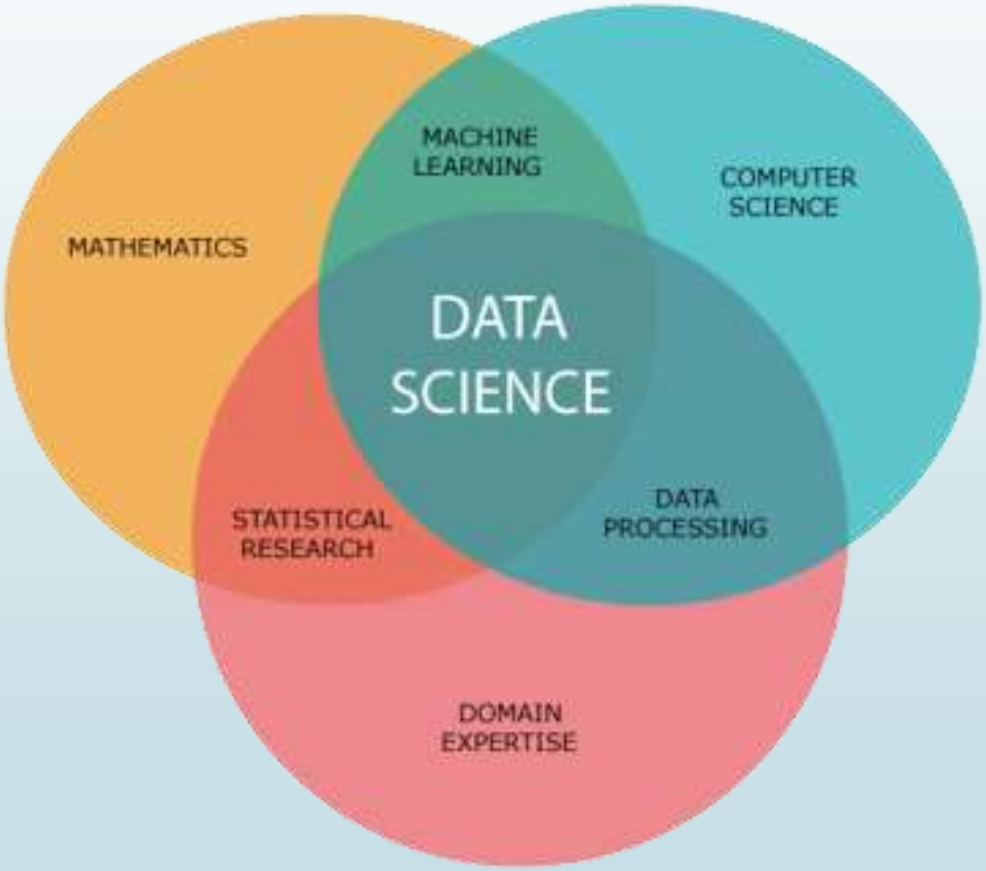
PROYECTO FINAL CODER HOUSE
DATA SCIENCE 2023



Guillermo Maiztegui Marcó

Problema de negocio

- Lograr segmentar a los clientes del ecommerce Amazon de acuerdo al género (Masculino - Femenino), con el fin de poder realizar análisis exhaustivos a ambos grupos sobre sus comportamientos durante 2020-2021

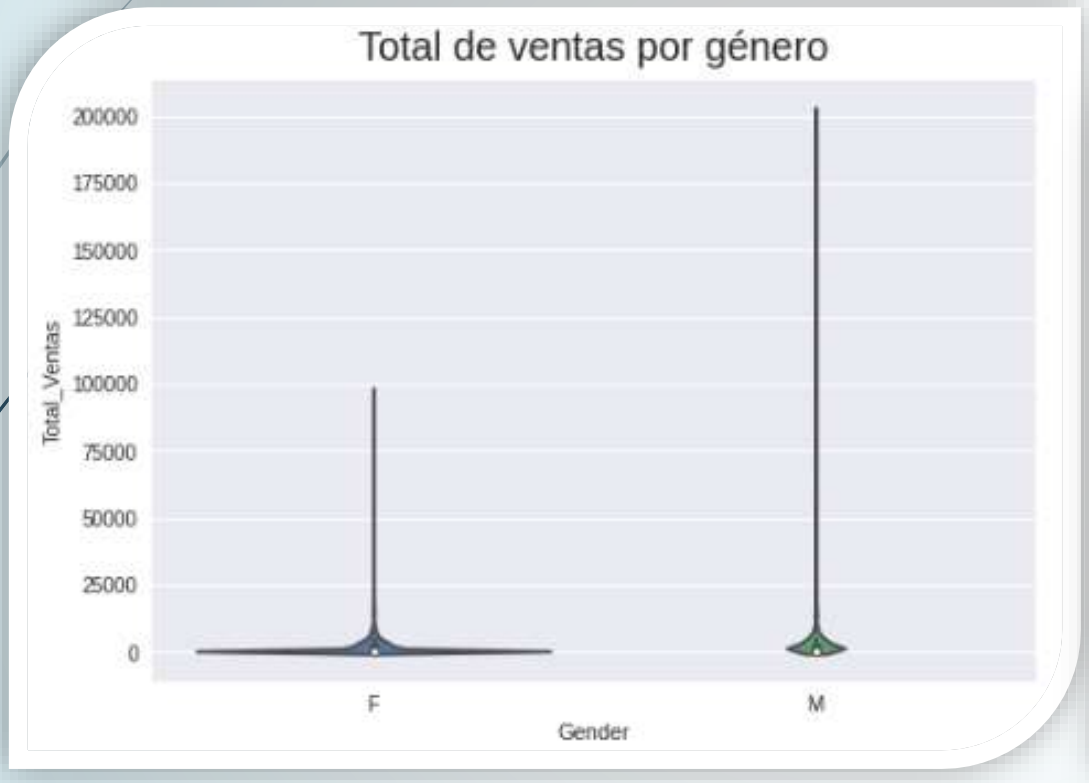
- Mejorar la toma de decisiones teniendo en cuenta dicha segmentación, tanto en en marketing como en lo operativo

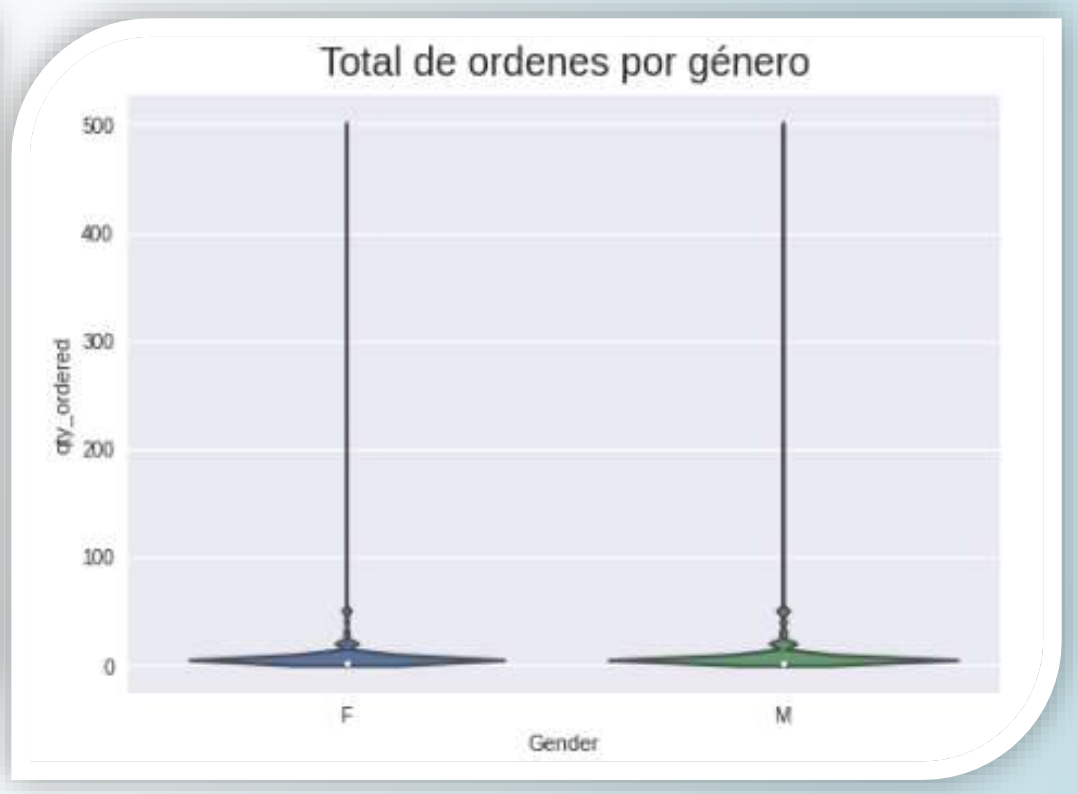
- Automatizar la plataforma lo más posible gracias a las predicciones del modelo de Machine Learning aplicado

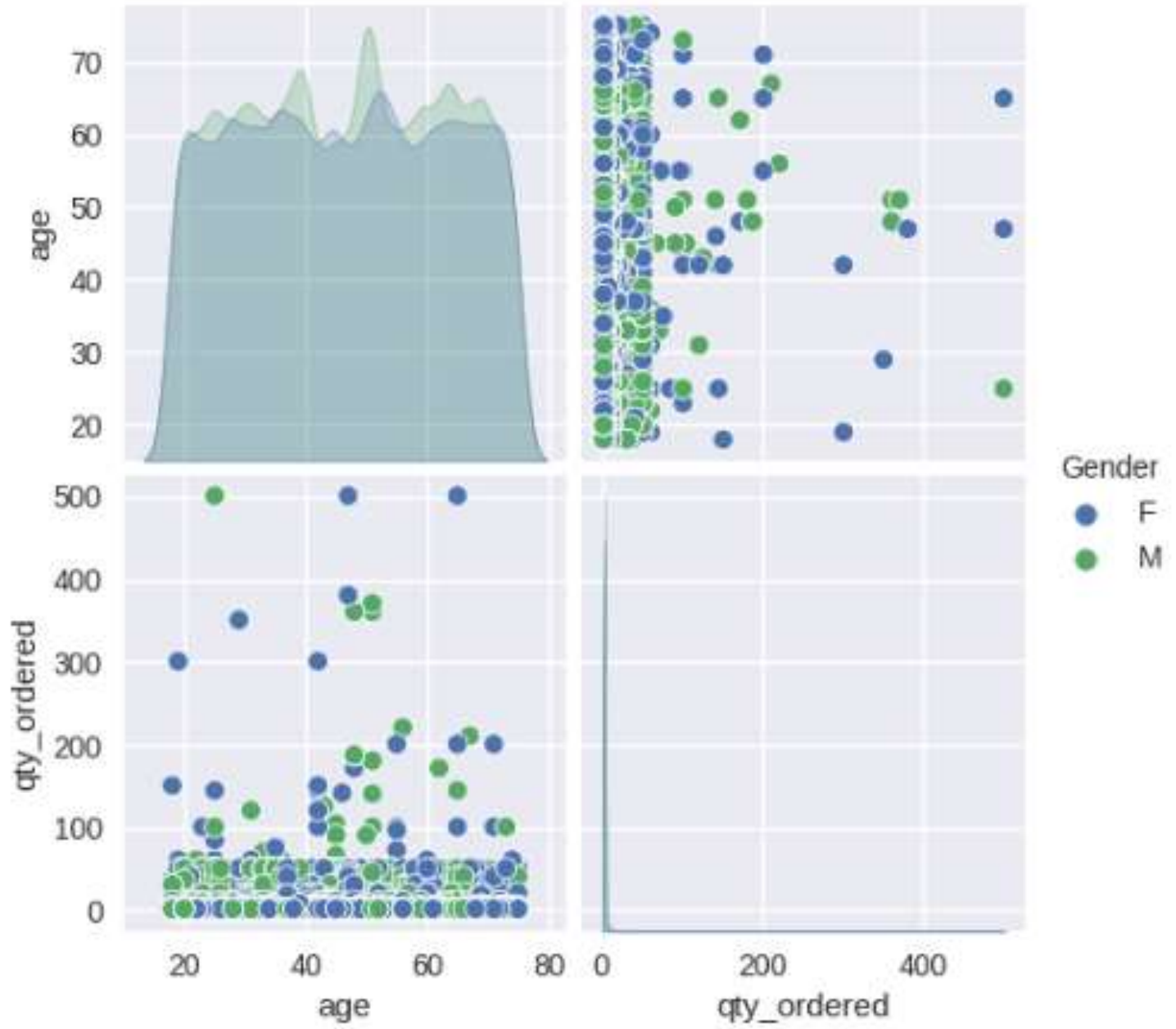
Análisis descriptivo

- Los hombres generan el doble en ventas que las mujeres. Sin embargo, la cantidad de pedidos que realizan son los mismos
- La cantidad de ventas en 2020 con respecto a los géneros fue prácticamente igual, levemente superior en las mujeres. Por otro lado, el 2021 subieron las ventas en general, y además las mujeres agrandaron esa brecha
- Men's & Women's Fashion se encuentran entre las categorías más consumidas
- El promedio de edad en cada género es de 46 años
- En cuanto a la edad en cada género, los hombre jóvenes realizan mayores compras que los mayores, mientras que en las mujeres ocurre lo opuesto

Visualizaciones relevantes






70
60
50
40
30
20
age
500
400
300
200
100
0
qty_ordered
20
40
60
80
age
0
200
400
qty_ordered
Gender
F
M

Variables utilizadas en el algoritmo

- Variable target (categórica) → 'Gender'
- Variables independientes → 'Qty_ordered'
  → 'Price'
  → 'Age'
- Otras variables tenidas en cuenta → 'Category'
  → 'State'
  → 'Date'
  → 'Payment Method'

Algoritmo utilizado

- Luego de probar diversos modelos, además del uso de técnicas y métodos complementarios a estos, se ha optado por el algoritmo supervisado de clasificación Decision Tree Classifier (árboles de decisión)

- Cabe destacar que, utilizando dicho algoritmo como base (estimador), mediante la aplicación del método de ensamble Bagging se obtuvieron las mejores métricas. Este caso puede representarse de la siguiente manera → Bagging + Decision Tree = Random Forest

- La métrica elegida para las observaciones fue la de Accuracy

Obtención de métricas más altas

```
[ ] gs.best_estimator_
```

▾ DecisionTreeClassifier
DecisionTreeClassifier()

❖ Overfitting (sobreajuste)

```
Decision Tree accuracy: 0.6931336091761379
Bagging accuracy: 0.7030849002252134
Cross-Validation in train set: 0.9248645864746574
Cross-Validation in Validation set: 0.6906766639238673
```

Posibles soluciones al Overfitting (alta varianza)

- Elegir un modelo más simple o "forzarlo" → Regularización
- Probar reducir o intercambiar los features
- Agregar datos (pocos datos en el conjunto de prueba "test")
- Evaluar desbalanceo

Conclusiones

- Gracias al Análisis Exploratorio de Datos se ha podido trabajar con el Dataset de manera clara y precisa
- Lograr la segmentación de clientes es una gran ventaja para el Business Intelligence, y esto es posible mediante la aplicación de Machine Learning
- Gracias al Data Wrangling se logró la obtención de una gran cantidad de detalles sobre los comportamientos de los consumidores
- Las visualizaciones son de gran ayuda a la hora de analizar los datos, compararlos e inferir resultados
- Lo importante es "hacer las preguntas correctas para poder lograr que los datos hablen", de acuerdo a nuestros intereses
- Gracias a las predicciones del modelo, se podrán llevar a cabo acciones concretas para mejorar las métricas del negocio (ventas, campañas publicitarias, retención y fidelidad del cliente, reducir costos y aumentar rentabilidad, etc)