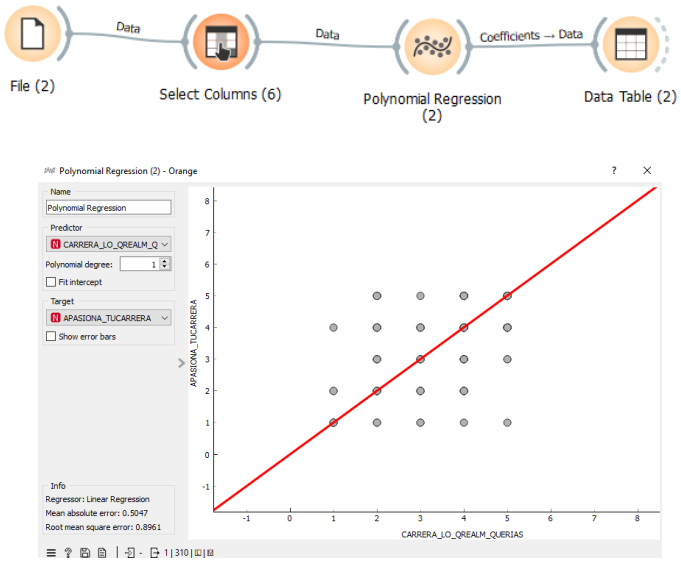
En estos ejemplos podemos observar que la distancia tiene relación con otros factores. Cuando se considera importante a la distancia, también se consideran importantes otros aspectos. Sin embargo, aunque muchos aspectos se consideran importantes, la decisión final no es mejor para estos alumnos. En la sugiente gráfica el eje horizontal corresponde a la distancia a la escuela y el eje vertical corresponde a la satisfacción de lo sestudiantes. Como se puede apreciar, el recuadro de la esquina superior derecha no es significativamente más grande que los otros, por lo que su decisión no es notoriamente mejor a oesar de darle más importancia a más factores.



Regresión lineal

La regresión lineal es un método estadístico utilizado para modelar la relación entre una variable dependiente y una o más variables independientes, asumiendo que esta relación puede aproximarse mediante una línea recta. El widget "Regresión Polinómica" en Orange expande el concepto de regresión lineal al permitir la aproximación de la relación mediante un polinomio de grado superior. Esto significa que, en lugar de ajustarse a una línea recta, el modelo se adapta a una curva polinómica, proporcionando mayor flexibilidad para capturar patrones más complejos en los datos.

La regresión lineal y polinómica requieren una cantidad significativa de datos para identificar patrones y realizar predicciones precisas. Cuando la cantidad de datos es limitada, estas técnicas pueden no proporcionar resultados robustos ni generalizables, ya que pueden ajustarse demasiado a los datos existentes, por eso es que para nuestro proyecto este algoritmo no es tan eficiente.



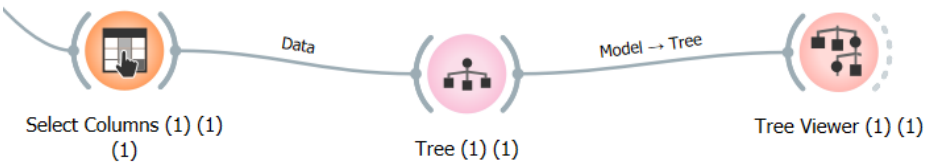
En conclusión, para el modelado descriptivo, el enfoque de mosaicos es el más adecuado dadas las limitaciones de datos específicas. Este nos proporcionará un modelado descriptivo más preciso al considerar de manera más efectiva las relaciones entre habilidades y calificaciones, incluso en áreas con pocos datos disponibles.

Modelo predictivo

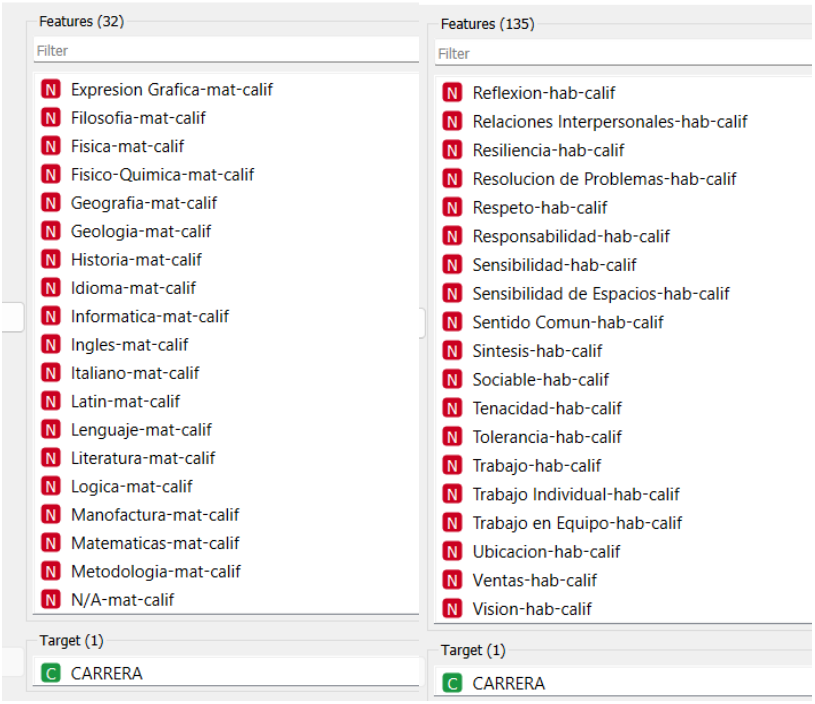
Para este modelo, verdaderamente realizamos dos modelos para predecir la carrera que podría ser a consideración del encuestado: por habilidades y por materias. Este enfoque parte de las necesidades ya no solo teóricas de las carreras, sino también de las habilidades soft, que forman al profesionista. El fin de estos modelos es proveer una herramienta de aproximación para futuros estudiantes que tengan que realizar la elección de carrera. No se busca determinar su carrera, sino sugerirle una carrera que podría adecuarse a sus habilidades e intereses, incitándolo a investigar primero sobre dicha carrera.

Arboles de decisión

Para la elaboración del modelo predictivo se tomaron los campos de materias y habilidades. Tras la homogenización, los posibles valores de estos campos fueron convertidos en columnas. Esto con el fin de tener una tabla donde se indique si una materia fue listada por el encuestado o no. Tras el tratamiento con Python realizamos esta conversión de valores a columnas. Así mismo, se enlazó la calificación asignada por el encuestado a su respectiva materia o habilidad. Habiendo acondicionado la colección para el uso del algoritmo de clasificación por árboles de decisión, empleamos los widtgets select column, Tree y Tree Viewer para obtener el modelo:

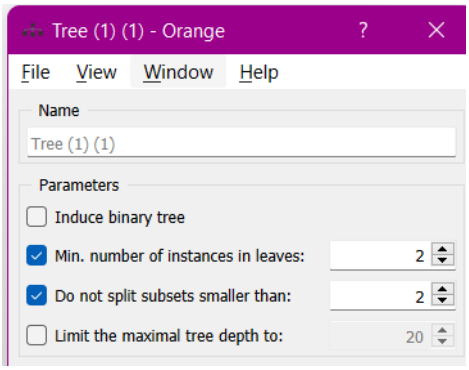


Las columnas seleccionadas en cada modelo corresponden a las columnas indicadas anteriormente. De esta forma, tenemos algo similar a lo que se muestra a continuación:

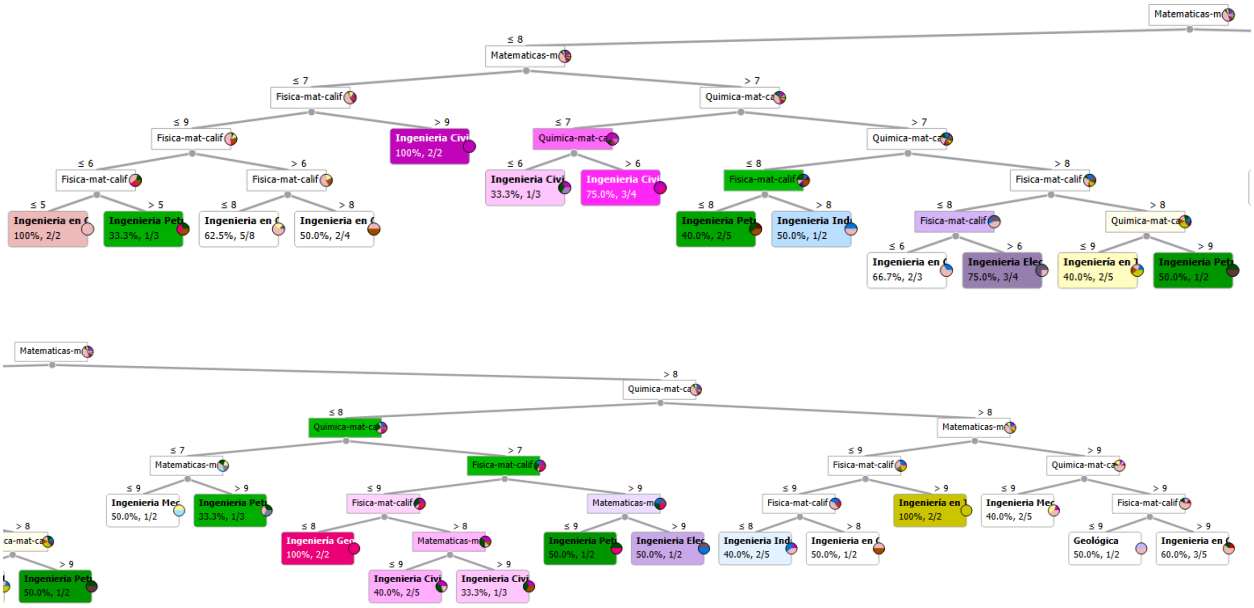


En ambos casos la variable objetivo es la Carrera, mientras que en los campos empleados se tienen aquellas que ya registran la calificación dada por el encuestado a las materias y habilidades listadas. Cabe recordar que las materias y habilidades se encuentran separadas en distinto modelo.

En los parámetros de los árboles a desarrollar tenemos variación respecto al número de instancias por hojas. En el caso del modelo predictivo con materias, se estableció en dos instancias. Esto porque el modelo mostró una mejor cobertura de las posibles carreras. Encontramos una mejor relación de las materias con el tipo de carrera según la percepción de los encuestados. Los parámetros del árbol para este modelo son los siguientes:

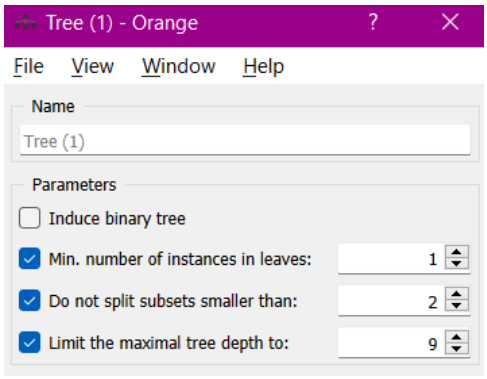


El árbol resultante sería el siguiente:

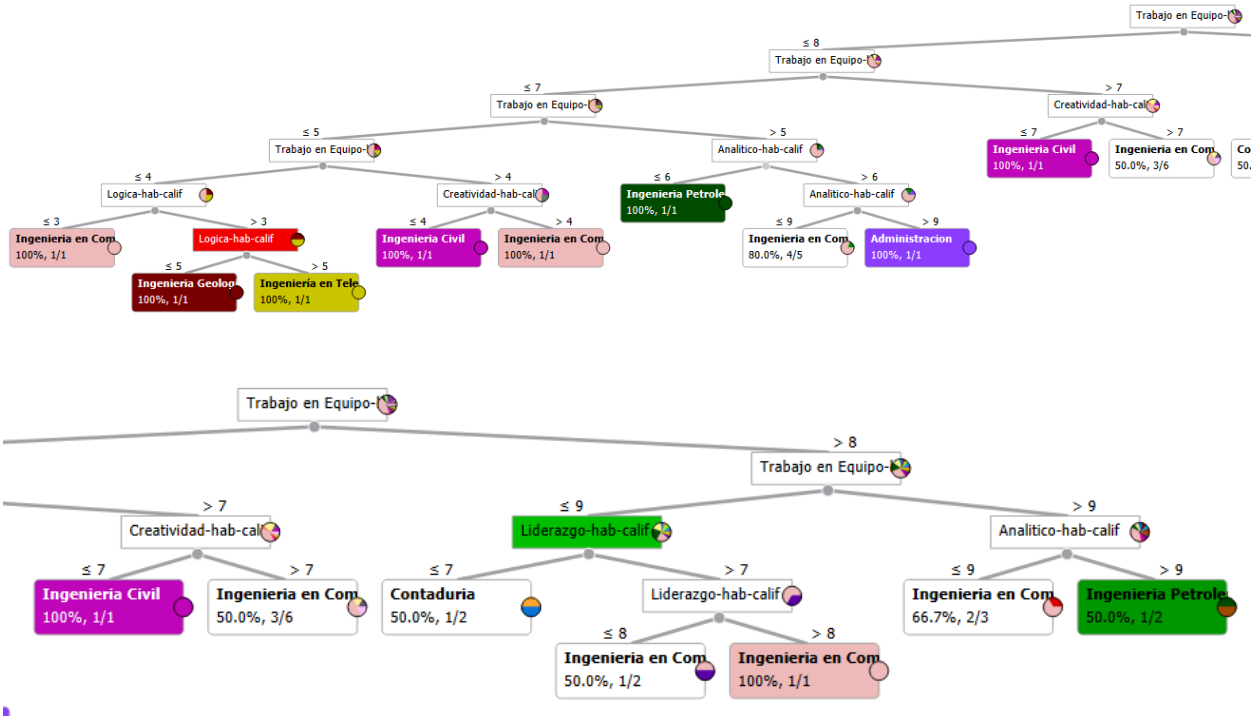


Se puede observar que existe una notable cantidad de repeticiones de carreras, como la Ingeniería en Computación o Ingeniería Petrolera. Así mismo, se observa que no se cubren todas las posibles carreras disponibles en nuestra colección. Esto es sencillo de explicar: poca variedad de datos sumado a un gran sesgo a ingeniería en computación. Respecto a la variedad de datos, tenemos que muchas carreras poseen un único registro en toda la colección, de ahí que primero, los valores de atributos pueden no ser de gran valor para el árbol y que de serlo, generaría un perfecto sobre ajuste para esa carrera en particular.

Por su parte, el modelo que considera las habilidades posee los siguientes parámetros:



El número de instancias por hojas se reduce a 1, lo que da la cobertura de un mayor rango de posibilidades. El árbol resultante es el siguiente:



En ambos modelos podría debatirse el uso de 1 instancia como mínimo por nodo hoja. Esto de forma natural podría tenerse como un sobre ajuste del árbol. Sin embargo, esto podría justificarse por la propia naturaleza de la colección de datos. Como vimos en la estadística de la información, un buen número de carreras contaba únicamente con un solo registro en el total de encuestas. Todas estas carreras se verían descartadas en automático, pues sus nodos hojas tenderán a tener una única instancia. En adición a esto, y como se puede observar en ambos modelos, la ingeniería en computación tiene una gran presencia. Esto se comprende fácilmente en la estadística de información, donde se visualizaba que el 42% de los registros correspondían a la carrera de ingeniería en computación. El sesgo de la muestra es evidente. Sin embargo, ya se cuenta con pocos

registros, por lo que el descartar aquellos de ingeniería en computación solo resultaría contraproducente.

Dado a que los modelos tienen un carácter de aproximación no buscamos asegurar un 100% de tasa de éxito. No se pretende indicar al encuestado qué carrera tomar sino cuál podría ser le de interés. Por ello mismo, incluso al no tener una tasa del 100% en la clasificación de los nodos hojas, resulta útil para sugerencias.

Una mejora necesaria del modelo no yace en el mismo modelo, sino en la colección de datos. Se plantea el realizar las encuestas nuevamente. En esta ocasión, se pretende realizar un mayor número de encuestas para aumentar el número de registros disponibles para el entrenamiento y validación. Así mismo, se busca una mejor distribución más uniforme entorno a la cantidad de estudiantes por carrera, para evitar así el sesgo ocasionado por la ingeniería en computación en este caso. Otro punto necesario, es volver obligatoria la necesidad de ingresar las 5 materias y 5 habilidades que consideren pertinente los encuestados. Esto para enriquecer para bien los modelos aquí planteados.

Este tipo de trabajo podría ser planteado ante institutos más grandes como la Facultad de Ingeniería, donde haciendo uso de las bases de datos, se puede solicitar a los estudiantes de los últimos 4 semestres (que ya cursaron una buena parte de la carrera y que tienen mayor probabilidad de concluirla) para asegurar los puntos anteriormente indicados. Así pues, la gran cantidad de carreras disponibles no son referentes a un único instituto como la Facultad de Ingeniería, sino también aborda otros como Facultad de Ciencias, de Ciencias Políticas, Contaduría, entre otras. De ahí que la nueva propuesta sea planteada a una escala universitaria completa, lo que enriquecería notoriamente el modelo y proveería mejores resultados.

COMPARATIVO DEL PLAN (REAL VS. ESTIMADO)

Se consideraron de 10 etapas. Cada una necesaria para los requisitos planteados para el proyecto, así como el correcto desarrollo de este. Se planteó entonces una estimación inicial de tiempo que permitiera coordinar el desarrollo de estos. La estimación inicial es la siguiente:

Actividad	Inicio	Fin	14-ago-23	22-sep-23	23-sep-23	07-oct-23	08-oct-23	11-oct-23	12-oct-23	28-oct-23	29-oct-23	30-oct-23	31-oct-23	11-nov-23	12-nov-23	23-nov-23	24-nov-23	27-nov-23	28-nov-23	29-nov-23
Selección de colección de datos	14-ago-23	22-sep-23																		
Exploración y caracterización de datos	23-sep-23	07-oct-23																		
Definición de catálogos de Materias y Habilidades	08-oct-23	11-oct-23																		
Homogenización de datos (Materias, Habilidades, Carreras, etc.)	12-oct-23	28-oct-23																		
Avance de Proyecto (Presentación)	29-oct-23	30-oct-23																		
Estadística de la información (Información transformada)	31-oct-23	11-nov-23																		
Modelos Descriptivos	12-nov-23	23-nov-23																		
Modelos Predictivos	12-nov-23	23-nov-23																		
Documentación escrita (reporte estadístico y reporte de resultados)	24-nov-23	27-nov-23																		
Presentación Final	28-nov-23	29-nov-23																		

Cada fase fue considerada importante en el desarrollo del proyecto. La selección de colección de datos fue adecuada pues nos permitió optar por aquella que nos representara un mayor sentido al momento de análisis, refiriéndonos no solo al interés en el tema sino también posibles análisis (sin entrar a detalle de los datos) que podríamos obtener de cada muestra. La exploración y caracterización de datos se presentó en el presente reporte escrito, donde se dio una exploración inicial de la información para comprender el estado de los datos para entender las transformaciones necesarias a realizar. Ejemplo de ello fue el plantear la homogenización de campos como género, carrera, entre otros. De esto es que entendíamos la necesidad de plantear catálogos con clases bien definidas para los campos. Este proceso, y como se detalló en el reporte, permitió definir sobre qué clase recaerían las distintas variaciones disponibles en los campos afectados. En este punto se plantearon decisiones como el colapsar las materias como cálculo, calculo vectorial, cálculo integral como una única materia “Matemáticas” para una mejor representación de la muestra. Posterior a la definición de catálogos, se planteaba la realización de la homogenización. Esta transformación sería esencial para el desarrollo de modelos tanto descriptivos como predictivos. Respecto a estos últimos, se planteo su desarrollo paralelo puesto que si bien mantienen cierto grado de relación, se buscan enfoques diferentes entre sí. La parte de la documentación escrita engloba el presente reporte y otros desarrollos más técnicos. Tanto el avance de proyecto como la presentación final son fechas marcadas por el profesor, por lo que sirvieron como marco de referencia en la asignación de tiempos.

Ahora bien, al momento del desarrollo del proyecto, si se encontraron variaciones de tiempo. En algunos casos se hubieron atrasos. Sin embargo, en estos se pudieron compensar tiempos o se ejerció el trabajo paralelo que provee el trabajo organizado en equipo. El plan real fue el siguiente:

Actividad	Inicio	Fin	14-ago-23	22-sep-23	23-sep-23	07-oct-23	08-oct-23	11-oct-23	12-oct-23	13-oct-23	28-oct-23	29-oct-23	30-oct-23	02-nov-23	03-oct-23	11-nov-23	12-nov-23	23-nov-23	24-nov-23	27-nov-23	28-nov-23	29-nov-23
Selección de colección de datos	14-ago-23	22-sep-23																				
Exploración y caracterización de datos	23-sep-23	07-oct-23																				
Definición de catálogos de Materias y Habilidades	08-oct-23	12-oct-23																				
Homogenización de datos (Materias, Habilidades, Carreras, etc.)	13-oct-23	02-oct-23																				
Avance de Proyecto (Presentación)	03-oct-23	30-oct-23																				
Estadística de la información (Información transformada)	03-oct-23	11-nov-23																				
Modelos Descriptivos	12-nov-23	24-nov-23																				
Modelos Predictivos	12-nov-23	24-nov-23																				
Documentación escrita (reporte estadístico y reporte de resultados)	24-nov-23	29-nov-23																				
Presentación Final	28-nov-23	29-nov-23																				

En este diagrama se marcó en rojo aquellas marcas de tiempo que sufrieron un atraso. Por ejemplo, en la definición de catálogos se dedicó un día extra a colapsar algunos valores en una misma clase para disminuir las variaciones en materias y habilidades. Este retraso afectó a la iniciación de la homogenización. Sin embargo, esta misma se extendió un total de 5 días. Esto se debió a que encontraron una alta cantidad de variaciones en los valores de habilidades y materias, que pese a tener ya considerados en los catálogos, requirió editar con el widget Edit Domain y verificar en repetidas ocasiones para evitar errores de dominio. Como pudo observarse, tanto el avance de proyecto como la presentación final son fechas inamovibles por los requerimientos del proyecto. La homogenización requirió extenderse por las razones mencionadas, esto generó que la estadística de la información tuviese un retraso de un día. Sin embargo, esta fase pudo cumplirse en la fecha límite planteada originalmente, esto compensó en gran medida todos los anteriores retrasos. El análisis estadístico no fue forzado a reducirse, simplemente se realizó más rápido que las estimaciones iniciales. Por su parte, el desarrollo de modelos tuvo un retraso de un día por correcciones finales que se realizaron con el fin de mejorar el modelo con lo comprendido en el modelo descriptivo. Por su parte, la documentación escrita inició en la fecha planteada. Esto porque resultó conveniente para realizar anotaciones de puntos clave observados en los modelos. Así mismo, se extendió más al momento de desarrollar la documentación. Estos fueron retrasos menores que fueron cubiertos fácilmente para concluir en la fecha límite del proyecto.

PROPUESTA DE NEGOCIO

La elección de carrera es una de las decisiones más importantes que deben tomar los estudiantes que quieren acceder a la educación superior. Esta decisión puede afectar a su futuro académico, profesional y personal, y puede tener consecuencias positivas o negativas según el grado de satisfacción y éxito que logren en su carrera. Sin embargo, muchos estudiantes no cuentan con la información, el asesoramiento o el apoyo adecuados para elegir la carrera que más se adapte a sus intereses, habilidades y expectativas.

Por otro lado, las universidades también se enfrentan al desafío de mejorar su oferta educativa y atraer y retener a los mejores talentos. Para ello, necesitan conocer las preferencias, las necesidades

y el perfil de los estudiantes potenciales, y ofrecerles información personalizada y relevante sobre sus programas académicos.

En este contexto, surge la oportunidad de aplicar la minería de datos, una técnica que permite extraer información valiosa de grandes conjuntos de datos, como los que se pueden recopilar de los estudiantes de una universidad. Con la minería de datos, se pueden descubrir patrones, relaciones y tendencias que pueden ayudar a tomar decisiones o generar conocimientos.

La metodología que se seguirá para desarrollar el proyecto es la siguiente:

- Recopilar los datos de los estudiantes de una universidad mediante una encuesta en línea. Los datos con los que contamos actualmente no son representativos de la población estudiantil en general ya que tenemos un gran sesgo hacia los ingenieros en computación. Debido a esto, realizar una nueva encuesta en línea poniendo más restricciones sobre las habilidades y materias que los alumnos pueden ingresar sería de gran utilidad para un negocio basado en estos algoritmos.
- Preprocesar los datos para eliminar valores faltantes, outliers o inconsistencias, y transformarlos en un formato adecuado para el análisis. De igual manera que en esta ocasión, la ejecución de este proyecto como un negocio involucraría la limpieza de los datos, pero con un formato mas restrictivo como el propuesto anteriormente este trabajo podría ser más fácilmente automatizado.
- Explorar los datos para obtener estadísticas descriptivas, visualizar la distribución y la correlación de las variables, y detectar posibles grupos o segmentos de estudiantes.
- Seleccionar las variables más relevantes para la predicción de la elección de carrera, utilizando técnicas de reducción de dimensionalidad o selección de características. En este caso, por nuestra experiencia consideramos que las materias y las habilidades serían las dimensiones más importantes para considerar.
- Aplicar diferentes algoritmos de minería de datos, como árboles de decisión, regresión logística, redes neuronales o máquinas de vectores de soporte, para crear modelos predictivos de la elección de carrera.
- Evaluar el rendimiento de los modelos, utilizando medidas como la precisión, la sensibilidad, la especificidad o el área bajo la curva ROC, y compararlos entre sí para elegir el mejor. Una de las áreas de mejora del proyecto actual sería realizar una mejor evaluación de los resultados de los modelos.
- Interpretar los resultados de los modelos, identificando las variables más influyentes, las reglas de decisión o los patrones de comportamiento que explican la elección de carrera de los estudiantes.
- Validar los modelos, utilizando datos nuevos o independientes, y verificar su robustez y generalización.

Finalmente, la etapa más importante de comercializar un modelo de este tipo sería crear una interfaz que sea amigable con el usuario. Por las características del proyecto una aplicación web sería lo ideal ya que permitiría tener el mayor alcance sin la necesidad de mantener varias bases de código.

Beneficios para los stakeholders

Los beneficios que se esperan obtener con este proyecto son de dos tipos: económicos y sociales.

Económicamente, se puede generar ingresos por la venta o el uso de los modelos y los análisis, cobrando una tarifa por cada consulta o una suscripción mensual o un pago único. Los clientes potenciales son los estudiantes que están buscando una carrera universitaria, y las universidades que quieren mejorar su oferta educativa.

El mayor cliente del producto serían universidades buscando atraer al mejor talento y retenerlo, ya que los estudiantes con potencial son los que eventualmente aumentaran el prestigio de la institución. Por otro lado, tener menor cantidad de bajas prematuras gracias a una correcta elección de carrera también tiene la potencia de mejorar la reputación de la universidad. En el sector privado esta estrategia es más relevante por la gran competencia del mercado. En el sector público el beneficio es la formación de profesionales mejor preparados y sobre todo más motivados a ejercer sus profesiones.

Socialmente, se puede contribuir al desarrollo educativo y profesional de los estudiantes, ayudándoles a elegir la carrera que más se adapte a sus intereses, habilidades y expectativas, y aumentando su motivación, rendimiento y satisfacción. También se puede contribuir al fortalecimiento de la calidad y la competitividad de las universidades, ofreciéndoles información personalizada y relevante sobre sus programas académicos, y facilitando su captación y fidelización de los mejores talentos. La organización con un modelo como este tendría una gran forma de promoverse públicamente como promotor del progreso social.

Inversiones

Para lanzar un proyecto como este se requerirían de varios recursos mínimos como los siguientes:

- Datos: Un conjunto de datos de mejor calidad sería necesario para entrenar modelos de calidad comercial. El principal problema con el conjunto actual es su poca variedad seguido por su pequeño tamaño. Ambos problemas deberían ser resueltos. Algo que cabe destacar es que no creemos llegar a un dataset que clasifique como BigData.
- Software: Para una solución comercial se debería preparar el software. Por un lado, no se podría continuar utilizando el entorno gráfico de Orange para realizar los cálculos, sino que probablemente sea mejor opción utilizar la biblioteca ofrecida por el mismo proyecto para Python. Esto lo que los permitiría sería realizar una mejor conexión con el resto de los componentes del sistema. Por otro lado, sería necesario crear un sitio web capaz de servir a la cantidad de usuario que decidan probar el software.
- Hardware: Debido a que creemos que no sería necesario un data set con características de BigData (Al menos en la primera iteración del proyecto) hardware modesto debería ser suficiente para entrenar los modelos y servir a los clientes. Soluciones escalables en la nube podrían ser una opción económica para lanzar el producto.
- Personal: Junto con los datos, la inversión en personal capacitado para crear y mantener el producto serían las inversiones más grandes.

Aspectos legales

Para llevar a cabo este proyecto, se deben tener en cuenta algunos aspectos legales importantes, como la protección de los datos personales de los estudiantes, y la propiedad intelectual de los modelos y los resultados.

La protección de los datos personales de los estudiantes implica respetar su privacidad y su confidencialidad, y contar con su consentimiento para usar sus datos con fines comerciales. También implica cumplir con la normativa vigente sobre protección de datos, como la Ley Federal de Protección de Datos Personales en Posesión de los Particulares (LFPDPPP) de México.

La propiedad intelectual de los modelos y los resultados implica definir quién es el dueño de estos, y cómo se van a proteger de posibles copias o usos indebidos. Para ello, se pueden recurrir a diferentes mecanismos legales, como las patentes, los derechos de autor o las licencias.

Precio de venta.

Colocar un precio de venta para este proyecto es un desafío bastante grande, en un principio planteamos un precio inicial entre $600,000 y $700,000 , calculando los costos de materia prima y mano de obra. Sin embargo, se notó un aspecto fundamental para poder hacer funcional el proyecto. Como vimos, los datos recolectados presentan desafíos significativos, como distribuciones desiguales y posibles inconsistencias.

Como se mencionó en la parte del modelo predictivo, la información es tan corta que nuestro modelo aún no está tan completo, por lo que sería necesario hacer una retroalimentación muy grande a la encuesta. La retroalimentación requerirá un análisis exhaustivo y la aplicación de técnicas avanzadas para corregir y homogeneizar la información, por ejemplo, cambiar las preguntas abiertas a preguntas nominales, creando un catálogo de respuestas posibles, haciendo obligatorio llenar las calificaciones que se les puede dará las materias y habilidades, además de enfocar a los grupos de estudiantes que se solicite las preguntas, teniendo una mayor colección de registros para que de esta manera nuestro modelo sea mucho más eficiente.

El equipo de trabajo, compuesto por cuatro personas altamente capacitadas en minería de datos, dedicará tiempo y esfuerzos considerables para abordar los retos específicos del proyecto. La experiencia y habilidades del equipo justifican una tarifa competitiva. Además, el proyecto tiene el potencial de generar información valiosa para las universidades, influyendo positivamente en la toma de decisiones estratégicas. Este impacto directo en el rendimiento y la eficiencia institucional agrega valor al proyecto. Por lo tanto, el precio que se considera para la venta de este proyecto es de $1,000,000.00 (Un millón de pesos M.N.).

Desglose del Costo de 1 Millón de Pesos:

- Análisis y Ajuste de Datos ($350,000.00): Incluye el tiempo dedicado a la identificación y corrección de problemas en los datos, así como la implementación de técnicas de preprocesamiento.

- Desarrollo de Modelos Específicos para Universidades ($250,000.00): Implica la creación de modelos predictivos y descriptivos adaptados a las características únicas del entorno universitario.
- Mano de Obra y Tiempo del Equipo ($200,000.00): Considera el tiempo y la experiencia del equipo, cuatro profesionales altamente capacitados en minería de datos.
- Asesoramiento y Retroalimentación Continua ($150,000.00): Incluye el soporte continuo para la interpretación de resultados, ajustes adicionales y recomendaciones estratégicas.
- Licencias de Software y Recursos Tecnológicos ($50,000.00): Cubre los costos asociados con el uso de herramientas especializadas y recursos tecnológicos necesarios para la ejecución del proyecto.

La valoración de 1 millón de pesos refleja no solo la complejidad técnica del proyecto, sino también su potencial impacto positivo en el ámbito universitario, respaldado por un equipo especializado y dedicado.

CONCLUSIONES

La minería de datos día a día se convierte en un aspecto fundamental para un gran número de empresas. No solo por obtener opiniones de productos, sino también para mejoras de calidad o campañas publicitarias que generan un mayor efecto sobre sus usuarios. En este proyecto se evaluó un conjunto de encuestas realizadas a estudiantes de distintas carreras, a fin de conocer qué determinó la elección de su carrera y si su elección, a su consideración, fue adecuada (nivel de pasión por su carrera). Un análisis estadístico (sin aplicar ningún algoritmo de aprendizaje) nos permitió comprender mejor la información. Entender la distribución de datos, conocer el sesgo de registros a la ingeniería en computación, entre otros. Así mismo, nos hizo comprender que con la edad de 17 o 18 años muchos de los encuestados demostraron tener mayor presencia de personas que no les apasiona su carrera. Estos valores, si bien son útiles para un primer análisis, quedan cortos en toda la información que nos podría proveer la colección. Con los modelos descriptivos ya no solo conocíamos el estado de los datos, su distribución y posibles valores. Ahora ya se podían encontrar relaciones entre estos. Por ejemplo, la distancia a la universidad jugó un papel importante en la elección de carrera, y nos da sentido lógico, pues viajar más lejos conlleva más horas y recursos invertidos para el transporte. Esta mayor inversión se vio reflejado en puntos a los que los encuestados daban más importancia, como el prestigio de la carrera, el sueldo que recibirían, y otros.

Esta información por sí misma es útil para tomar decisiones a corto plazo. Por ejemplo, si queremos incentivar a una mejor elección de carrera para personas que viven más lejos de la universidad se podría optar por apoyos económicos o incluso viviendas para estudiantes a modo que la distancia no juegue un papel tan importante al momento de la elección. Como esta, los modelos descriptivos permitieron encontrar relaciones ocultas a simple vista entre los datos. Tan solo esto ya nos provee una información más enriquecedora que la observada en la exploración inicial.

Por otra parte, los modelos predictivos no solo encuentran relaciones entre campos, sino que también permiten definir bien estas relaciones a modo que según los valores de este campo se

puedan determinar clases adecuadas que sean descritas por los valores dados. En nuestros modelos de materias o habilidades, es precisamente lo que se planteaba con el árbol de decisión. Las calificaciones que daban cada encuestado a la materia o habilidad permitía aproximar más al encuestado a una clase (carrera) dada. Este modelo provee no solo un modelo que utilice algoritmos de inteligencia artificial, sino que provee un valor útil para la sociedad. Al momento de seleccionar la carrera, los jóvenes pueden estar indecisos en su elección. Incluso con las ferias de carreras (presentación de carreras) muchas dudas quedan escondidas. Muchas personas pueden dejarse llevar por parámetros como el prestigio de la carrera, cuánto ganarían trabajando, entre otros. Sin embargo, ignoran qué se encontrarán en el desarrollo de esta, decepcionándose en el camino y generando un bajo nivel de pasión por su carrera. Esto afecta en la calidad de profesionistas. Bien dice el dicho "el mejor trabajo es el que te apasiona", si no lo hiciera, la calidad del trabajo será carente y afectará a largo plazo.

Referencias

Orange. (14 de Enero de 2021). *k-Means*. Recuperado el 22 de Octubre de 2023, de https://orangedatamining.com/widget-catalog/unsupervised/kmeans/

González del Real, M. (2020). Regresión lineal simple parte 2 (U1A14). Recuperado el 3 de noviembre de 2023, de https://rpubs.com/Marijose/U1A14

Pacheco Ortiz, J., Rodriguez Mazahua, L., & Alor Hernández, G. (2023). Reglas de asociación como estrategia de selección de ítems en exámenes adaptativos computarizados. *Ciencia Ergo Sum, 30*(2), 1-9. Obtenido de https://doaj.org/article/57041d2ec3e64b10a74f112519fede33