MACHINE LEARNING PROJECT CUSTOMER CLASSIFICATION

Objective of the project:

“Customer retention is a crucial aspect for the survival and growth of any company. In this context, The company is experiencing a customer churn rate that could be detrimental if not controlled. Taking into account the various attributes of customers, such as age, whether they have a spouse, dependents, type of contract, among others, we need to develop a machine learning model that can accurately predict whether a customer will leave the company in the near future. . “This model will help us identify customers at risk of churn and take proactive measures to improve their satisfaction and retention.”

Data description:

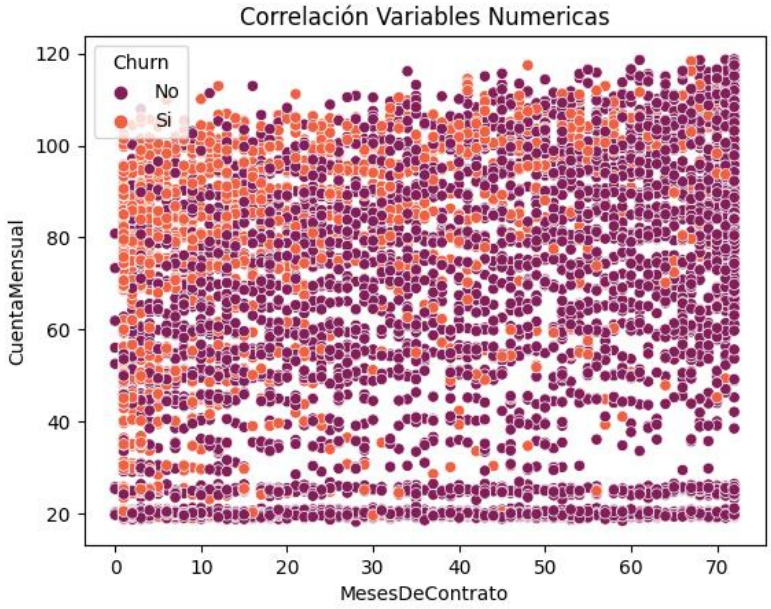
This DataFrame is a set of customer data, from a telecommunications company, with 7043 entries and 18 different characteristics. Here is a description of each column:

1. **Older60Years** : This column indicates if the client is over 60 years old.
2. **Spouse** : This column indicates whether the client has a spouse.
3. **Dependents** : This column indicates whether the client has dependents.
4. **ContractMonths** : This column represents the length of the customer's contract in months.
5. **Landline** : This column indicates whether the client has a landline.
6. **MultipleTelephoneLines** : This column indicates if the client has multiple phone lines.
7. **InternetService** : This column indicates the type of Internet service the customer has.
8. **OnlineSecurity** : This column indicates whether the client has online security.
9. **BackupOnline** : This column indicates whether the client has online backup.
10. **InsuranceOnDevice** : This column indicates whether the customer has insurance on their device.
11. **Technical Support** : This column indicates if the client has technical support.
12. **TVCable** : This column indicates if the client has cable TV.
13. **Streaming** : This column indicates whether the client has streaming services .
14. **ContractType** : This column indicates the type of contract that the client has.
15. **PagoOnline** : This column indicates whether the customer makes payments online.
16. **Payment Method** : This column indicates the customer's payment method.
17. **MonthlyAccount** : This column indicates the amount of the client's monthly account.
18. **Churn** : This column indicates whether the customer has unsubscribed or not.

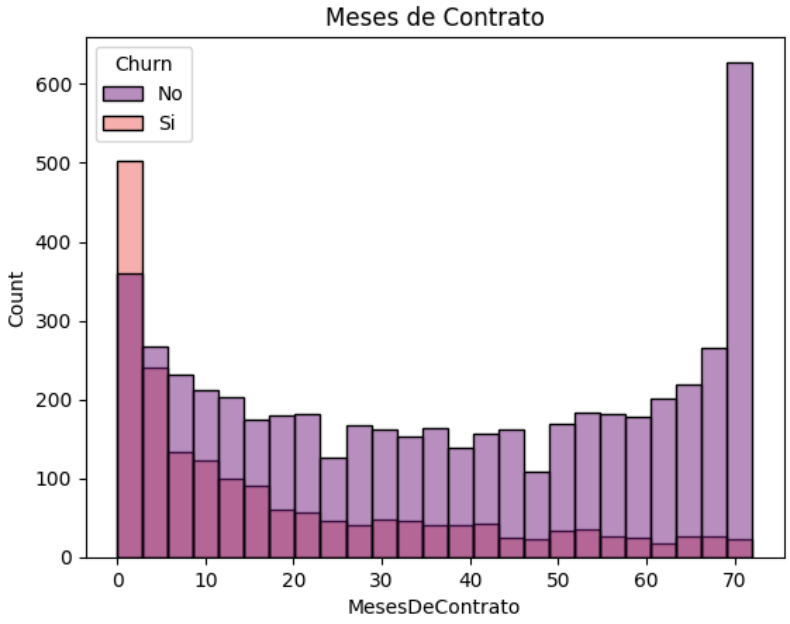
Exploratory analysis

The following findings were found in the dataset :

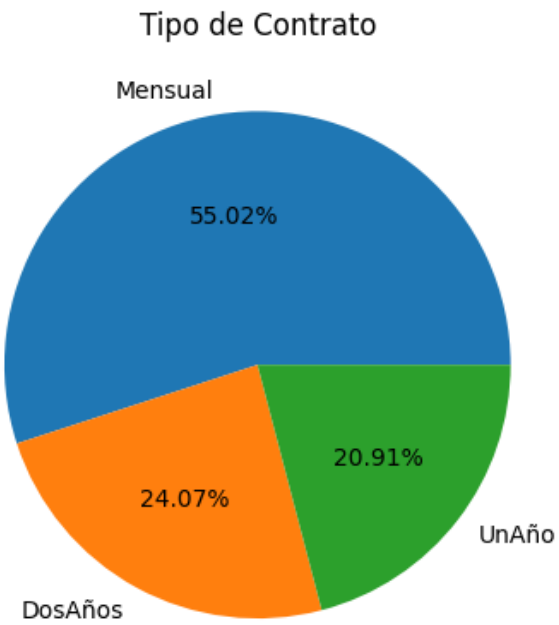
There is no correlation between contract month data and monthly fee.



The withdrawal of the company's clients occurs mostly in the first months.



The type of contract that has the most clients is the monthly contract with a total of 55.02%, followed by the two-year contract with 24.07% and finally the one-year contract with 20.91%.



Data processing

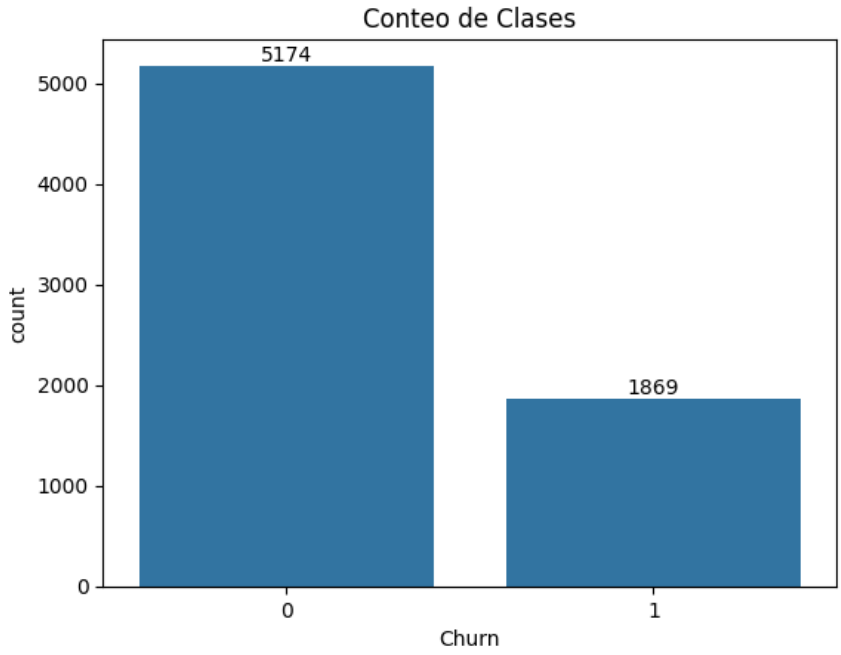
The data was converted to dummies to fit the models, the dataframe was left with these columns:

1. **Conyuge**
2. **Dependientes**
3. **TelefonoFijo**
4. **PagoOnline**
5. **Churn**
6. **Mayor60Años**
7. **MesesDeContrato**
8. **CuentaMensual**
9. **VariasLineasTelefonicas_No**
10. **VariasLineasTelefonicas_Si**
11. **VariasLineasTelefonicas_SinServicioTelefonico**
12. **ServicioDeInternet_DSL**
13. **ServicioDeInternet_FibraOptica**
14. **ServicioDeInternet_No**
15. **SeguridadOnline_No**
16. **SeguridadOnline_Si**
17. **SeguridadOnline_SinServicioDeInternet**
18. **BackupOnline_No**
19. **BackupOnline_Si**
20. **BackupOnline_SinServicioDeInternet**
21. **SeguroEnDispositivo_No**
22. **SeguroEnDispositivo_Si**
23. **SeguroEnDispositivo_SinServicioDeInternet**
24. **SoporteTecnico_No**
25. **SoporteTecnico_Si**
26. **SoporteTecnico_SinServicioDeInternet**

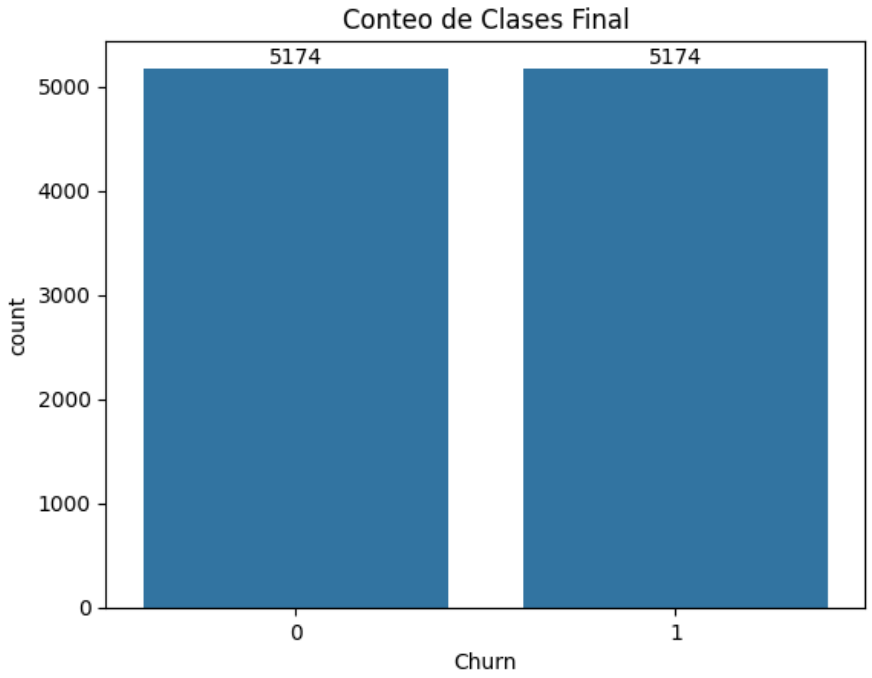
27. **TVCable_No**
28. **TVCable_Si**
29. **TVCable_SinServicioDeInternet**
30. **Streaming_No**
31. **Streaming_Si**
32. **Streaming_SinServicioDeInternet**
33. **TipoDeContrato_DosAños**
34. **TipoDeContrato_Mensual**
35. **TipoDeContrato_UnAño**
36. **FormaDePago_ChequeDigital**
37. **FormaDePago_ChequePapel**
38. **FormaDePago_DebitoEnCuenta**
39. **FormaDePago_TarjetaDeCredito**

Data balancing

The classes had the following distribution:



oversampling technique was used to balance the classes by generating synthetic records using the imblearn library . The distribution of the final classes is as follows:



Classification models

The models chosen for the project are:

- **KNN**
- **Naive Bayes Bernoulli**
- **Decision tree**

Naive Bayes Bernoulli model, another treatment of the data was carried out from the median to convert them to binary.

Results

- **KNN**

confusion matrix

1241	**328**
251	**1285**

Accuracy : 0.8135

Precision: 0.7966

Recall : 0.8365

- **Naive Bayes Bernoulli**

confusion matrix

1048	**521**
240	**1296**

Accuracy : 0.7549

Precision: 0.7132

Recall : 0.8437

- **decision tree** _

confusion matrix

1243	**326**
281	**1255**

Accuracy : 0.8045

Precision: 0.7938

Recall : 0.8170

Model optimization

The models that were sought to be optimized were KNN and Decision Tree. The RandomizedSearchCV class was used to find the best parameters. The results obtained were.

KNN parameters: {'weights': 'uniform', 'n_neighbors': 9, 'leaf_size': 10, 'algorithm': 'ball_tree'}

KNN Precision: 0.7986

Decision Tree Parameters: {'splitter': 'best', 'min_samples_split': 18, 'min_samples_leaf': 6, 'max_features': 'sqrt', 'max_depth': 12, 'criterion': 'entropy'}

Decision Tree Precision : 0.8069

Final Election

The model with the best performance was the **Decision Tree** and the metric on which it was based to choose the final model was precision since the objective was measure the proportion of positive outcomes predicted for truly know how many clients could leave the company.