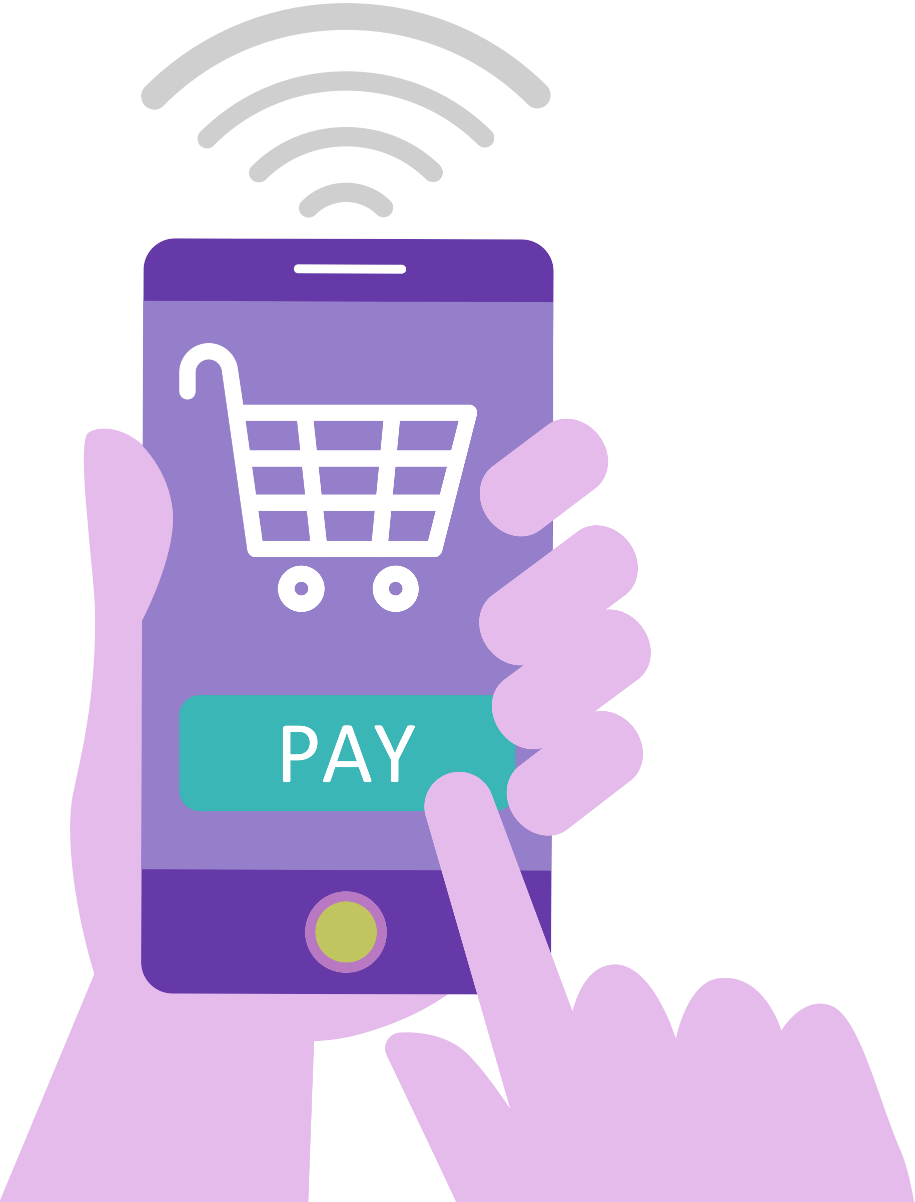
Università di Pisa

Artificial Intelligence and Data Engineering

CHURNPREDICT

Data Mining and Machine Learning project

Denny Meini

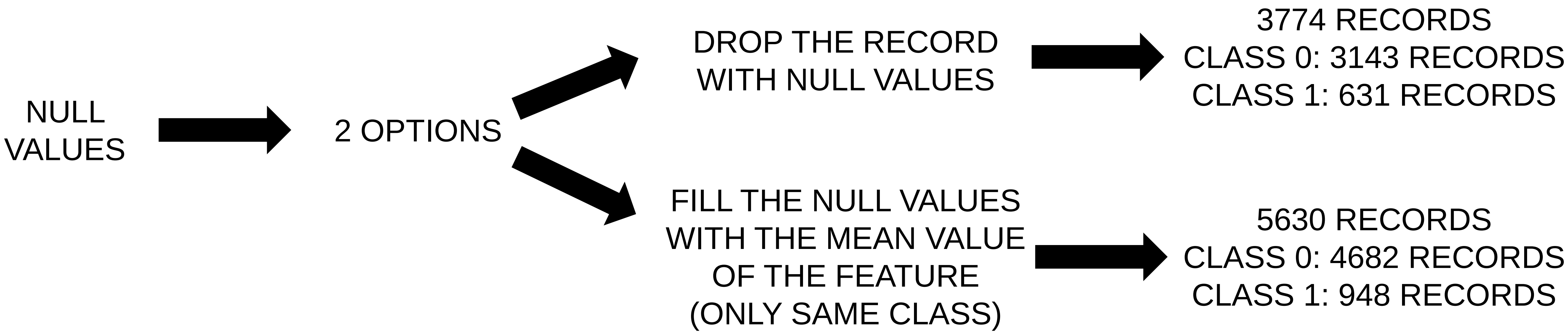


PROJECT GOALS

The goals of this project are to compare a list of classifiers in order to discover which of them gives us the best performance in terms of some parameters (f-score and average accuracy). Another goal is to build an application that uses the classifier which gave the best results. The dataset has some null values and an additional aim is to know if is better to drop the incomplete records or to complete them.

DATASET

I took the dataset from kaggle
Link: https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction
5630 records with 19 features and 1 binary target



NULL VALUES

2 OPTIONS

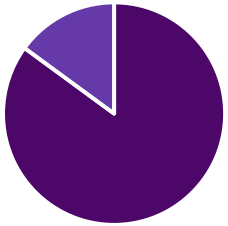
DROP THE RECORD WITH NULL VALUES

3774 RECORDS
CLASS 0: 3143 RECORDS
CLASS 1: 631 RECORDS

FILL THE NULL VALUES WITH THE MEAN VALUE OF THE FEATURE (ONLY SAME CLASS)

5630 RECORDS
CLASS 0: 4682 RECORDS
CLASS 1: 948 RECORDS

IMBALANCED DATASET



DATASET

Numerical feature
Categorical feature

CustomerID

Churn

Output class [0: did not left, 1 left].

Tenure

The number of years the customer has been a customer.

PreferredLoginDevice

CityTier

The tier of the city (Chinese model) [1: big, 2: medium, 3: small].

WarehouseToHome

Distance between the warehouse and the customer's house.

PreferredPaymentMode

Gender

HourSpendOnApp

Number of hours the customer has spent on the app.

NumberOfDeviceRegistered

PreferredOrderCat

SatisfactionScore

Grade of satisfaction of the customer from 1 (not satisfied) to 5 (very satisfied).

MaritalStatus

NumberOfAddress

Number of address added by a customer.

Complain

If the customer complained or not during the last month [0: no, 1: yes].

OrderAmountHikeFromlastYear

Percentage increase regarding the orders from the previous year.

CouponUsed

Number of coupon the customer used last month.

OrderCount

Number of order the customer placed during last month.

DaySinceLastOrder

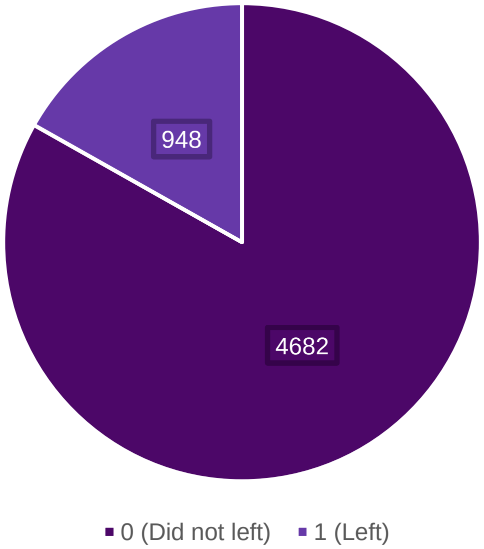
CashbackAmount

Average cashback of the customer last month.

PREPROCESSING

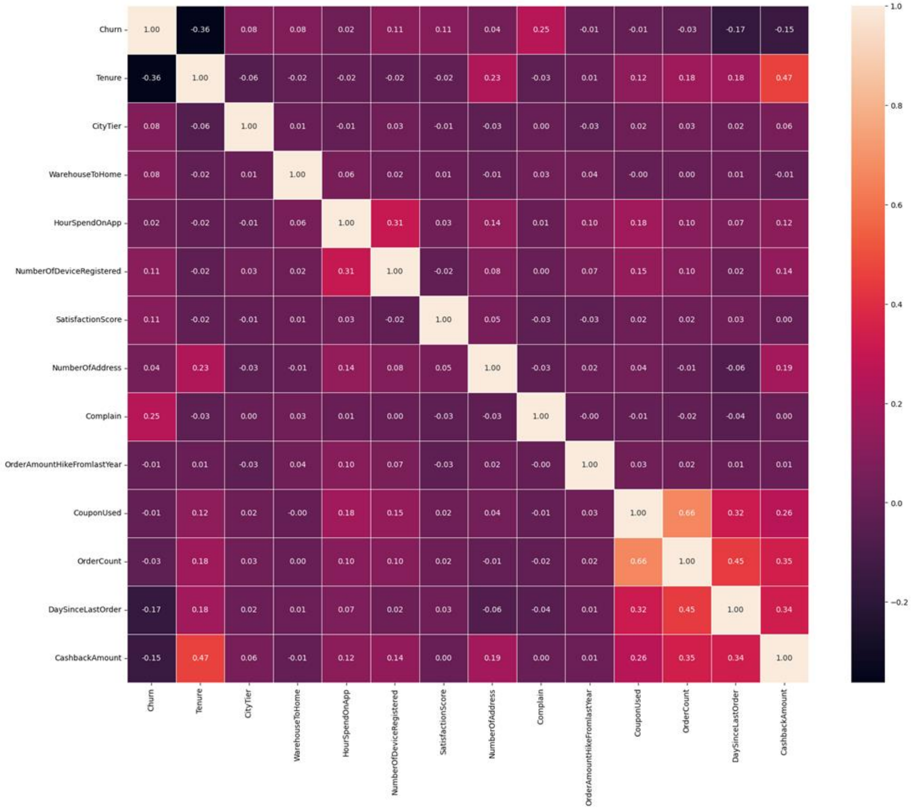
Null values problem:
Creation of the two datasets

Balance analysis:
Imbalanced Dataset



Drop of useless features:
CustomerID

Correlation analysis:
no correlated features



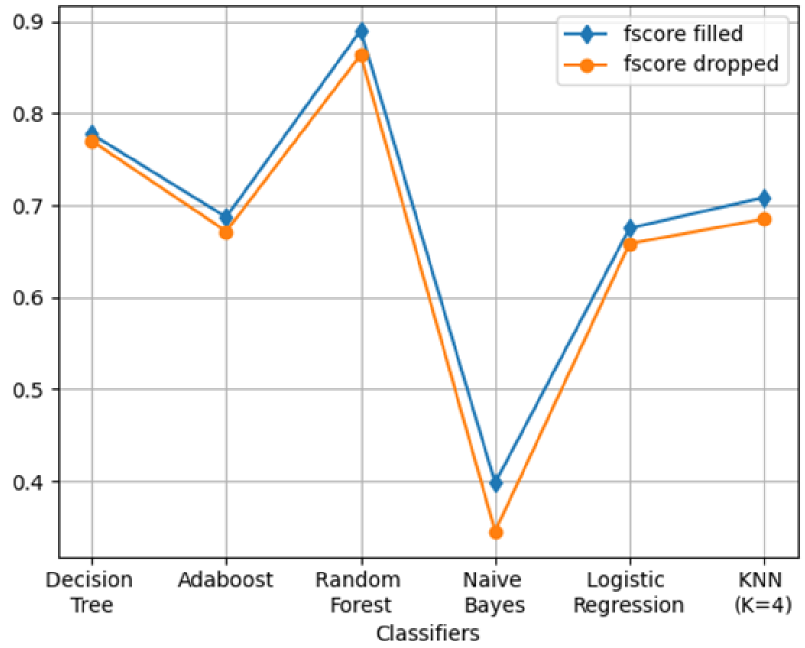
Feature extraction:
Obtaining numerical representation of categorical features: getDummies()

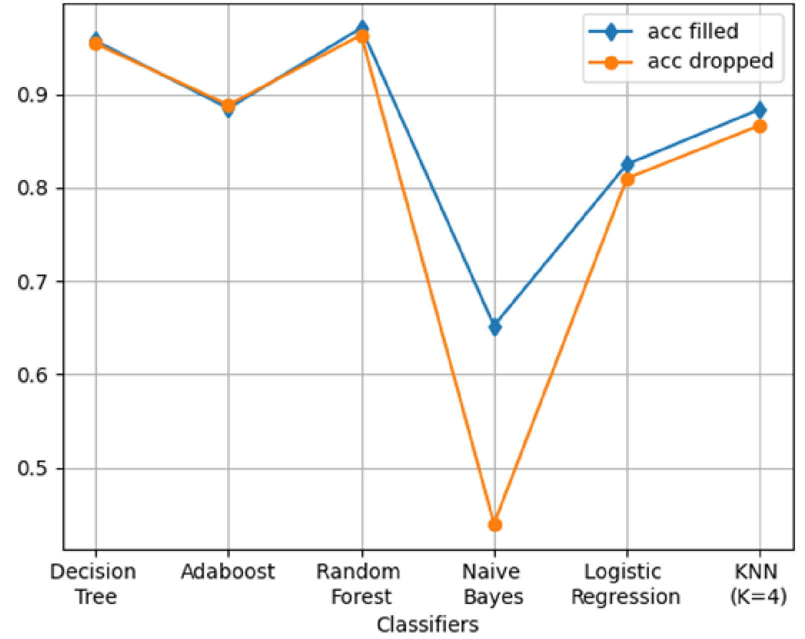
CLASSIFICATION

Cross validation using Stratified K Fold (k=5)
Rebalancing using SMOTE
For each classifier I got f-score and Avg Accuracy

Classifiers		f-score	Avg accuracy
Decision Tree	Filled	0.777	0.957
	Dropped	0.770	0.954
AdaBoost	Filled	0.687	0.885
	Dropped	0.672	0.888
Random Forest	Filled	0.890	0.971
	Dropped	0.864	0.963
Gaussian Naive Bayes	Filled	0.398	0.652
	Dropped	0.345	0.440
Logistic Regression	Filled	0.675	0.825
	Dropped	0.658	0.810
KNN (K=4)	Filled	0.708	0.884
	Dropped	0.685	0.866

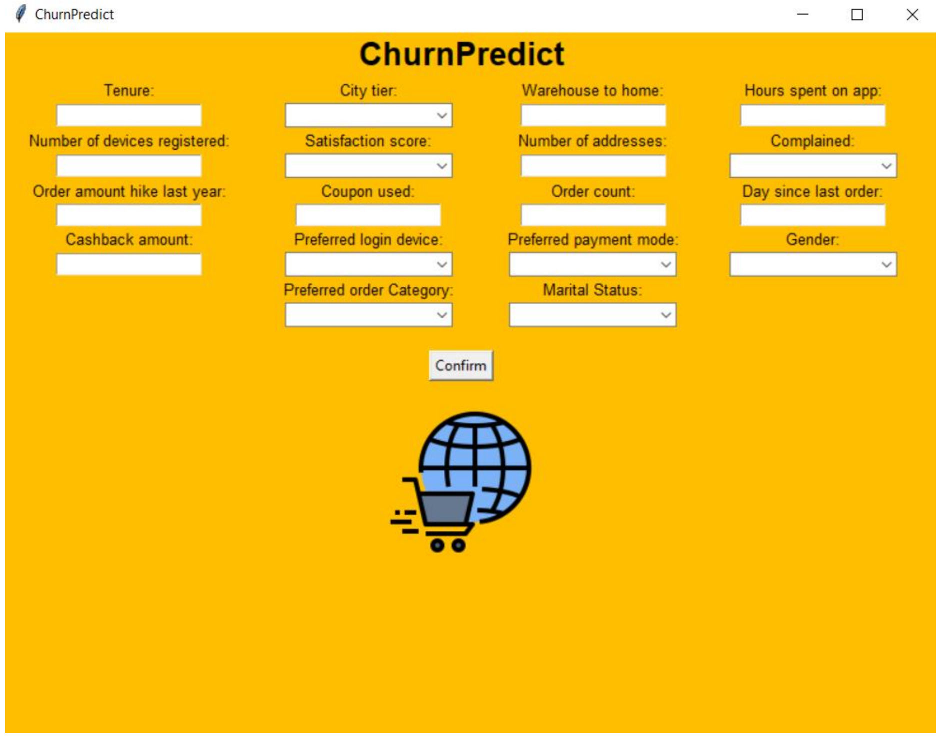
Wilcoxon test: pvalue=0.0625
↓
No statistical relevance





APPLICATION IMPLEMENTATION

Input: The values of the features of the dataset
Output: Presence or not of churn risk

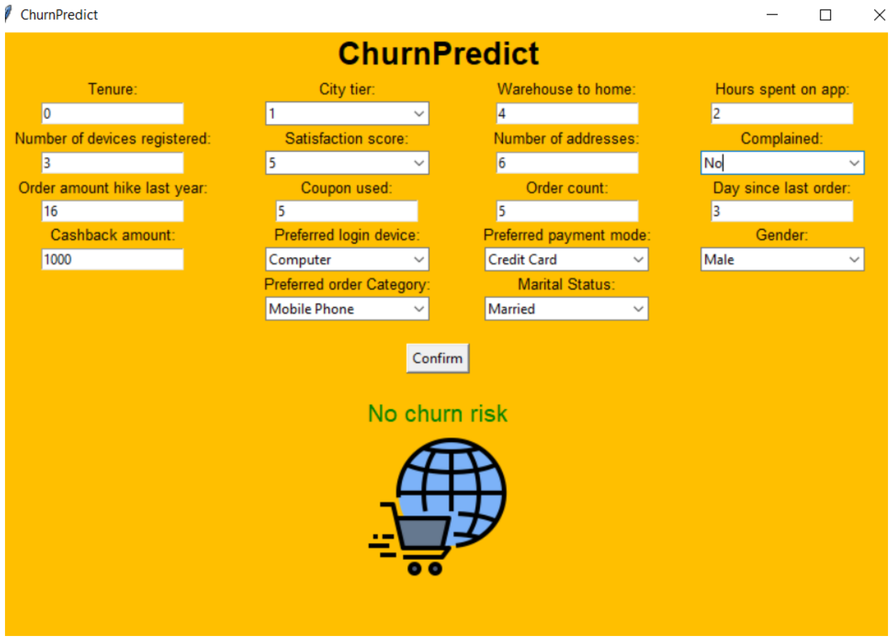


ChurnPredict

ChurnPredict

Tenure:
City tier:
Warehouse to home:
Hours spent on app:
Number of devices registered:
Satisfaction score:
Number of addresses:
Complained:
Order amount hike last year:
Coupon used:
Order count:
Day since last order:
Cashback amount:
Preferred login device:
Preferred payment mode:
Gender:
Preferred order Category:
Marital Status:

NO RISK



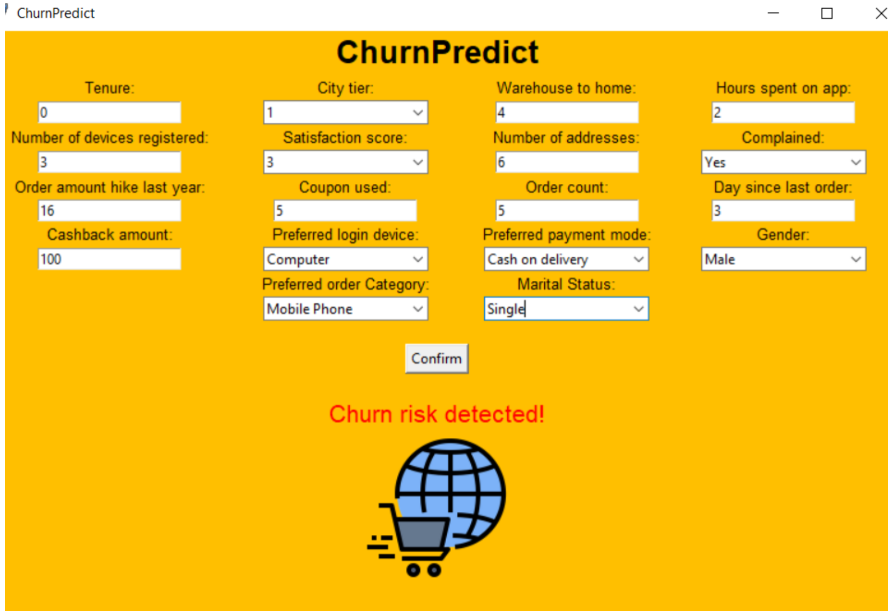
ChurnPredict

ChurnPredict

Tenure: 0
City tier: 1
Warehouse to home: 4
Hours spent on app: 2
Number of devices registered: 3
Satisfaction score: 5
Number of addresses: 6
Complained: No
Order amount hike last year: 16
Coupon used: 5
Order count: 5
Day since last order: 3
Cashback amount: 1000
Preferred login device: Computer
Preferred payment mode: Credit Card
Gender: Male
Preferred order Category: Mobile Phone
Marital Status: Married

No churn risk

RISK



ChurnPredict

ChurnPredict

Tenure: 0
City tier: 1
Warehouse to home: 4
Hours spent on app: 2
Number of devices registered: 3
Satisfaction score: 3
Number of addresses: 6
Complained: Yes
Order amount hike last year: 16
Coupon used: 5
Order count: 5
Day since last order: 3
Cashback amount: 100
Preferred login device: Computer
Preferred payment mode: Cash on delivery
Gender: Male
Preferred order Category: Mobile Phone
Marital Status: Single

Churn risk detected!

SHOP

THANK YOU FOR YOUR ATTENTION