
THE TRAVEL EXPERIENCE

AIM OF THE ANALYSIS

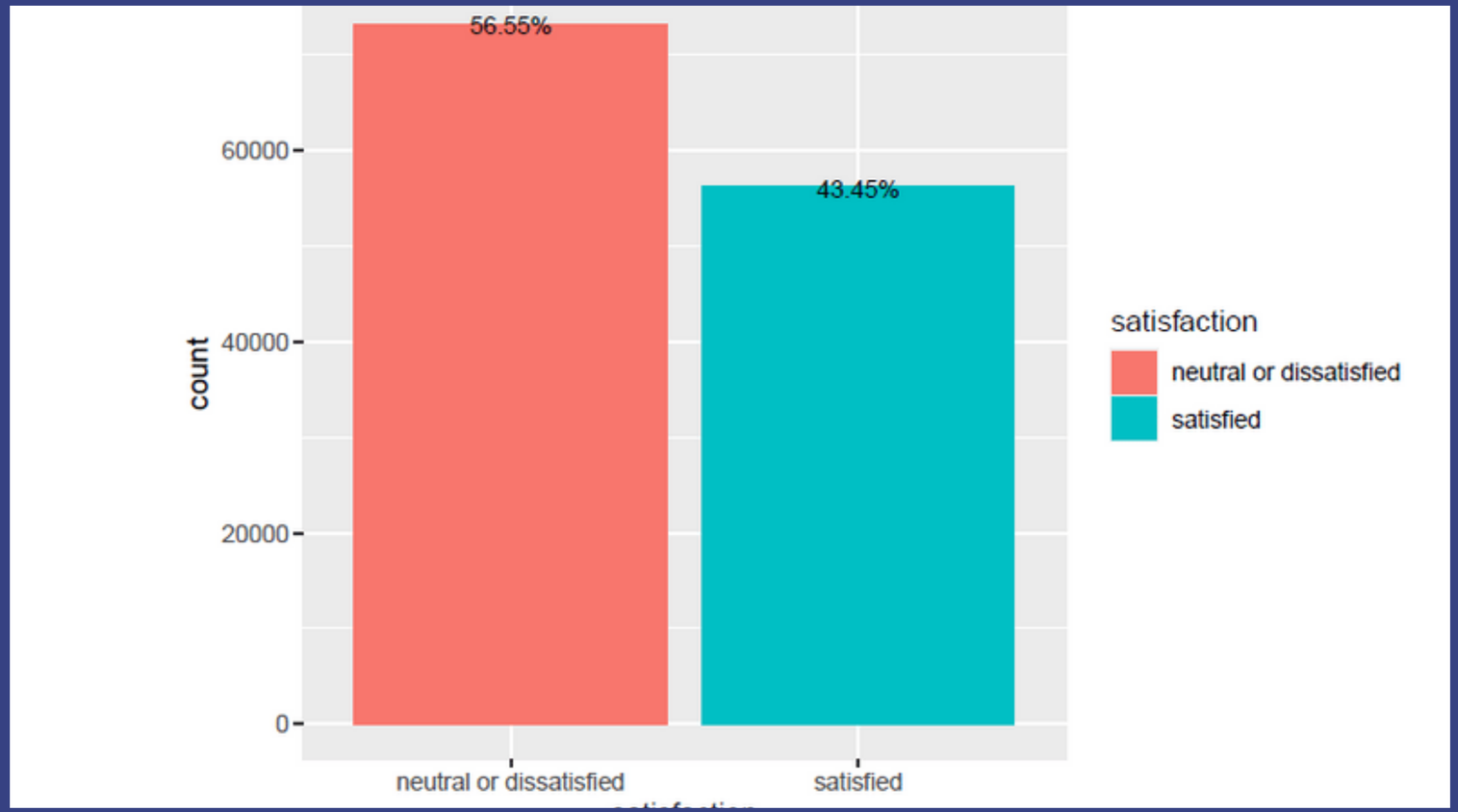
Airlines market has reached
high levels of competition.
For this reason,
the customer satisfaction is
a key feature to survive.
Using a dataset of kaggle, the analysis aims
to develop a model
able to accurately predict
the satisfaction of the passengers
and which are the most relevant variables.

FEATURES OF THE DATASET

Almost 130000 observations are present, classified with the following variables:

- **satisfaction** = the dependent variable, dichotomous variable, *satisfied* or *neutral or dissatisfied.*
- **gender** = dichotomous variable, *male* or *female.*
- **customer_type** = dichotomous variable, *loyal customer* or *disloyal customer.*
- **age** = continuous variable.
- **type_of_travel** = dichotomous variable, *business travel* or *personal travel.*
- **customer class** = dichotomous variable, *business* or *eco.*
- **flight_distance** = continuous variable.
- **inflight_wifi_service** = discrete variable, assigned score between 0 and 5.
- **department_arrival_time_convenient** = discrete variable, assigned score between 0 and 5.
- **ease_of_online_booking** = discrete variable, assigned score between 0 and 5.
- **gate_location** = discrete variable, assigned score between 0 and 5.
- **food_and_drink** = discrete variable, assigned score between 0 and 5.
- **online_boarding** = discrete variable, assigned score between 0 and 5.
- **seat_comfort** = discrete variable, assigned score between 0 and 5.
- **inflight_entertainment** = discrete variable, assigned score between 0 and 5.
- **onboard_service** = discrete variable, assigned score between 0 and 5.
- **leg_room_service** = discrete variable, assigned score between 0 and 5.
- **baggage_handling** = discrete variable, assigned score between 0 and 5.
- **checkin_service** = discrete variable, assigned score between 0 and 5.
- **inflight_service** = discrete variable, assigned score between 0 and 5.
- **cleanliness** = discrete variable, assigned score between 0 and 5.
- **departure_delay_in_minutes** = continuous variable, in minutes.
- **arrival_delay_in_minutes** = continuous variable, in minutes.

I made descriptive statistics for all the variables, to check the balance of the dependent variable (satisfaction) and which variables have an high probability to be associated with a satisfied customer.



The satisfaction is balanced between the two discrete values.

STEPS OF THE ANALYSIS

Logistic regression

Tree predictors

K-NN algorithms

LOGISTIC REGRESSION

FULL MODEL

First of all, I ran the full model
considering all the variables.
Since the huge number of categorical variables,
75 variables are included in the first model.

DETECTING MULTICOLLINEARITY

Therefore, I checked for potential multicollinearity
amnog the variables, considering 10 as threshold
for the variance inflation factor.
Then, at each step, I discarded from the model
the variable with the high vif,
until an empty 'high correlation' set.

```
## High Correlation
##
##                                 Term     VIF Increased SE Tolerance
##  departure_arrival_time_convenient   31.69          5.63      0.03
##             ease_of_online_booking 1334.29         36.53      0.00
##                      gate_location   11.82          3.44      0.08
##                     food_and_drink   19.06          4.37      0.05
##                    online_boarding   95.20          9.76      0.01
##             inflight_entertainment  125.38         11.20      0.01
##                   leg_room_service   24.96          5.00      0.04
##                        cleanliness   16.39          4.05      0.06
##         departure_delay_in_minutes   15.48          3.93      0.06
##           arrival_delay_in_minutes   15.52          3.94      0.06
```

These are the VIFs for the full model. ***Ease_of_online_booking*** is the first discarded variables. Then, the next full models selected, in order: ***inflight_entertainment***, ***departure_delay_in_minutes*** and ***departure_arrival_time_convinient***.

I continued the logistic regression analysis without them.

LOGISTIC REGRESSION STEPS

ACCURACY WITH VALIDATION SET

I used a validation set approch, splitting the starting dataset with a 0.7 training size.

ADDING ROBUSTNESS

I added robustness in several ways:

- changing two times data partition
- changing the training size
- 10-fold corss validation

NORMALIZING VARIABLES

Furthermore, I tried normalizing vaiables, even if the units of measure were quite similar.

```
##           Reference
## Prediction    No   Yes
##        No  48699  3372
##        Yes  2558 36011
##
##                Accuracy : 0.9346
##                  95% CI : (0.9329, 0.9362)
##     No Information Rate : 0.5655
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8665
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9144
##             Specificity : 0.9501
##          Pos Pred Value : 0.9337
##          Neg Pred Value : 0.9352
##              Prevalence : 0.4345
##          Detection Rate : 0.3973
##    Detection Prevalence : 0.4255
##       Balanced Accuracy : 0.9322
##
##        'Positive' Class : Yes
##
```

```
##           Reference
## Prediction    No   Yes
##        No  20824  1507
##        Yes  1143 15371
##
##                Accuracy : 0.9318
##                  95% CI : (0.9292, 0.9343)
##     No Information Rate : 0.5655
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8608
##
##  Mcnemar's Test P-Value : 1.77e-12
##
##             Sensitivity : 0.9107
##             Specificity : 0.9480
##          Pos Pred Value : 0.9308
##          Neg Pred Value : 0.9325
##              Prevalence : 0.4345
##          Detection Rate : 0.3957
##    Detection Prevalence : 0.4251
##       Balanced Accuracy : 0.9293
##
##        'Positive' Class : Yes
##
```

Here the confusion matrices are reported, the training one on the left and the test one on the right. The test accuracy is really high (0.9318), as the sensitivity and the specificity. All the checks give consistent result around 0.93, with a peak equal to 0.9369 with the bigger training size. Results did not change with normalized variables.

Logistic regression

Logistic regression gives good predictive performances, but I'm looking for more interpretable results.

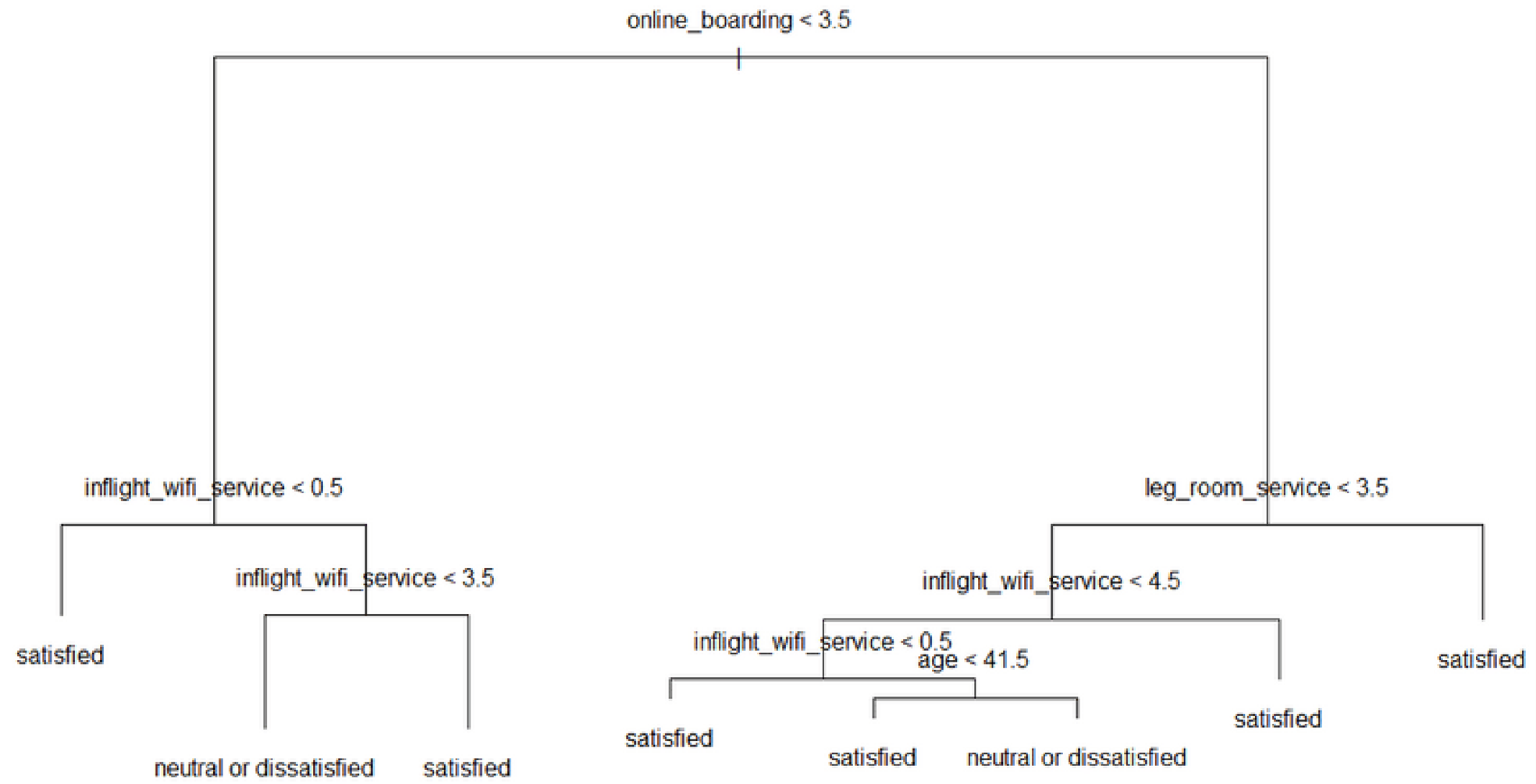
Trees

Trees have these features, therefore I used them to visualize relevant variables and to understand how they work.

TREES

I started growing the full tree. After have looked at it, the need to prune it came out. According to the miscalissification rate, I passed the total number of final nodes from 11 to 8. The result becomes more clear, but the prediction accuracy stay always around 0.85.

PRUNED TREE



Online_boarding plays the role of the first split, then the presence of ***inflight_wifi_service*** is relevant to reach the final nodes. The results are the same without considering the variables discarded in the logistic regression section.

SUMMARY

LOGISTIC REGRESSION

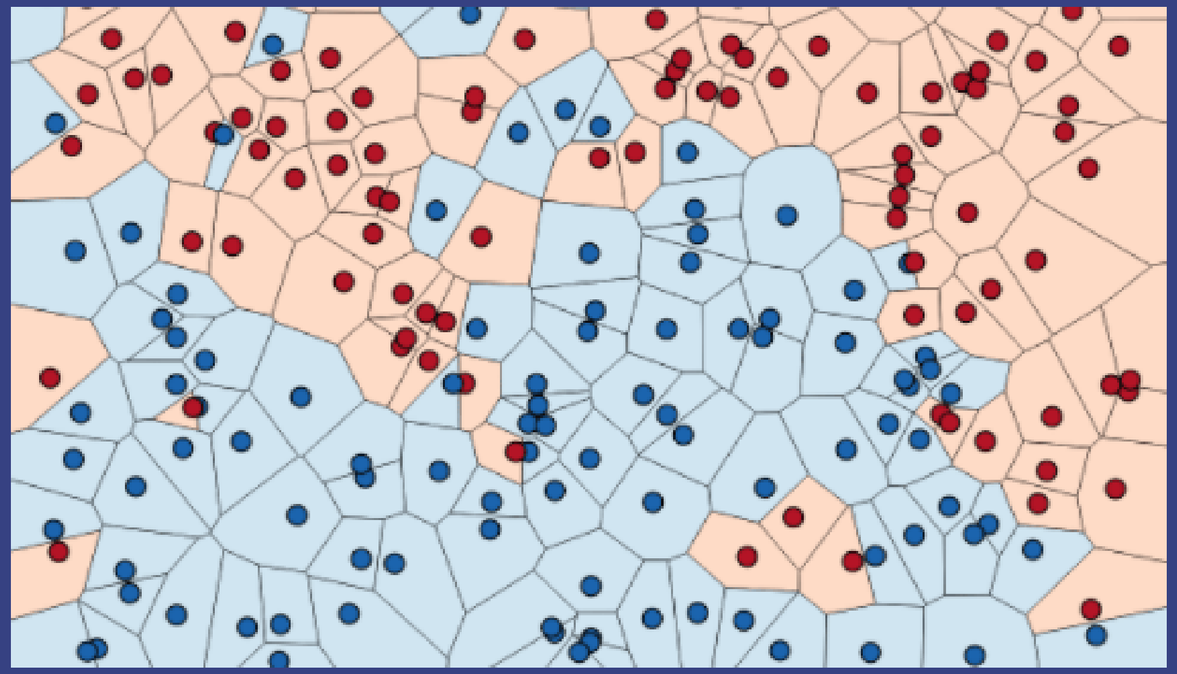
Good predictive performances, but less easy interpretation of the variables.

TREES

Useful to interpret the role of the variables, but shrunk predictive performance...

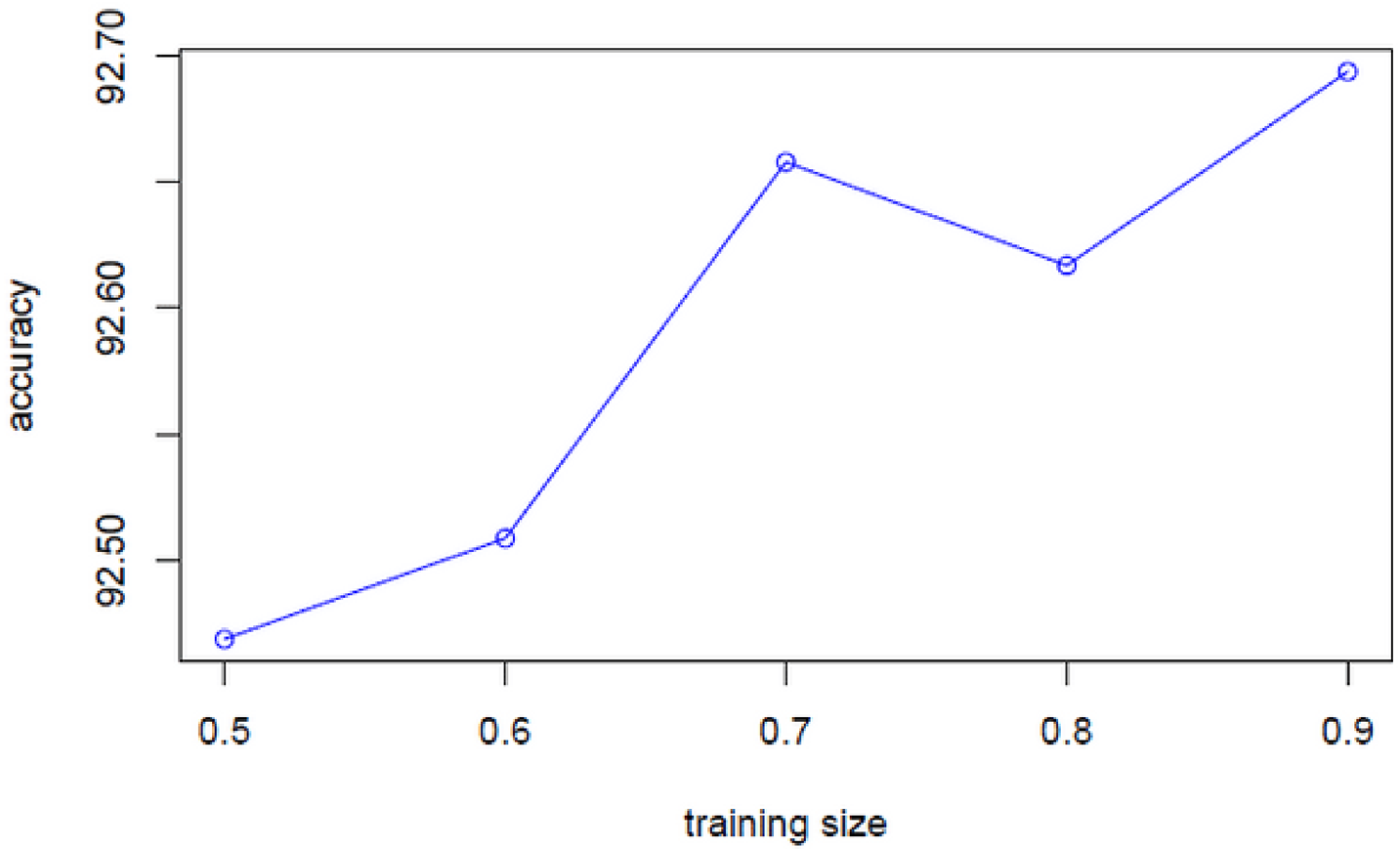
K-NN

To bring back the accuracy as in the logistic regression, I used the K-nearest neighbours



K-NN algorithms

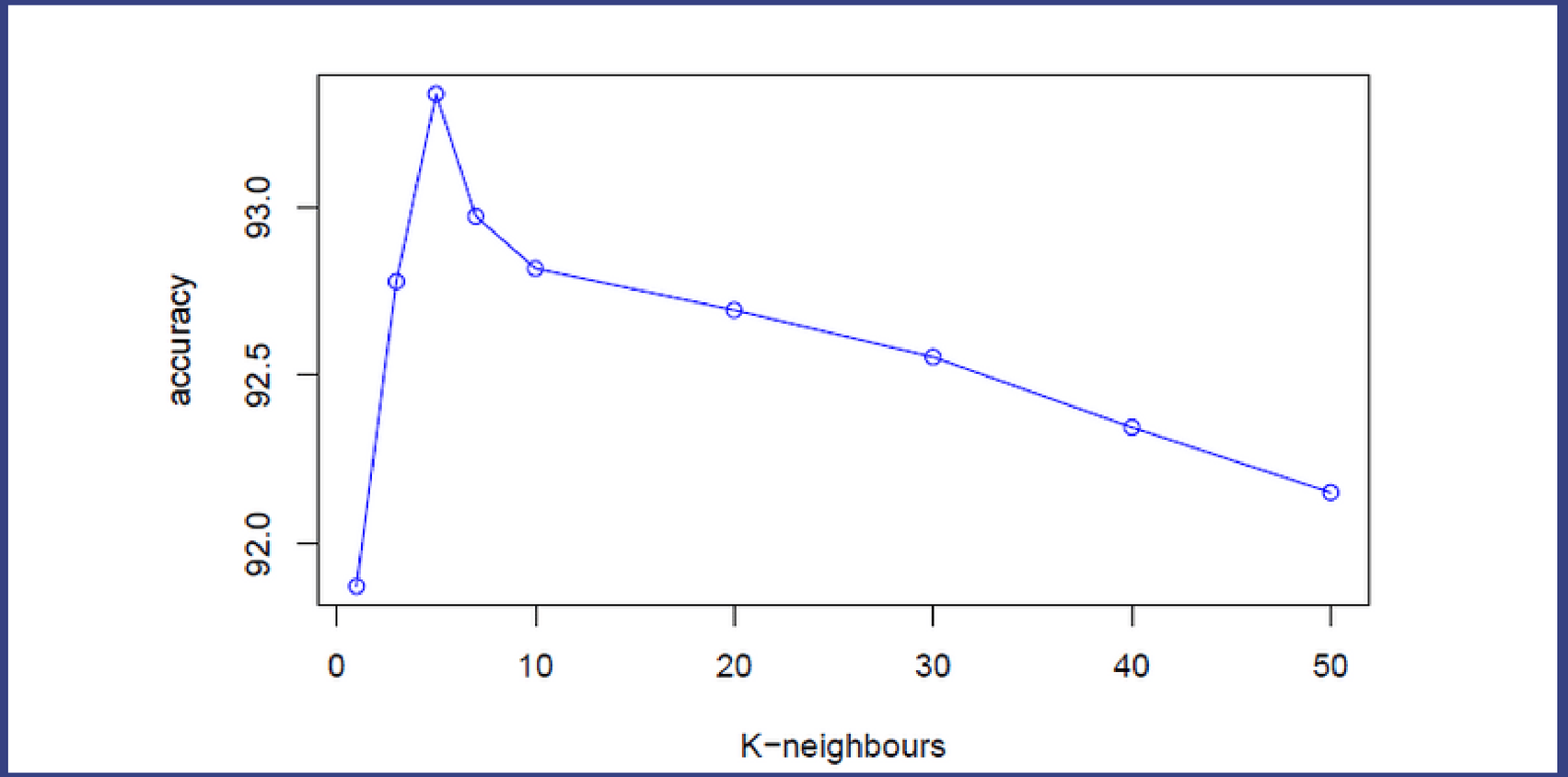
After having normalized the variables, I chose a random k (equal to 20) and I ran the 20-NN algortihm with different training sizes (0.5, 0.6, 0.7, 0.8, 0.9) to select the best one.



The findings are good: the prediction accuracy with k = 20 seems to be near to the one of the logistic regression analysis. The peak is reached with a training size equal to 0.9 with respect to the starting dataset. In this point, the test accuracy is equal to 0.9265.

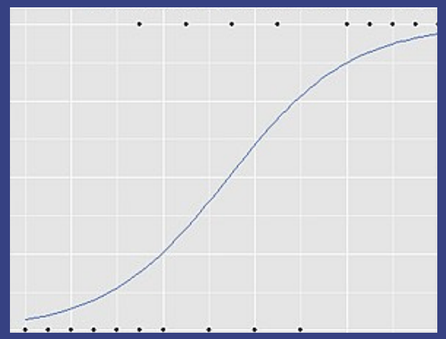
I went on with the analysis considering 0.9 as training size and by varying the number of considered k to get the parameter with the best test accuracy.

K-NN test accuracy



Considering a discrete set of values for k (1, 3, 5, 7, 10, 20, 30, 40, 50), I obtained the test accuracies for these different values. After a situation of overfitting with k = 1, the peak is quickly reached with k equal to 5 and, after a downward leap, the decrease becomes constant. The test accuracy, when k = 5 and the training size is equal to 0.9, is 0.9333 as it was in the logistic regression. The results remain constant without the four variables previously discarded.

CONCLUSIONS

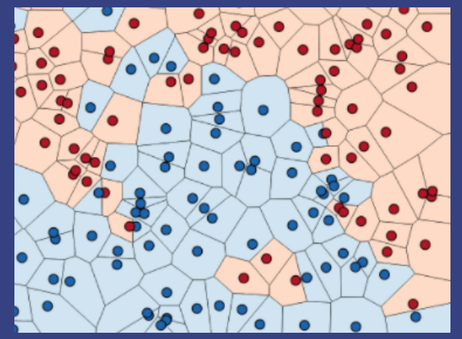


LOGISTIC REGRESSION

Good test predictive performances, around 0.93, robust result with all the checks.

TREES

Online_boarding and **inflight_wifi_service** as two main variables to classify satisfied and neutral/dissatisfied customers.



K-NN

Test performances good as in the logistic regression, thid dds robustness to the 0.93 accuracy.

THANK YOU FOR THE ATTENTION!