# Bank Marketing Classification

## مقدمه

**این داده‌ها به کمپین‌های بازاریابی مستقیم یک مؤسسه بانکی پرتغالی مرتبط است. کمپین‌های بازاریابی بر اساس تماس‌های تلفنی انجام شده‌اند. و مشتری یا محصول بانک (سپرده مدت‌دار بانکی) را دریافت می کرد ('بله') یا رد می کرد('خیر').اغلب برای دسترسی به این موضوع، بیش از یک تماس با یک مشتری لازم بود.**

**لینک دیتاست**
https://archive.ics.uci.edu/dataset/222/bank+marketing

- **تعداد feature :۱۶**
- **نمونه‌ها: ۴۵۲۱۱**
- **آیا missing value وجود دارد؟ خیر**

## اطلاعات بیش تر در مورد feature ها

۱. **سن (age)(numeric)**
۲. **شغل (job)(categorical)**
۳. **وضعیت تاهل (marital)(categorical)**
۴. **تحصیلات (education)(categorical)**
۵. **اعتبار پیش فرض (credit)(binary)**

۶. **میانگین موجودی حساب (numeric)(balance):** میانگین موجودی سالانه حساب یه یورو

۷. **وام مسکن (binary)(housing)**

۸. **وام شخصی (binary)(loan)**

۹. **وسیله تماس (categorical)(contact):** وسیله ارتباطی مثل گوشی، تلفن و ...

۱۰. **روز (numeric)(day) :** آخرین تماس در چه روزی از ماه بوده است

۱۱. **ماه (categorical)(month) :** آخرین تماس در چه ماهی انجام شده است

۱۲.**مدت (numeric)(duration):** مدت زمان آخرین تماس به ثانیه

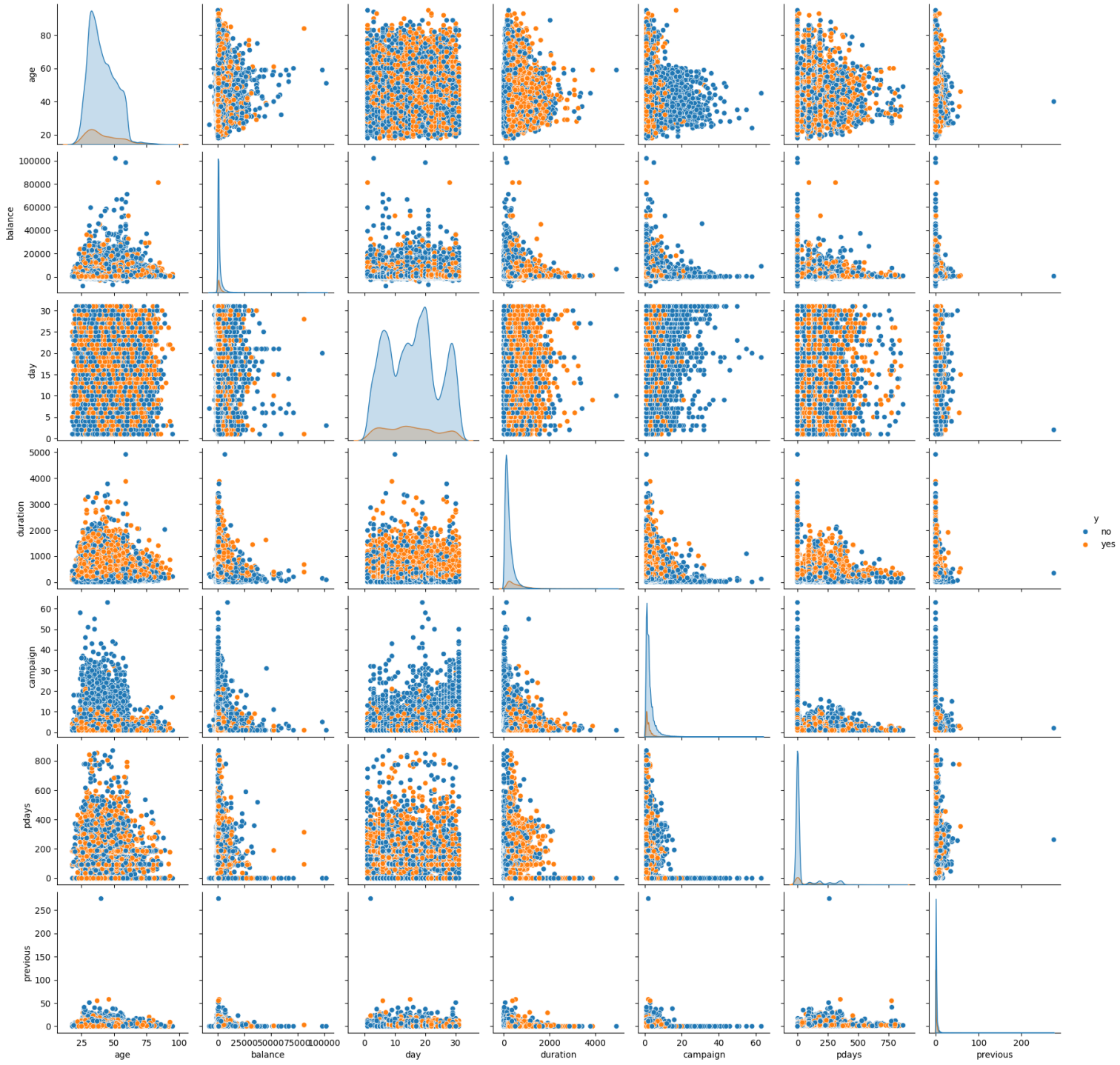
۱۳. **تماس های این کمپین (numeric)(campaign) :** تعداد تماس های انجام شده برای این کمپین و برای این مشتری

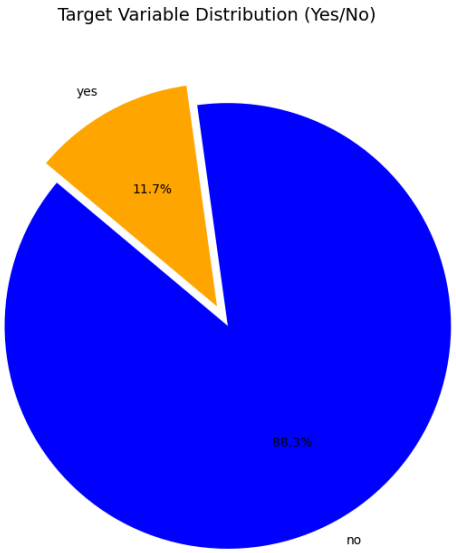
۱۴.**مدت زمان گذشته از آخرین تماس (numeric)(pdays):** تعداد روزهای گذشته از اخرین دفعه ای که با مشتری برای کمپین قبلی تماس گرفته شد

۱۵. **تماس های قبل این کمپین (numeric)(previous) :تعداد تماس های انجام شده با مشتری قبل از این کمپین**

۱۶. **نتیجه کمپین قبل (categorical)(poutcome) :**نتیجه مارکتینگ کمپین قبلی

# Scatter plot





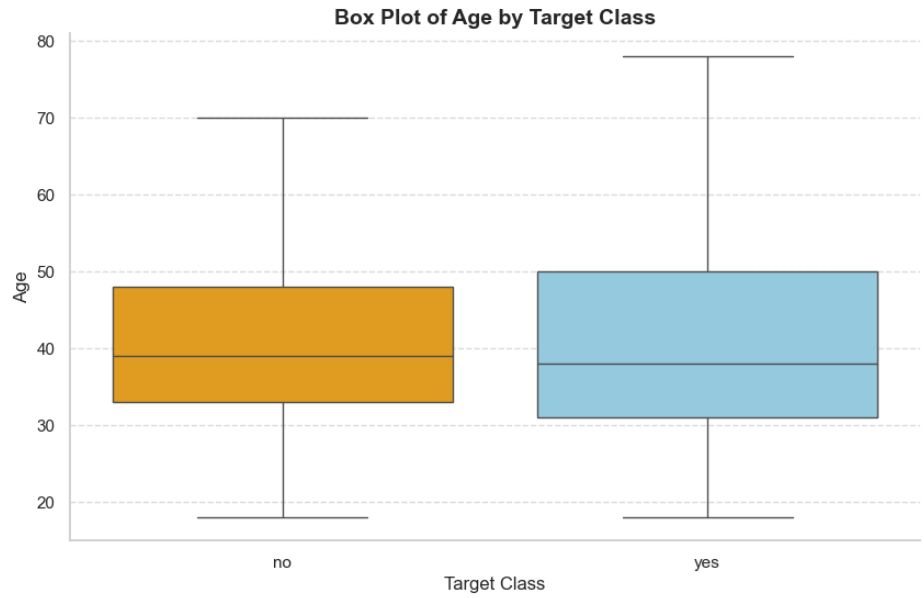

- در بسیاری از نمودارها، نقاط آبی ("نه") بر مناطق خاصی تسلط دارند، که نشان می دهد اکثر مجموعه داده ها ممکن است به سمت نتایج منفی سوگیری داشته باشند و بعد از بررسی داده ها مشخص شد که تقریبا ۸۸% داده ها "نه" هستند و فقط ۱۲% داده ها "بله" هستند که نشان از بالانس نبودن دیتاست دارد.

- به نظر می‌رسد مقادیر طولانی‌تر زمان تماس بیشتر با پاسخ‌های «بله» مرتبط هستند (متغیر هدف = بله). این نشان می دهد که تماس های طولانی تر ممکن است ارتباط مثبتی با احتمال موفقیت داشته باشد.

- در حالی که هر دو پاسخ "بله" و "خیر" در بین میزان موجودی پراکنده هستند، موجودی حساب های بسیار بالا نادر هستند و به نظر می رسد که ارتباطی با دسته "بله" ندارند.

- متغیرهایی مانند balance، duration و pdays ها دارای مقادیر پرت شدید هستند.

# Box plots

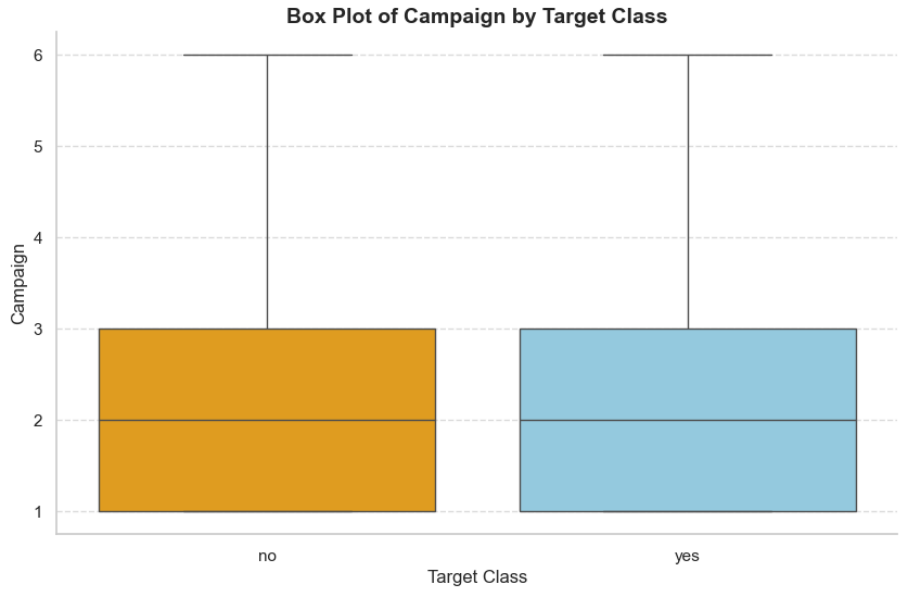



سن: هر دو کلاس "بله" و "خیر" توزیع سنی مشابهی دارند و طیف وسیعی را در بر می گیرند، بدون هیچ نقطه پرت قابل توجهی برای هر یک از کلاس ها. و در نتیجه سن تفاوت مشخصی بین طبقات نشان نمی‌دهد و ممکن است به شدت بر هدف تأثیر نداشته باشد.

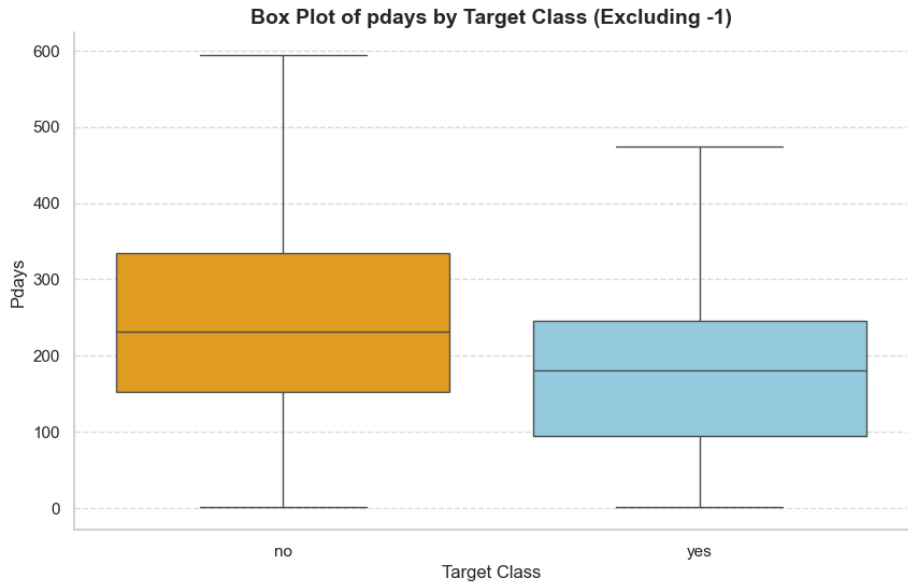

موجودی حساب:توزیع: «نه» در مقایسه با «بله» محدوده بین چارکی (IQR) باریک تری دارد، که نشان دهنده تنوع کمتری در تعادل برای کسانی است که این سرویس را انتخاب نکرده اند و کلاس "بله" به طور کلی توازن کمی بالاتر را نشان می دهد، با اندکی موارد پرت. در نتیجه balance های بالاتر ممکن است کمی با پاسخ "بله" مرتبط باشد.


روز: هر دو کلاس میانه ها و IQR تقریباً یکسان را نشان می دهند، که نشان می دهد تفاوت عمده ای در متغیر روز وجود ندارد. شواهدی مبنی بر تفاوت بین گروه ها وجود ندارد, و احتمالا روز تأثیر کمتری بر متغیر هدف دارد.


مدت تماس: کلاس "بله" در مقایسه با "نه" دارای میانگین مدت زمان بسیار بالاتر و IQR گسترده تر است. گروه "بله" شامل مقادیر بالاتر است و به نظر می رسد که به مدت طولانی تر منحرف شده است. مدت زمان تماس طولانی‌تر به شدت با پاسخ «بله» مرتبط است و این feature را به یک feature مهم تبدیل می‌کند.

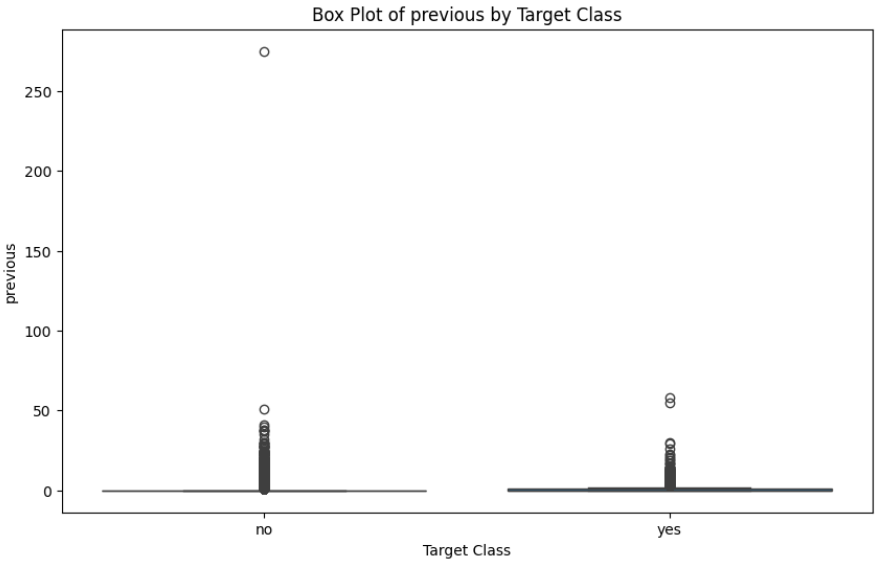



کمپین: هر دو کلاس دارای میانه های مشابه و IQR های باریک هستند. کمپین تأثیر کمتری بر تمایز بین «بله» و «نه» دارد.

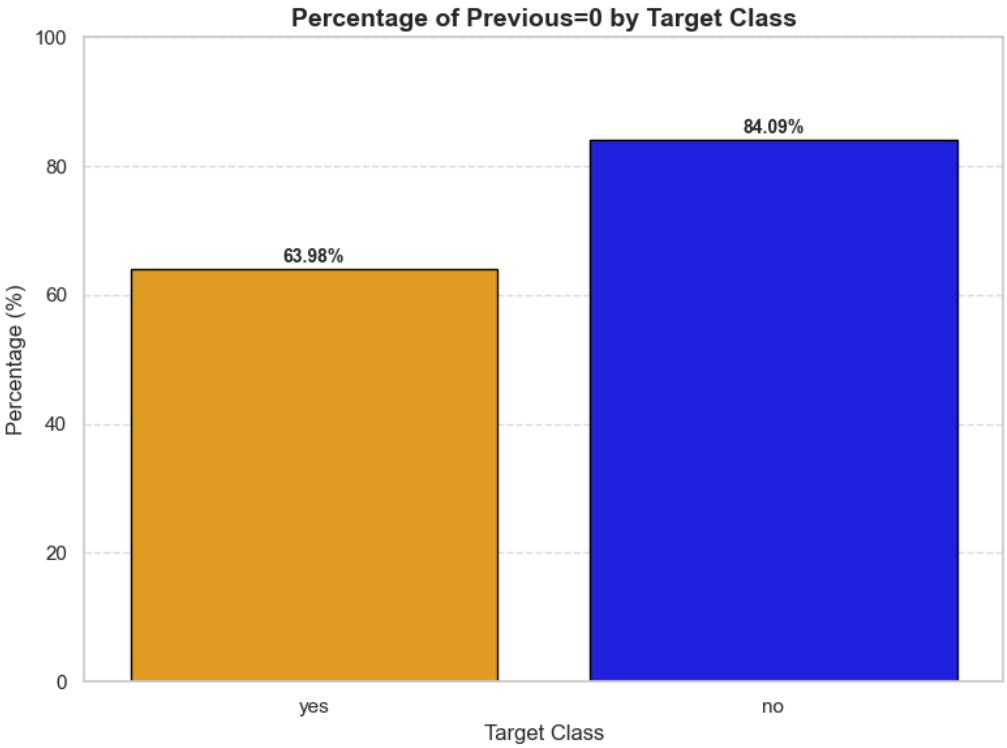



pdays: کلاس‌های «بله» و «خیر» تفاوت‌هایی را نشان می‌دهند، با کلاس «بله» مقادیر pdays بالاتر (در صورت معتبر بودن) و IQR گسترده‌تر.

کلاس "no" به طور کلی مقادیر pdays کمتری دارد، با مقادیر پرت کمتر. تعداد روزهای پس از آخرین تماس کمپین قبلی ممکن است در پیش‌بینی «بله» نقش داشته باشد.

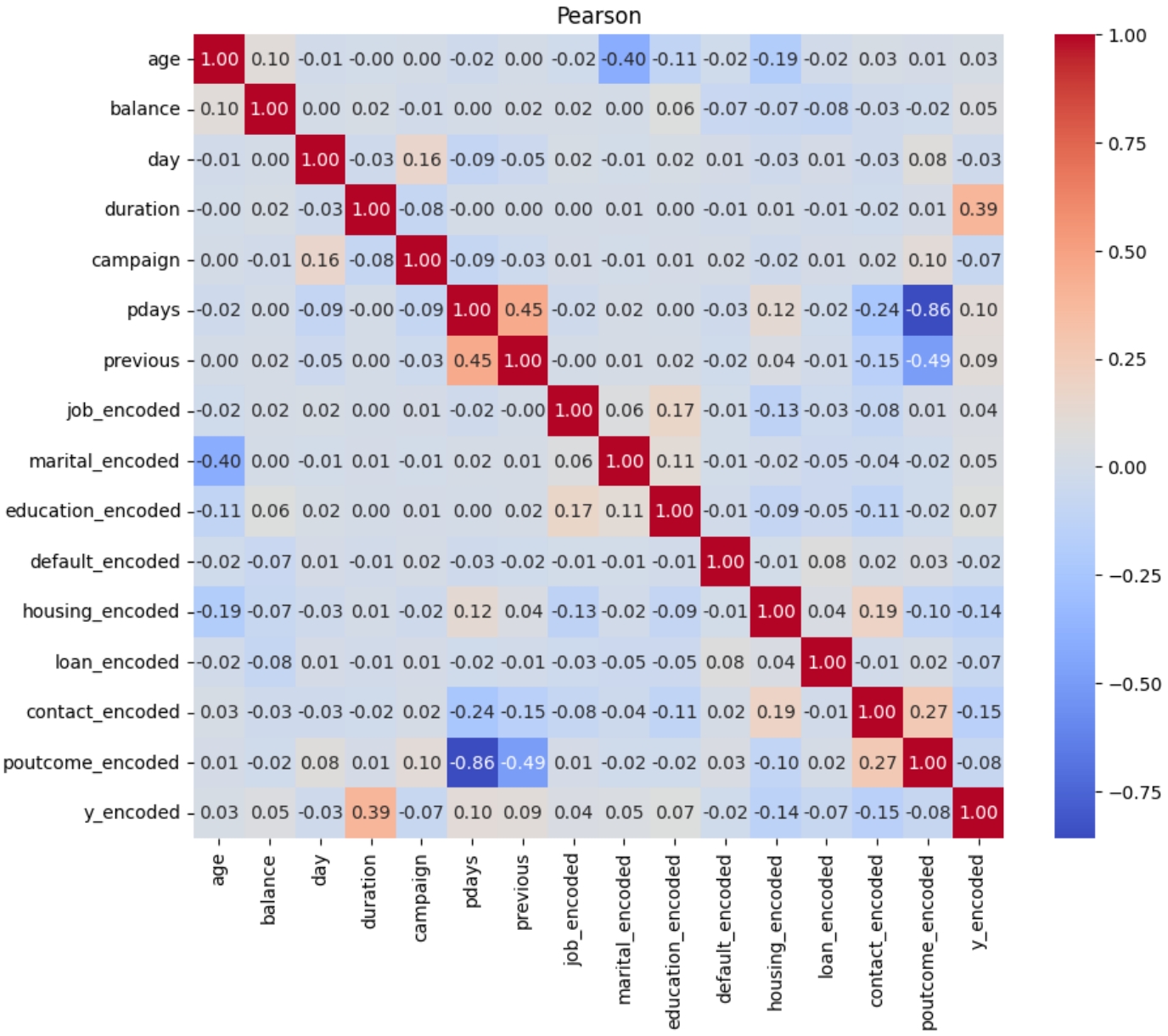


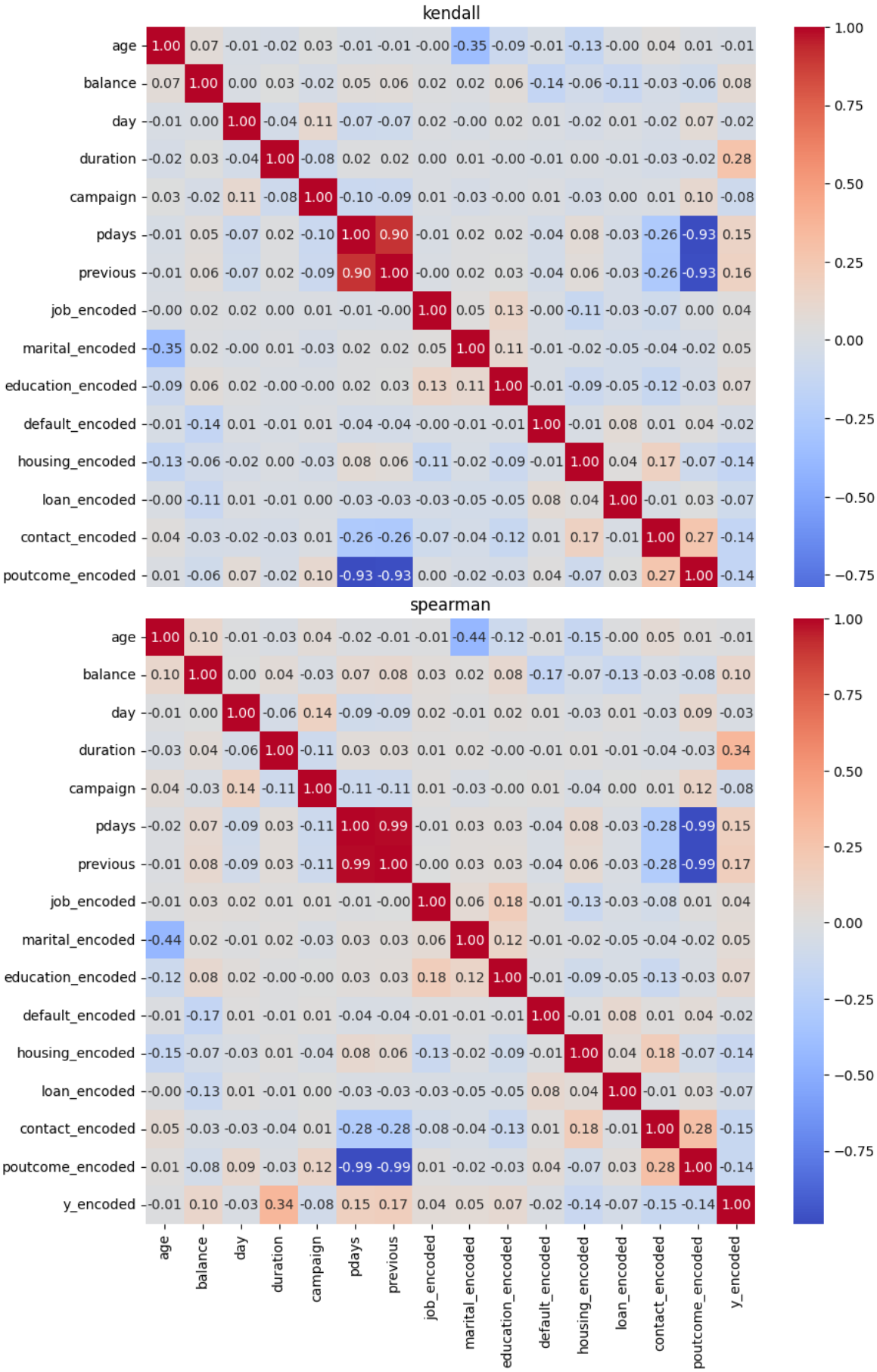

هر دو کلاس هدف (نه و بله) اکثر مقادیر "قبلی"اش ۰ است ، به این معنی که قبلاً به ندرت با مشتریان تماس گرفته می شد. در هر دو کلاس هدف مقادیر پرت قابل توجهی وجود دارد. برای کلاس no، ماکزیمم پرت به بالای ۲۵۰ می رسد که انحراف قابل توجهی از بقیه داده ها است. کلاس yes نیز دارای مقادیر پرت است اما در مقادیر کمتر در مقایسه با کلاس no.




همان طور که مشخص است، در هردو کلاس تعداد بالایی ۰ وجود دارد.

# HeatMap


Pearson
age 1.00 0.10 -0.01 -0.00 0.00 -0.02 0.00 -0.02 -0.40 -0.11 -0.02 -0.19 -0.02 0.03 0.01 0.03
balance 0.10 1.00 0.00 0.02 -0.01 0.00 0.02 0.02 0.00 0.06 -0.07 -0.07 -0.08 -0.03 -0.02 0.05
day -0.01 0.00 1.00 -0.03 0.16 -0.09 -0.05 0.02 -0.01 0.02 0.01 -0.03 0.01 -0.03 0.08 -0.03
duration -0.00 0.02 -0.03 1.00 -0.08 -0.00 0.00 0.00 0.01 0.00 -0.01 0.01 -0.01 -0.02 0.01 0.39
campaign 0.00 -0.01 0.16 -0.08 1.00 -0.09 -0.03 0.01 -0.01 0.01 0.02 -0.02 0.01 0.02 0.10 -0.07
pdays -0.02 0.00 -0.09 -0.00 -0.09 1.00 0.45 -0.02 0.02 0.00 -0.03 0.12 -0.02 -0.24 -0.86 0.10
previous 0.00 0.02 -0.05 0.00 -0.03 0.45 1.00 -0.00 0.01 0.02 -0.02 0.04 -0.01 -0.15 -0.49 0.09
job_encoded -0.02 0.02 0.02 0.00 0.01 -0.02 -0.00 1.00 0.06 0.17 -0.01 -0.13 -0.03 -0.08 0.01 0.04
marital_encoded -0.40 0.00 -0.01 0.01 -0.01 0.02 0.01 0.06 1.00 0.11 -0.01 -0.02 -0.05 -0.04 -0.02 0.05
education_encoded -0.11 0.06 0.02 0.00 0.01 0.00 0.02 0.17 0.11 1.00 -0.01 -0.09 -0.05 -0.11 -0.02 0.07
default_encoded -0.02 -0.07 0.01 -0.01 0.02 -0.03 -0.02 -0.01 -0.01 -0.01 1.00 -0.01 0.08 0.02 0.03 -0.02
housing_encoded -0.19 -0.07 -0.03 0.01 -0.02 0.12 0.04 -0.13 -0.02 -0.09 -0.01 1.00 0.04 0.19 -0.10 -0.14
loan_encoded -0.02 -0.08 0.01 -0.01 0.01 -0.02 -0.01 -0.03 -0.05 -0.05 0.08 0.04 1.00 -0.01 0.02 -0.07
contact_encoded 0.03 -0.03 -0.03 -0.02 0.02 -0.24 -0.15 -0.08 -0.04 -0.11 0.02 0.19 -0.01 1.00 0.27 -0.15
poutcome_encoded 0.01 -0.02 0.08 0.01 0.10 -0.86 -0.49 0.01 -0.02 -0.02 0.03 -0.10 0.02 0.27 1.00 -0.08
y_encoded 0.03 0.05 -0.03 0.39 -0.07 0.10 0.09 0.04 0.05 0.07 -0.02 -0.14 -0.07 -0.15 -0.08 1.00
age balance day duration campaign pdays previous job_encoded marital_encoded education_encoded default_encoded housing_encoded loan_encoded contact_encoded poutcome_encoded y_encoded
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75

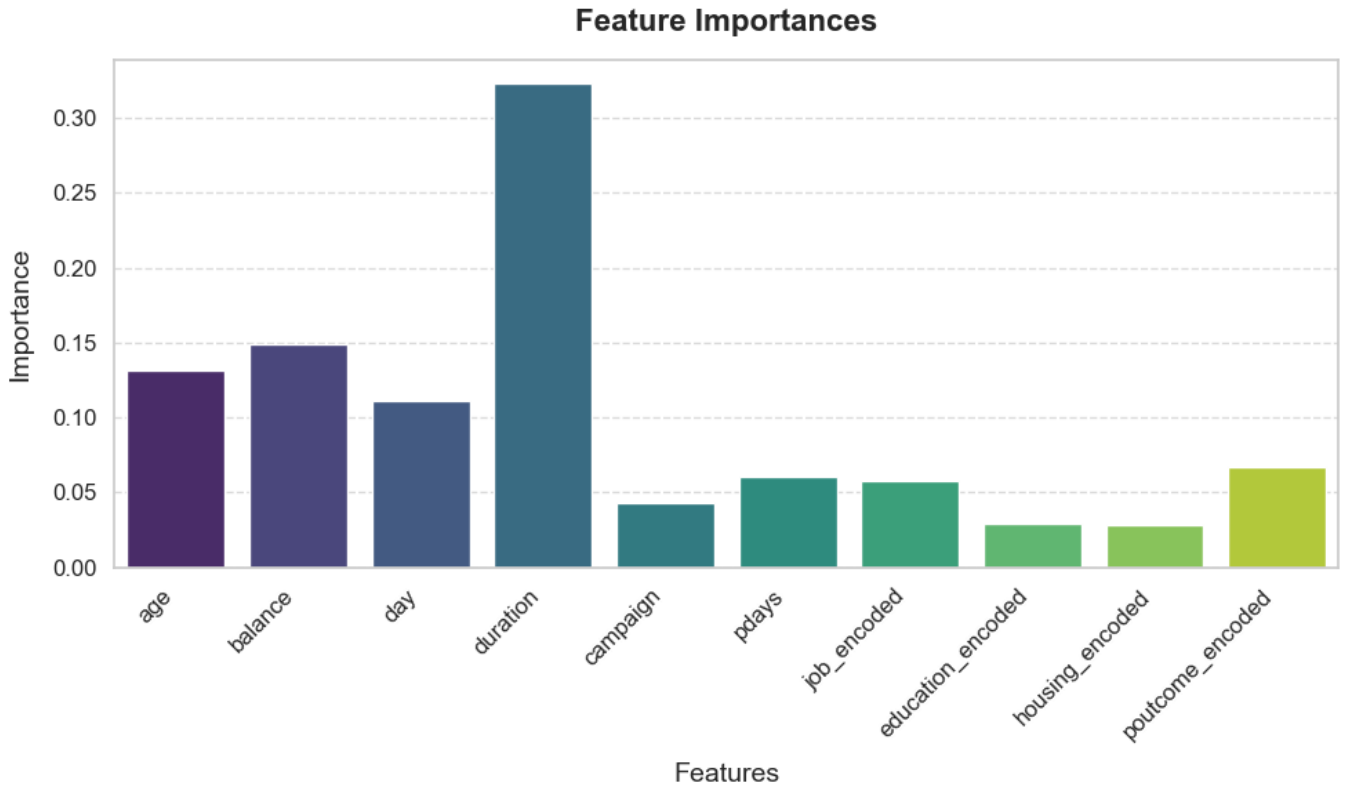

kendall
age 1.00 0.07 -0.01 -0.02 0.03 -0.01 -0.01 -0.00 -0.35 -0.09 -0.01 -0.13 -0.00 0.04 0.01 -0.01
balance 0.07 1.00 0.00 0.03 -0.02 0.05 0.06 0.02 0.02 0.06 -0.14 -0.06 -0.11 -0.03 -0.06 0.08
day -0.01 0.00 1.00 -0.04 0.11 -0.07 -0.07 0.02 -0.00 0.02 0.01 -0.02 0.01 -0.02 0.07 -0.02
duration -0.02 0.03 -0.04 1.00 -0.08 0.02 0.02 0.00 0.01 -0.00 -0.01 0.00 -0.01 -0.03 -0.02 0.28
campaign 0.03 -0.02 0.11 -0.08 1.00 -0.10 -0.09 0.01 -0.03 -0.00 0.01 -0.03 0.00 0.01 0.10 -0.08
pdays -0.01 0.05 -0.07 0.02 -0.10 1.00 0.90 -0.01 0.02 0.02 -0.04 0.08 -0.03 -0.26 -0.93 0.15
previous -0.01 0.06 -0.07 0.02 -0.09 0.90 1.00 -0.00 0.02 0.03 -0.04 0.06 -0.03 -0.26 -0.93 0.16
job_encoded -0.00 0.02 0.02 0.00 0.01 -0.01 -0.00 1.00 0.05 0.13 -0.00 -0.11 -0.03 -0.07 0.00 0.04
marital_encoded -0.35 0.02 -0.00 0.01 -0.03 0.02 0.02 0.05 1.00 0.11 -0.01 -0.02 -0.05 -0.04 -0.02 0.05
education_encoded -0.09 0.06 0.02 -0.00 -0.00 0.02 0.03 0.13 0.11 1.00 -0.01 -0.09 -0.05 -0.12 -0.03 0.07
default_encoded -0.01 -0.14 0.01 -0.01 0.01 -0.04 -0.04 -0.00 -0.01 -0.01 1.00 -0.01 0.08 0.01 0.04 -0.02
housing_encoded -0.13 -0.06 -0.02 0.00 -0.03 0.08 0.06 -0.11 -0.02 -0.09 -0.01 1.00 0.04 0.17 -0.07 -0.14
loan_encoded -0.00 -0.11 0.01 -0.01 0.00 -0.03 -0.03 -0.03 -0.05 -0.05 0.08 0.04 1.00 -0.01 0.03 -0.07
contact_encoded 0.04 -0.03 -0.02 -0.03 0.01 -0.26 -0.26 -0.07 -0.04 -0.12 0.01 0.17 -0.01 1.00 0.27 -0.14
poutcome_encoded 0.01 -0.06 0.07 -0.02 0.10 -0.93 -0.93 0.00 -0.02 -0.03 0.04 -0.07 0.03 0.27 1.00 -0.14
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
spearman
age 1.00 0.10 -0.01 -0.03 0.04 -0.02 -0.01 -0.01 -0.44 -0.12 -0.01 -0.15 -0.00 0.05 0.01 -0.01
balance 0.10 1.00 0.00 0.04 -0.03 0.07 0.08 0.03 0.02 0.08 -0.17 -0.07 -0.13 -0.03 -0.08 0.10
day -0.01 0.00 1.00 -0.06 0.14 -0.09 -0.09 0.02 -0.01 0.02 0.01 -0.03 0.01 -0.03 0.09 -0.03
duration -0.03 0.04 -0.06 1.00 -0.11 0.03 0.03 0.01 0.02 -0.00 -0.01 0.01 -0.01 -0.04 -0.03 0.34
campaign 0.04 -0.03 0.14 -0.11 1.00 -0.11 -0.11 0.01 -0.03 -0.00 0.01 -0.04 0.00 0.01 0.12 -0.08
pdays -0.02 0.07 -0.09 0.03 -0.11 1.00 0.99 -0.01 0.03 0.03 -0.04 0.08 -0.03 -0.28 -0.99 0.15
previous -0.01 0.08 -0.09 0.03 -0.11 0.99 1.00 -0.00 0.03 0.03 -0.04 0.06 -0.03 -0.28 -0.99 0.17
job_encoded -0.01 0.03 0.02 0.01 0.01 -0.01 -0.00 1.00 0.06 0.18 -0.01 -0.13 -0.03 -0.08 0.01 0.04
marital_encoded -0.44 0.02 -0.01 0.02 -0.03 0.03 0.03 0.06 1.00 0.12 -0.01 -0.02 -0.05 -0.04 -0.02 0.05
education_encoded -0.12 0.08 0.02 -0.00 -0.00 0.03 0.03 0.18 0.12 1.00 -0.01 -0.09 -0.05 -0.13 -0.03 0.07
default_encoded -0.01 -0.17 0.01 -0.01 0.01 -0.04 -0.04 -0.01 -0.01 -0.01 1.00 -0.01 0.08 0.01 0.04 -0.02
housing_encoded -0.15 -0.07 -0.03 0.01 -0.04 0.08 0.06 -0.13 -0.02 -0.09 -0.01 1.00 0.04 0.18 -0.07 -0.14
loan_encoded -0.00 -0.13 0.01 -0.01 0.00 -0.03 -0.03 -0.03 -0.05 -0.05 0.08 0.04 1.00 -0.01 0.03 -0.07
contact_encoded 0.05 -0.03 -0.03 -0.04 0.01 -0.28 -0.28 -0.08 -0.04 -0.13 0.01 0.18 -0.01 1.00 0.28 -0.15
poutcome_encoded 0.01 -0.08 0.09 -0.03 0.12 -0.99 -0.99 0.01 -0.02 -0.03 0.04 -0.07 0.03 0.28 1.00 -0.14
y_encoded -0.01 0.10 -0.03 0.34 -0.08 0.15 0.17 0.04 0.05 0.07 -0.02 -0.14 -0.07 -0.15 -0.14 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
age
balance
day
duration
campaign
pdays
previous
job_encoded
marital_encoded
education_encoded
default_encoded
housing_encoded
loan_encoded
contact_encoded
poutcome_encoded
y_encoded

برای ترسیم heatmap علاوه بر pearson از Kendall و spearman هم استفاده شده است، زیرا برخلاف پیرسون، spearman خطی بودن را فرض نمی‌کند، بنابراین روابط را بر اساس rank order می‌گیرد. این باعث می شود آن را نسبت به وابستگی های غیر خطی قوی تر کند.

- متغیر های pdays و previous حتی در تجزیه و تحلیل مبتنی بر rank، به شدت correlation مثبت دارند، که نشان می‌دهد بدون توجه به خطی بودن، رفتار مشابهی دارند.
- poutcome_encoded با دو متغیر pdays و previous همبستگی قوی منفی دارد.
- متغیر duration با همبستگی متوسط نشان می دهد که مدت زمان طولانی تر هنوز با هدف مرتبط است. اسپیرمن . کندال مانند پیرسون، این را به عنوان یک ویژگی کلیدی شناسایی می کند.
- سایر ویژگی‌ها مانند previous و pdays همبستگی‌های ضعیف‌تر، اما همچنان مرتبط را نشان می‌دهند.
- بسیاری از feature ها دارای همبستگی نزدیک به ۰ با متغیر هدف هستند که نشان دهنده روابط ضعیف است (به عنوان مثال، سن، job_encoded، default_encoded).

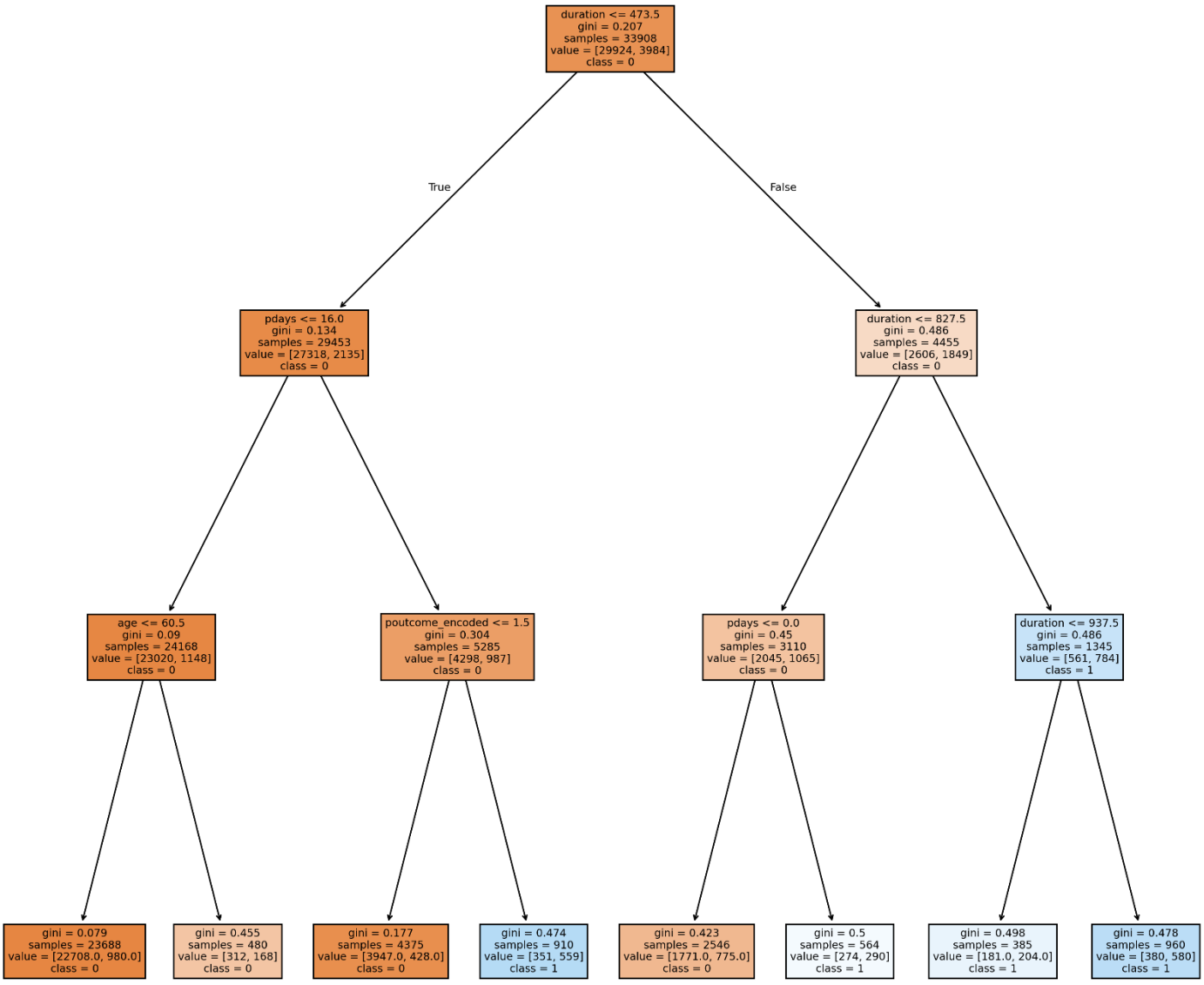
# Feature selection

همان طور که در heatmap دیده شد، تعداد feature های قوی خیلی کم بود و حتی همان feature های خوب هم correlation چندان بالایی نداشتند.برای مثال بهترین feature این دیتاست، duration یا همان مدت تماس می باشد که دارای corr = ۰.۳۹ می باشد. برای همین تلاش کردم تا اثر feature ها را به طور مستقیم در f۱ و دقت ببینم. این نتیجه حاصل شد که متغیر default_encoded هیچ تاثیری در دقت و f۱ score ندارد، در نتیجه حذف شد. اما داستان برای متغیرهای دیگر کمی فرق دارد. اول اینکه به جز feature های اصلی حذف کردن یا نکردن بقیه feature ها تاثیر چندانی در دقت و f۱ score ندارد و بازه تغییرات بین ۰.۵ تا ۳ درصد است. دوم اینکه حذف کردن یا نکردن feature های غیر مهم تاثیر یکسانی روی متود های مختلف نداشت. برای مثال حذف کردن یک featureباعث افزایش f۱score در روش knn می شود ولی در عین حال باعث کاهش در روش decision tree و random forest می شود. در انتها برای انتخاب feature ها از random forest و rfe استفاده شد تا ۱۰ feature مهم بدست بیاید. نتیجه به صورت زیر شد:



از ۱۰ feature بالا برای train مدل ها استفاده شد.

# Decision tree



# مقایسه نتایج با استفاده از روش‌های مختلف

| Metric/Method | KNN | Decision Tree | Random Forest | Logistic Regression | SVM Linear | SVM RBF | MLP |
|---|---|---|---|---|---|---|---|
| F1 score | 0.422 | 0.476 | 0.468 | 0.458 | 0.455 | 0.393 | 0.451 |
| Accuracy | 0.897 | 0.895 | 0.896 | 0.793 | 0.789 | 0.900 | 0.898 |

# اقدامات جهت بهبود f۱ score

همان طور که مشاهده می کنید، accuracy برای اکثر روش ها ۹۰ می باشد، به این معنی که کلاس اکثریت به خوبی پیش بینی می شود، اما برای اقلیت نتیجه خوبی ندارد و پایین بودن recall و در نتیجه پایین بودن f۱ score حاکی از آن است. به همین دلیل باید اقدامات مناسب با دیتاست نابالانس روی آن اجرا شود. از بین روش های موجود،(SMOTE(Synthetic minority oversampling ، Undersampling و ADASYN(adaptive synthetic) پیاده و اجرا شد و نتایج آن را در زیر مشاهده می کنید:

## SMOTE

| Metric/Method | KNN | Decision Tree | Random Forest | Logistic Regression | SVM Linear | SVM RBF | MLP |
|---|---|---|---|---|---|---|---|
| F1 score | 0.465 | 0.469 | 0.499 | 0.444 | 0.443 | 0.477 | 0.355 |
| Accuracy | 0.827 | 0.800 | 0.856 | 0.798 | 0.791 | 0.816 | 0.615 |

# Undersampling

| Metric/Method | KNN | Decision Tree | Random Forest | Logistic Regression | SVM Linear | SVM RBF | MLP |
|---|---|---|---|---|---|---|---|
| F1 score | 0.430 | 0.478 | 0.500 | 0.459 | 0.457 | 0.484 | 0.486 |
| Accuracy | 0.785 | 0.796 | 0.810 | 0.795 | 0.792 | 0.796 | 0.799 |

# ADASYN

| Metric/Method | KNN | Decision Tree | Random Forest | Logistic Regression | SVM Linear | SVM RBF | MLP |
|---|---|---|---|---|---|---|---|
| F1 score | 0.453 | 0.450 | 0.493 | 0.458 | 0.405 | 0.458 | 0.468 |
| Accuracy | 0.823 | 0.776 | 0.842 | 0.793 | 0.738 | 0.795 | 0.797 |

# نتیجه

بر خلاف انتظار روش های بالا تاثیر چندانی در بهبود f۱ score نداشتند که بیش تر به نظر می رسد به دلیل ضعیف بودن feature ها باشد تا کم بودن تعداد داده های کلاس "بله". بیش ترین f۱ score برای random forest در smote و undersampling می باشد با عدد ۰.۵۰ و بیش ترین accuracy برای svm rbf در دیتاست بدون تغییر می باشد.