1. Introduction

Project objective

- Detect the probability of customer churn using supervised models.
- Segment customers into **churn** and **non-churn** groups by clustering (K Means).

Problem context

Customer **retention is a crucial** point in any profit-oriented business strategy. Retaining existing customers has a **considerably lower cost** than acquiring new ones, which directly impacts the efficiency and sustainability of the company.

In this specific case, customers at risk of churn represent **USD 2,862,926.9**, i.e. **27% of our total number of users.** This is evidence of the strategic priority of proactively identifying customers at risk of churn before they leave the company.

However, it is not enough to **identify customers at risk of leakage.** It is equally important:

1. **Understand your typology:** Identify the different customer profiles that are more likely to drop out.
2. **Understand the possible causes of churn:** factors such as quality of service, price, competition or others.

With this information, concrete actions can be implemented to **reduce the drop-out rate**, such as:

- Improve our services and the quality of our products.
- Develop specific **loyalty** strategies tailored to the segments identified.

Applied methodology

- Use of classification algorithms (Random Forest, Logistic Regression, XGBoost).
- Choice of a **Voting Classifier** to improve the balance between Recall and Accuracy and to create a more robust model for an unbalanced target.
- Application of **KMeans** to segment customers into groups within the results obtained.

2. Exploratory Data Analysis

Description of data

- **Identification variables:** customerID.
- **Demographic variables:** gender, SeniorCitizen, Partner, Dependents.
- **Service-related variables:** PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.
- **Contractual variables:** Contract, PaperlessBilling, PaymentMethod.
- **Consumption variables:** tenure (meses de suscripción), MonthlyCharges, TotalCharges.

- **Target:** Churn (Sí/No).

Data pre-processing

Elimination of irrelevant variables:

- The customerID column was removed as it did not provide predictive value.

Coding of categorical variables:

- One-Hot Encoding was applied to columns with multiple categories: Gender, MultipleLines, InternetService, OnlineSecurity, OnlineSecurityCopy, DeviceProtection, TechSupport, StreamingTV, StreamingMovies and PaymentMethod.
- The Contract column was initially encoded ordinally, but later transformed with One-Hot Encoding to better capture contract information.

Conversion of binary variables:

- Columns with ‘Yes/No’ values (Partner, Dependents, PhoneService, PaperlessBilling, Churn) were converted to True/False and then to numeric values (1/0).

Transformation of numeric columns:

- TotalCharges: Converted to float with error handling.
- Resulting null values were filled in with the median.

Scaling and cleaning:

- Boolean columns were completely transformed to values 1 and 0 to ensure compatibility with the models.

Visualización inicial

	Feature	Importance
7	TotalCharges	0.182142
6	MonthlyCharges	0.171455
3	tenure	0.157688
21	contract_MtoM	0.064010
10	InternetService_Fiber optic	0.039744
19	PaymentMethod_Electronic check	0.037753
8	gender_Male	0.029074
5	PaperlessBilling	0.027799
12	OnlineSecurity_Yes	0.023841
1	Partner	0.023366
13	OnlineBackup_Yes	0.021972
15	TechSupport_Yes	0.021376
9	MultipleLines_Yes	0.020969
14	DeviceProtection_Yes	0.020382
23	contract_TwoYear	0.020308
0	SeniorCitizen	0.020195
2	Dependents	0.019890
17	StreamingMovies_Yes	0.018411
16	StreamingTV_Yes	0.017725
11	InternetService_No	0.016937
18	PaymentMethod_Credit card (automatic)	0.013818
20	PaymentMethod_Mailed check	0.012432
22	contract_OneYear	0.011480
4	PhoneService	0.007234

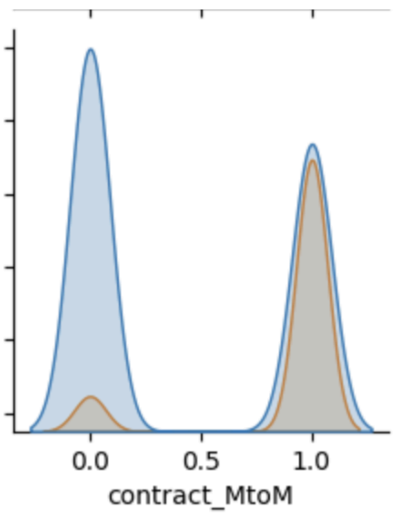
- TotalCharges, MonthlyCharges and tenure are the most relevant characteristics for predicting churn.

- The contract_MtoM variable stands out as a key factor.

The density graph shows the **distribution of customers** with ‘month-to-month’ contracts in relation to churn.

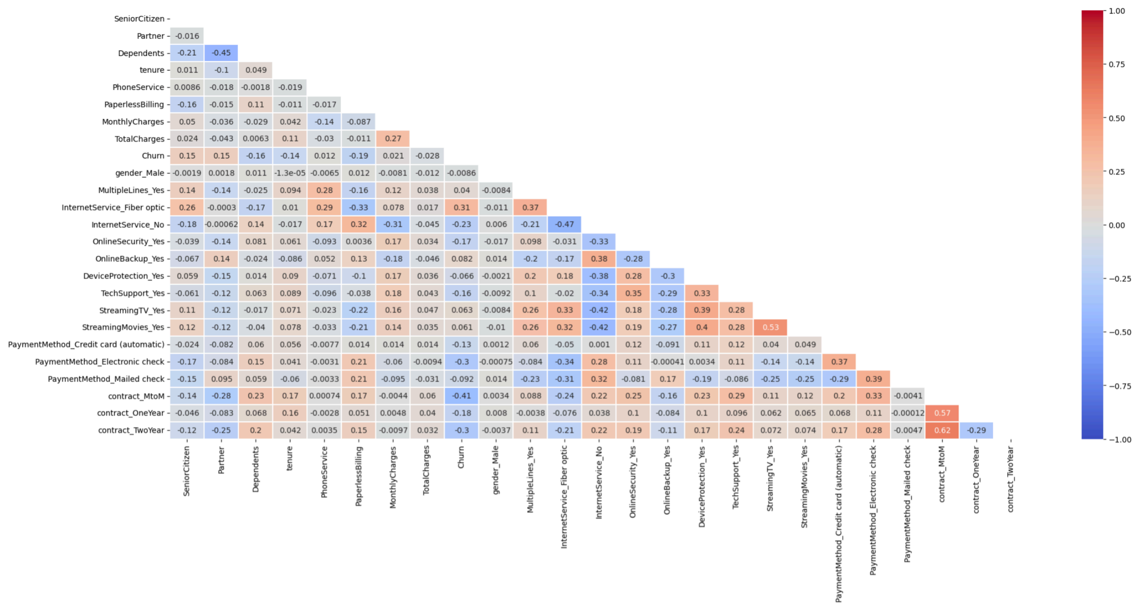
Key remarks:

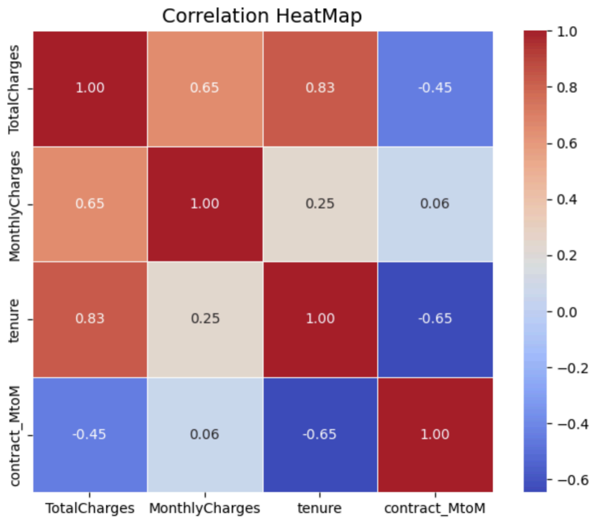
- There is a notable concentration of **customers with positive churn** in monthly contracts.
- This trend reinforces the need to analyse this specific group for retention strategies.



Correlation map
Key remarks:

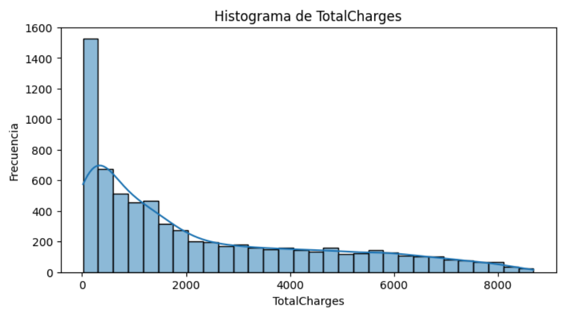
- TotalCharges, MonthlyCharges and tenure show high positive correlation between them, suggesting redundancy.
- The variable Churn shows negative correlations with tenure and positive correlations with contract_MtoM and InternetService_Fiber optic, indicating a possible relationship with **customer churn** on shorter contracts.

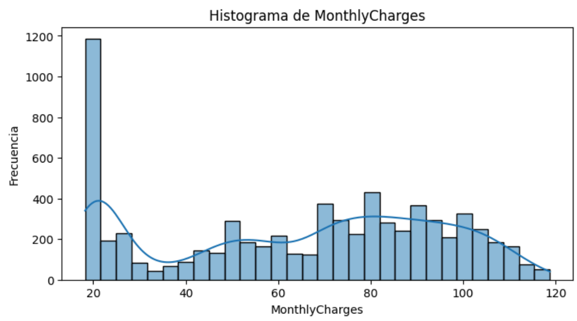


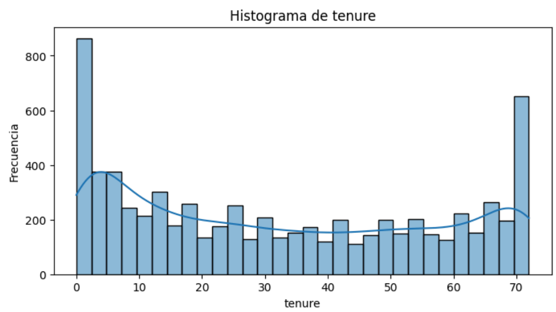


TotalCharges correlates strongly with tenure and **MonthlyCharges,** but negatively with **contract_MtoM,** indicating possible short-term customers with monthly contracts.

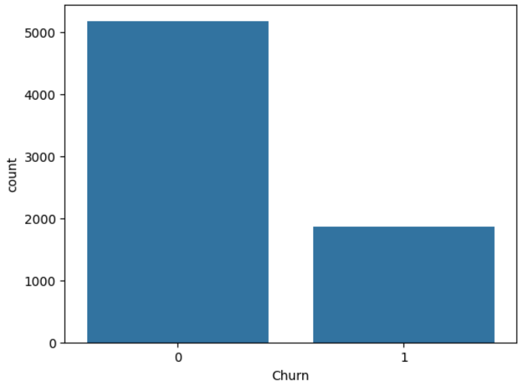
Apart from this, we dismiss the idea of a possible redundancy between these variables.







Due to the skewness and strong leftward bias, a **logarithmic transformation** has been applied to TotalCharges and MonthlyCharges to normalise the data and improve their behaviour in the models.



Most customers **have not churned** (negative class).
This imbalance should be considered when training the models to avoid predictive bias.

3. Construction of the Classification Model (Churn Detection)

Evaluated models

- Random Forest
- Logistic Regression
- XG Boost

Selection of the final model

The main reason was to find a **balance between recall and accuracy**. With **Logistic Regression** a better recall was obtained, achieving a favourable balance based on the F1-Score metric, while with **XGBoost** a **higher accuracy** was achieved. Combining these models using a Voting Classifier allows the strengths of each model to be exploited and overall performance to be improved.

The **Voting Classifier** combines the predictions of several models and selects the final outcome using soft voting probabilities.

Comparativa mejores resultados

Modelo (best)	**Precisión**	**Recall**	**ROC-AUC**
Random Forest(M5)	(N)0.91 (P)0.50	(N)0.71 (P)0.82	0.7623
Logistic Regression(M5)	(N)0.87 (P)0.51	(N)0.77 (P)0.68	0.7620
XGBoost(M2)	(N)0.84 (P)0.67	(N)0.91 (P)0.52	0.7147
Voting Classifier	**(N)0.88 (P)0.59**	**(N)0.82 (P)0.70**	**0.8507**

Final model results with all data

```
Matriz de Confusión:
 [[4349  825]
 [ 554 1315]]

Informe de Clasificación:
              precision    recall  f1-score   support

           0       0.89      0.84      0.86      5174
           1       0.61      0.70      0.66      1869

    accuracy                           0.80      7043
   macro avg       0.75      0.77      0.76      7043
weighted avg       0.81      0.80      0.81      7043


ROC-AUC: 0.8655741149671476
```

4. Customer Segmentation with K Means

Target

- Detect subgroups within the customers classified as **Churn** and **No Churn.**

Methodology

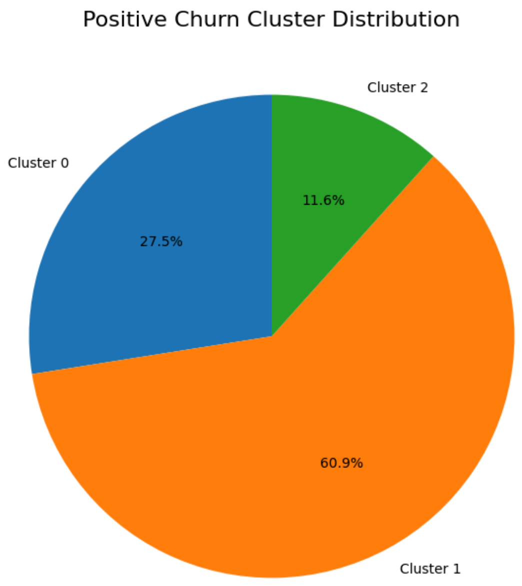
- Application of **K Means** separately in churn and non-churn groups.
- Selection of the optimal number of clusters using the **Elbow method** and the **Silhouette Score**.

Clustering results

Characteristics of customers with POSITIVE churn (who drop out)

Main features:

- Customers with **low seniority** and **few contracted services.**
- High proportion of **monthly contracts** (easy cancellation).
- Predominance of electronic invoicing and **high monthly charges.**
- Low take-up of additional services such as **TechSupport or OnlineSecurity.**
- Increased use of **Electronic Check** payment methods.



Three customer **clusters** were identified with a probability of Churn (positive):

Cluster 0 (27.5%): New customers, with **low monthly charges**, **few** contracted **services** and **monthly contracts.**

- **Cause:** Low perceived value.
- **Strategy:** Basic service offerings with limited promotions.

Cluster 1 (60.9%): Customers with **high charges**, monthly contracts and some additional services.

- **Cause:** High price and lack of customisation.
- **Strategy:** Modular plans and customised discounts.

Cluster 2 (11.6%): Long-standing customers with **high total expenditure** and **multiple contracted services.**

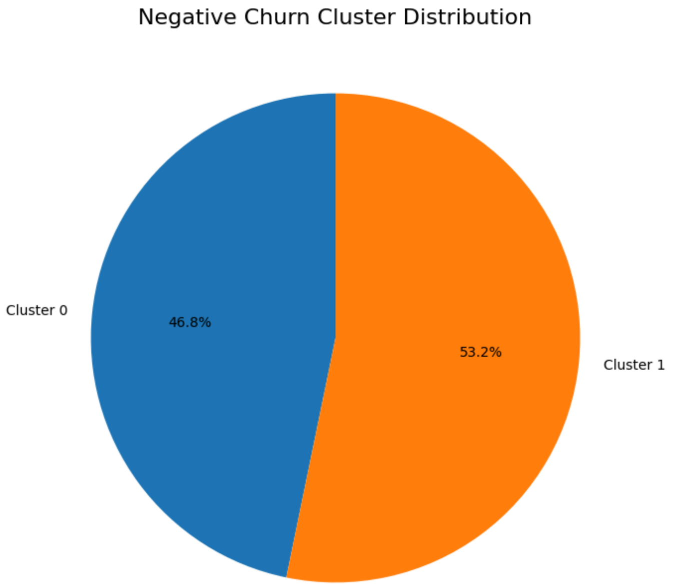
- **Cause:** Lack of active engagement.
- **Strategy:** Loyalty programmes and proactive contact.

Characteristics of customers with NEGATIVE churn (loyal customers)

Main features:

- Increased proportion of customers with **long-term contracts** (annual or biannual).
- High usage of additional services such as OnlineBackup, TechSupport and StreamingTV.
- **Lower monthly charges** and higher satisfaction.
- Predominance of the **Credit Card Automatic** payment method.

- High proportion of clients with **families** or **dependents.**



2 clusters were identified in the customers with a probability of Churn (positive):

- **Cluster 0 (46.8%):** Loyal, **long-standing customers** with multiple contracted services, including Internet and fibre.
  - They are customers with varied contracts (monthly, annual, biannual).

- **Cluster 1 (53.2%): Newer customers** with lower monthly charges, who tend not to sign up for fibre and additional services.
  - Younger/adult segment, less committed to additional services.

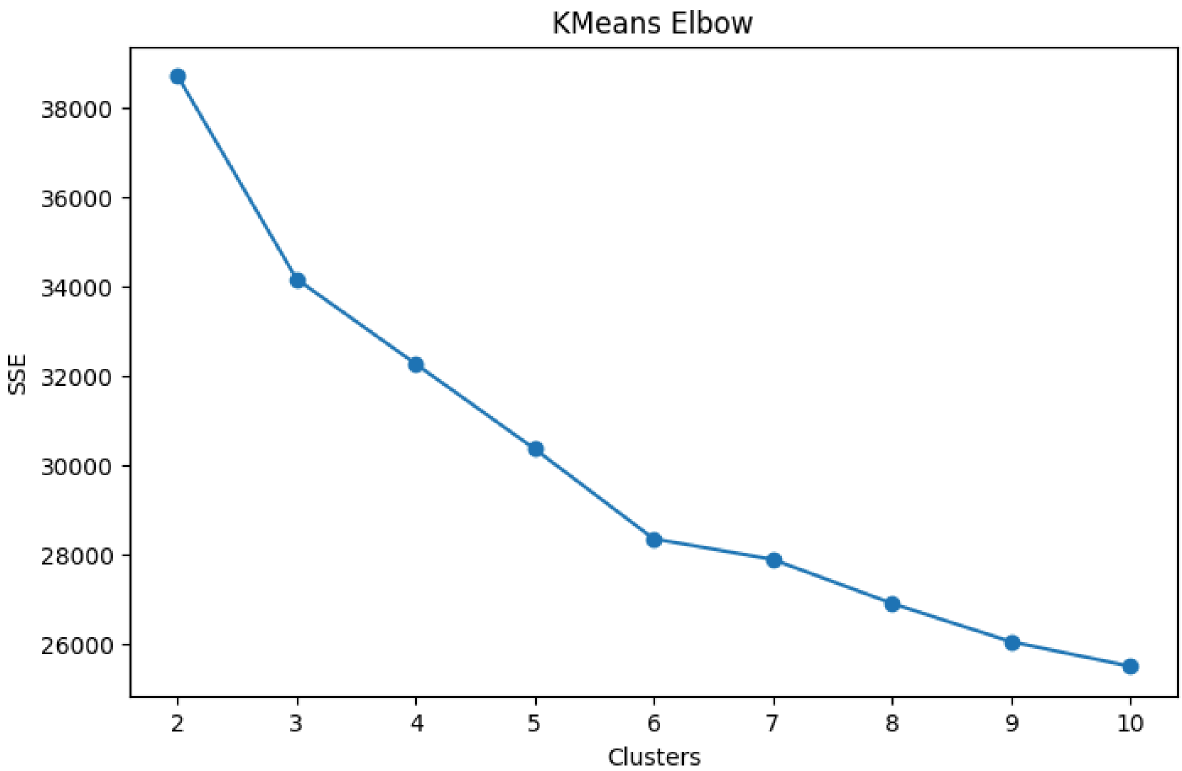
5. Evaluation and Conclusions

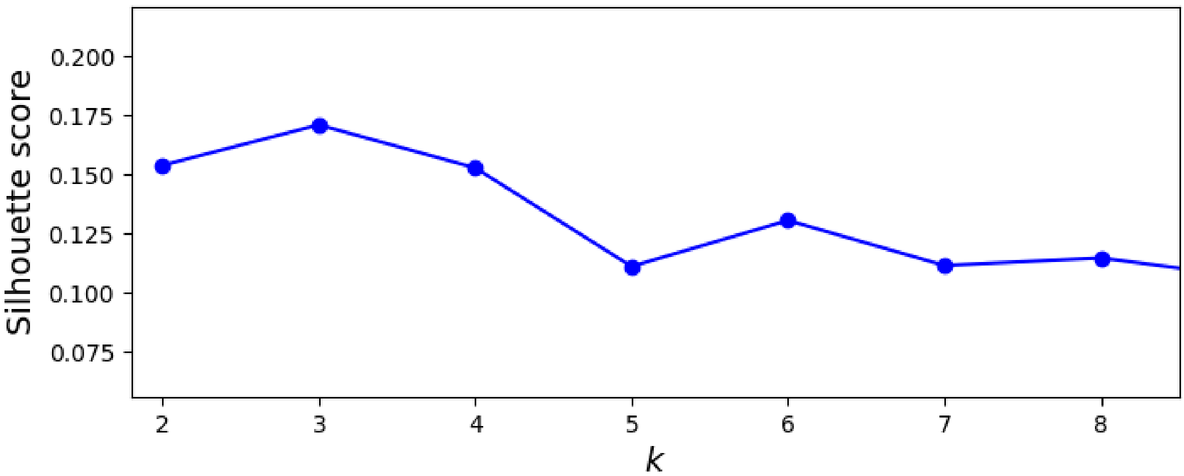
Evaluation of the model

- In tests, the **Voting Classifier** proves to be a robust predictor, achieving a **good balance for generalising to new data.** With a **Train Score of 0.80** and a **Test Score of 0.78,** the model offers confidence even in the face of class imbalance in the target, ensuring consistent performance when incorporating new data.

Segmentation assessment

- **Three clusters** were selected for the **POSITIVE churn** analysis, as they provide a clear and generalisable segmentation; however, the possibility of a fourth cluster will be reviewed in the future to capture additional characteristics.





- For the **NEGATIVE churn, 2 clusters** were selected, as the segmentation has generalised well and provides clear and consistent interpretations. The choice provides confidence in the results and allows an accurate identification of loyal groups.

Final results

- It shows the usefulness of the project:
  - Ability to effectively predict churn.

    - Identification of subgroups that allow customisation of retention strategies.

Key conclusions

Main findings:

- Customers with **monthly contracts**, high charges and few additional services have a higher risk of churn.
- **Loyal customers** tend to have long contracts and multiple services.

Benefits of Voting Classifier:

- Better balance between **recall and precision**, achieving robust performance.

Benefits of KMeans:

- Clear segmentation that allows for strategies tailored to the **specific needs** of each group.