T5o10 Data Science Bootcamp

Albraa Baghdadi

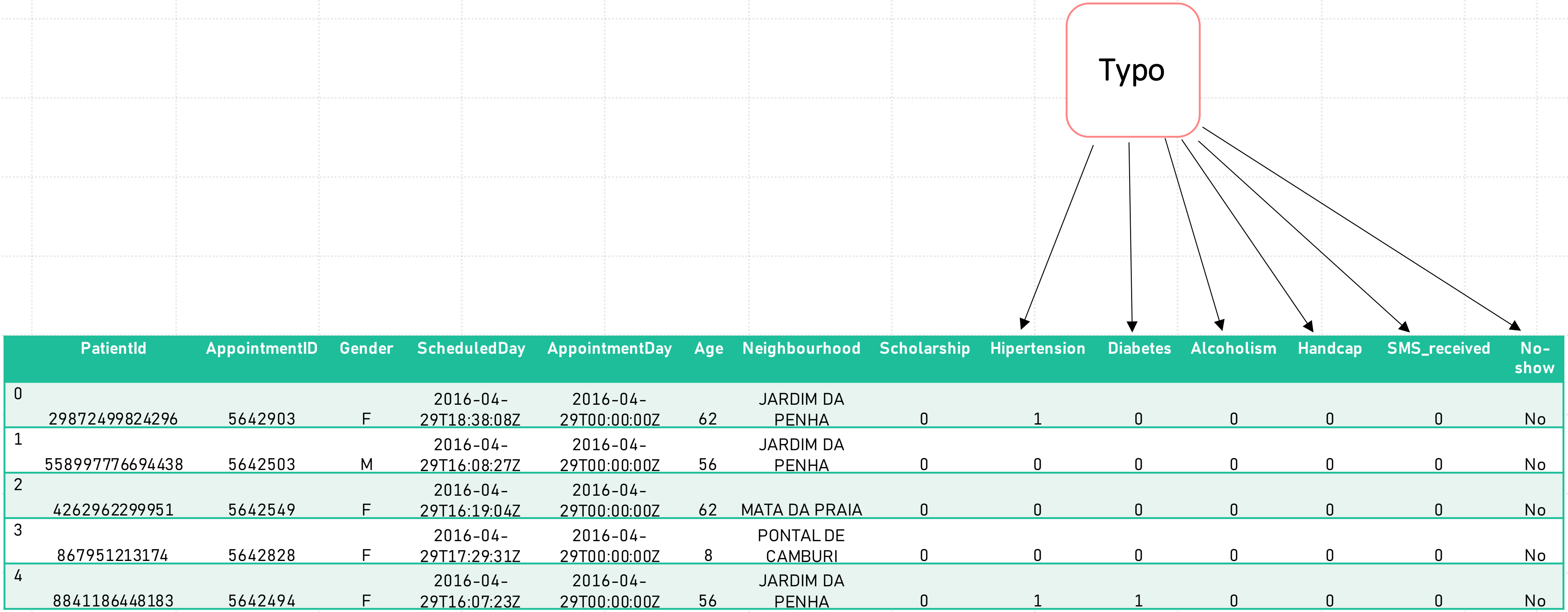
INTRODUCTION

- Missing appointment happens a lot and sometimes affect businesses that depend on appointments to complete the service such as Hospitals. In this project we will build a model based on our dataset to predict the probability of the patient to miss the appointment.
- The benefit of this model that the hospital could predict the probability of the patient to miss the appointment and then increase number of appointment for that day.

DATASET INFORMATION

PatientId
AppointmentID
Age
Gender
Scheduled Day
Appointment Day
Neighborhood
Scholarship
Hypertension
Diabetes
Alcoholism
Handicap
SMS_received
No-show

- The dataset is provided by Kaggle and contains more than 100000+ records with 14 features



Typo

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	4262962299951	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	867951213174	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	8841186448183	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

DATA CLEANING

DATA CLEANING

Change No to "0" and Yes to "1"

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	4262962299951	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	867951213174	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	8841186448183	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

DATA CLEANING

Remove the records with > 0 Age

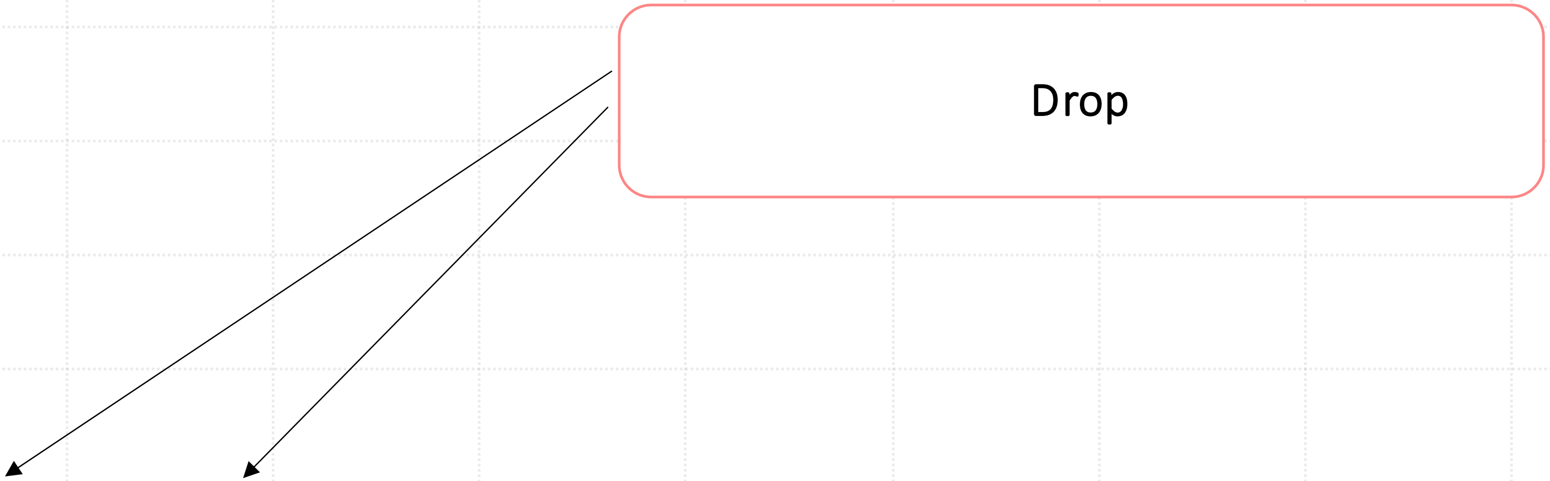
	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	4262962299951	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	867951213174	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	8841186448183	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

Change the type to date type

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	4262962299951	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	867951213174	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	8841186448183	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

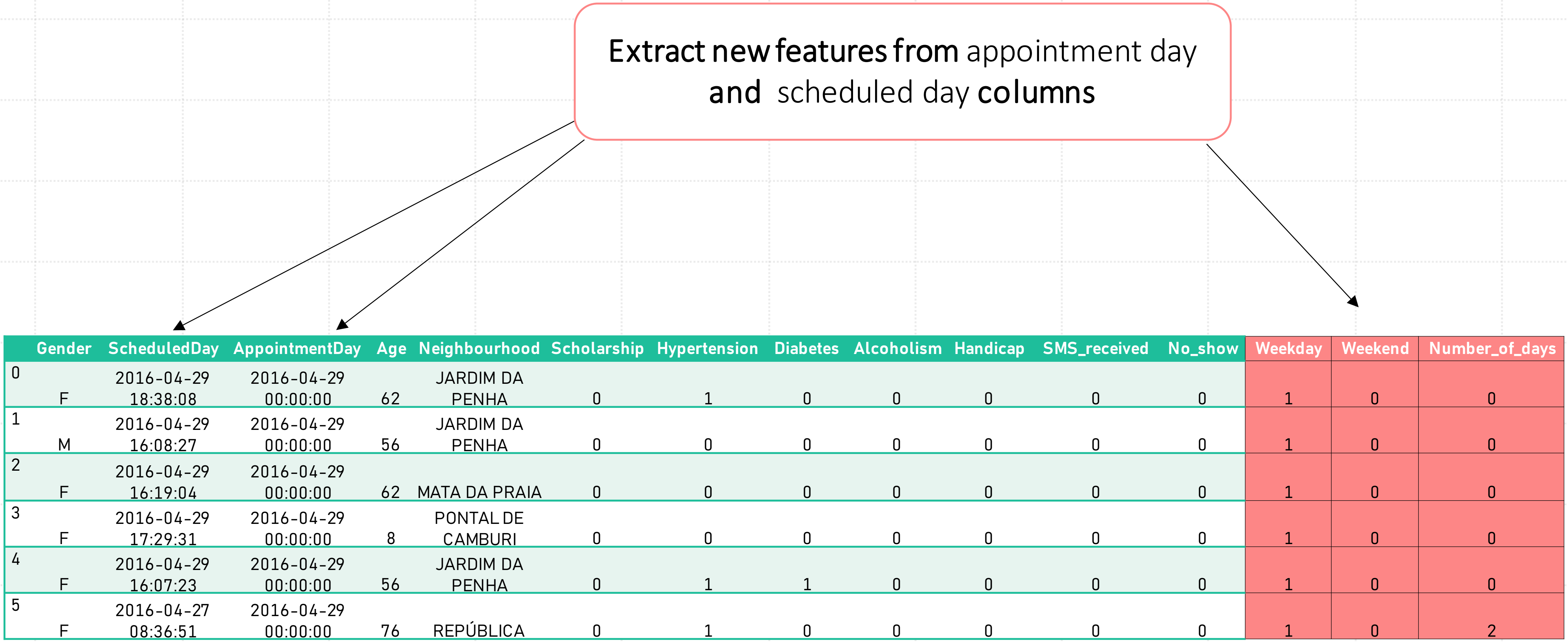
DATA CLEANING

DATA CLEANING



Drop

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handcap	SMS_received	No-show
0	29872499824296	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	0	0	No
1	558997776694438	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	0	0	No
2	4262962299951	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	0	0	No
3	867951213174	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	0	0	No
4	8841186448183	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	0	0	No

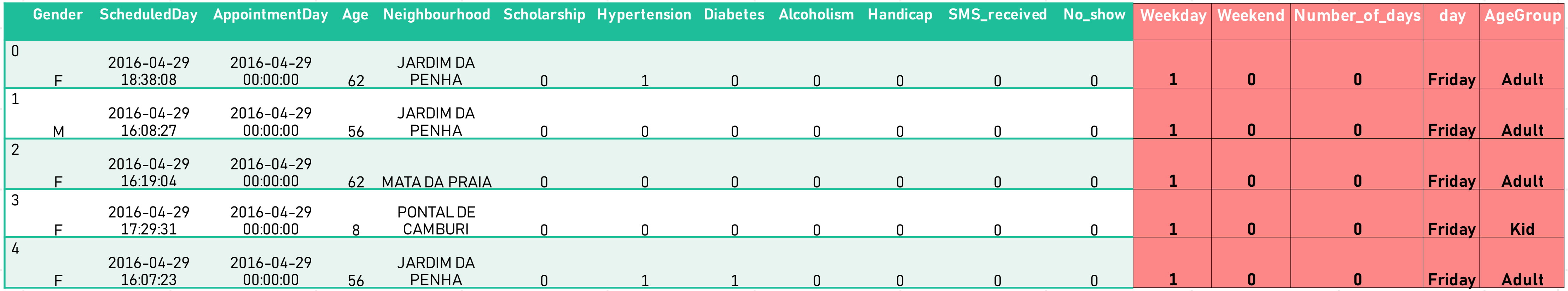


Extract new features from appointment day and scheduled day columns

	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No_show	Weekday	Weekend	Number_of_days
0	F	2016-04-29 18:38:08	2016-04-29 00:00:00	62	JARDIM DA PENHA	0	1	0	0	0	0	0	1	0	0
1	M	2016-04-29 16:08:27	2016-04-29 00:00:00	56	JARDIM DA PENHA	0	0	0	0	0	0	0	1	0	0
2	F	2016-04-29 16:19:04	2016-04-29 00:00:00	62	MATA DA PRAIA	0	0	0	0	0	0	0	1	0	0
3	F	2016-04-29 17:29:31	2016-04-29 00:00:00	8	PONTAL DE CAMBURI	0	0	0	0	0	0	0	1	0	0
4	F	2016-04-29 16:07:23	2016-04-29 00:00:00	56	JARDIM DA PENHA	0	1	1	0	0	0	0	1	0	0
5	F	2016-04-27 08:36:51	2016-04-29 00:00:00	76	REPÚBLICA	0	1	0	0	0	0	0	1	0	2

DATA CLEANING

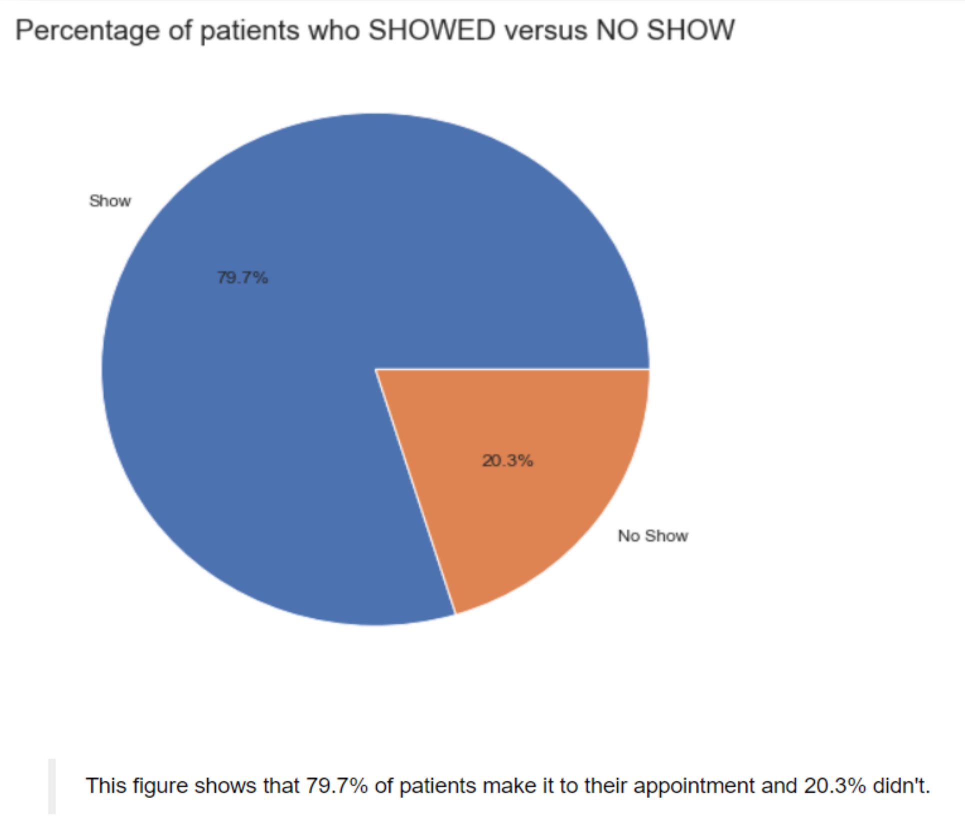
Extract AgeGroup from Age column



	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	No_show	Weekday	Weekend	Number_of_days	day	AgeGroup
0	F	2016-04-29 18:38:08	2016-04-29 00:00:00	62	JARDIM DA PENHA	0	1	0	0	0	0	0	1	0	0	Friday	Adult
1	M	2016-04-29 16:08:27	2016-04-29 00:00:00	56	JARDIM DA PENHA	0	0	0	0	0	0	0	1	0	0	Friday	Adult
2	F	2016-04-29 16:19:04	2016-04-29 00:00:00	62	MATA DA PRAIA	0	0	0	0	0	0	0	1	0	0	Friday	Adult
3	F	2016-04-29 17:29:31	2016-04-29 00:00:00	8	PONTAL DE CAMBURI	0	0	0	0	0	0	0	1	0	0	Friday	Kid
4	F	2016-04-29 16:07:23	2016-04-29 00:00:00	56	JARDIM DA PENHA	0	1	1	0	0	0	0	1	0	0	Friday	Adult

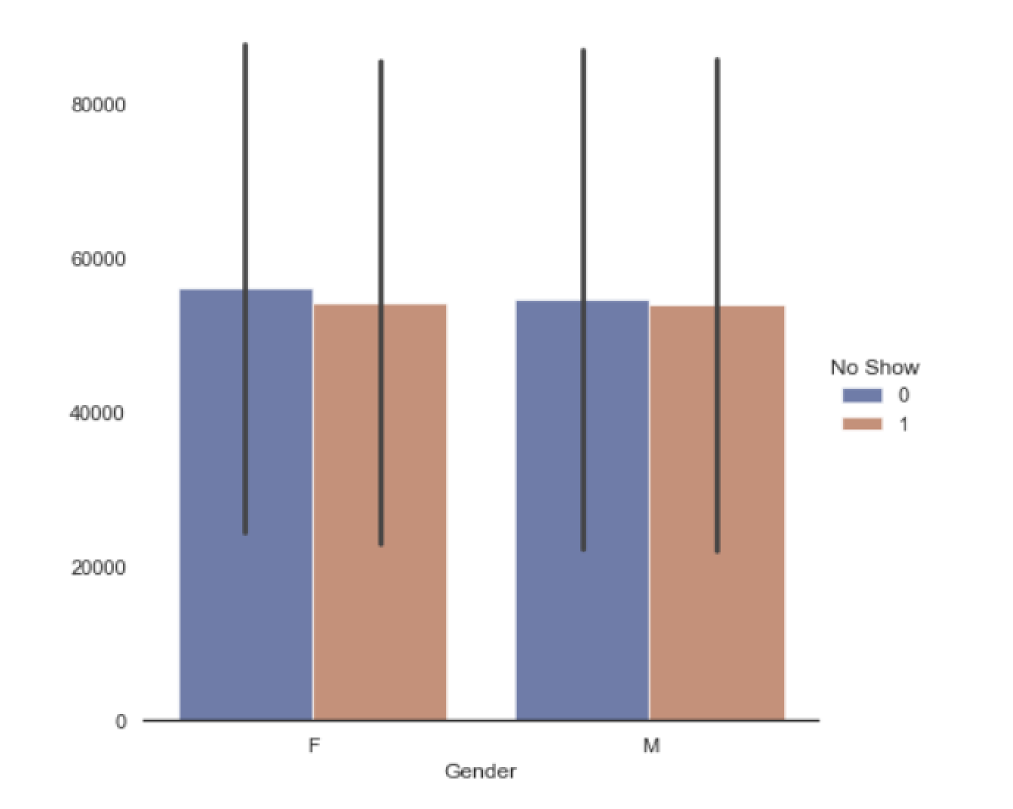
DATA CLEANING

DATA ANALYST



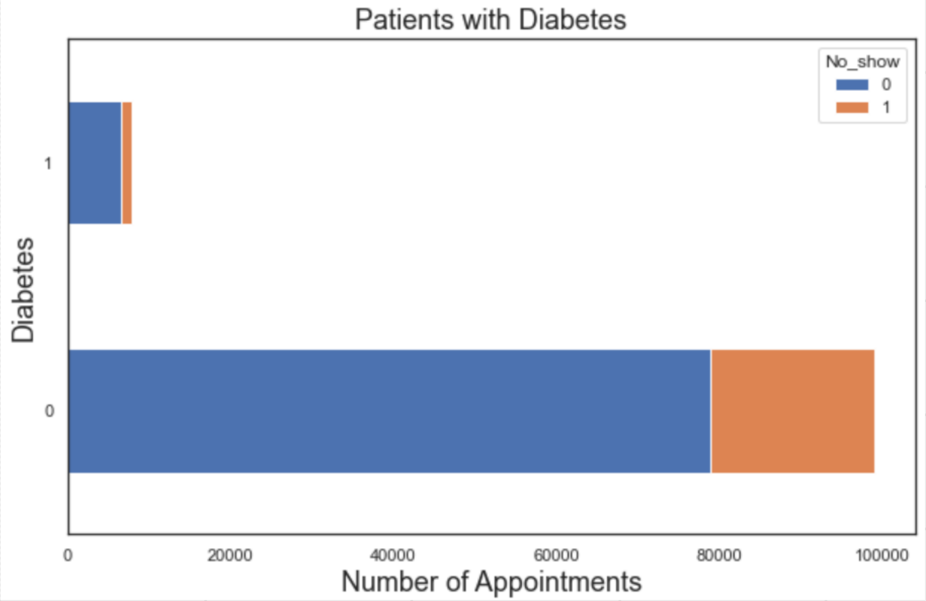
Percentage of patients who SHOWED versus NO SHOW

This figure shows that 79.7% of patients make it to their appointment and 20.3% didn't.

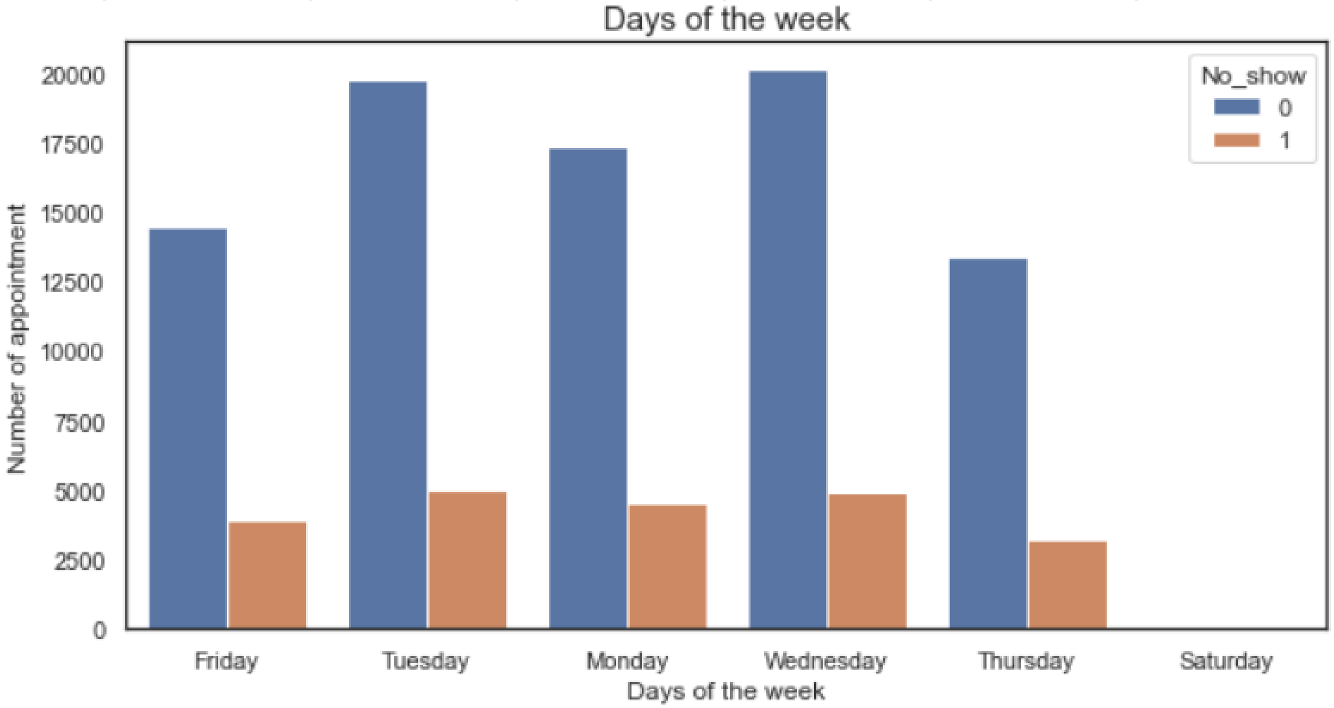


DATA ANALYST

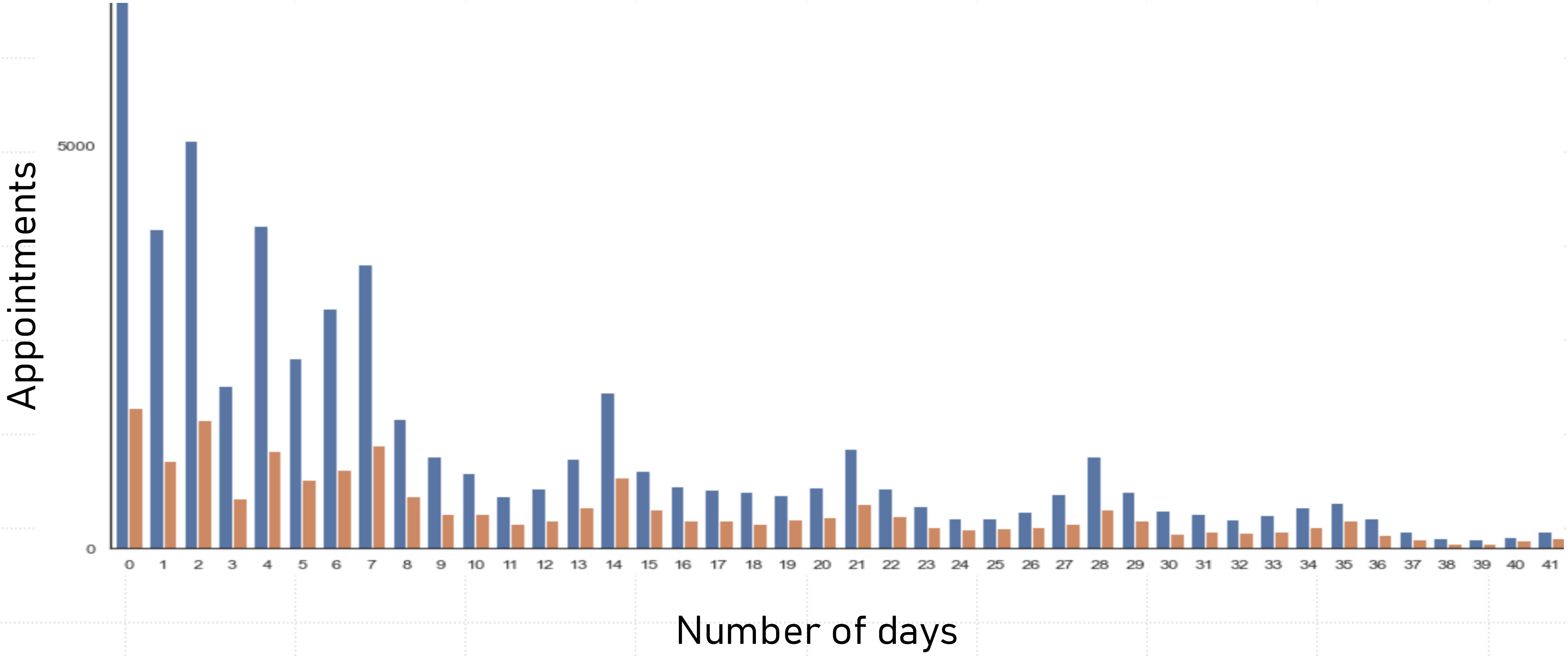
- The graph shows that patients with diabetes are more likely to make it to their appointment. that means patients with diabetes have a high chance to attend their appointment.



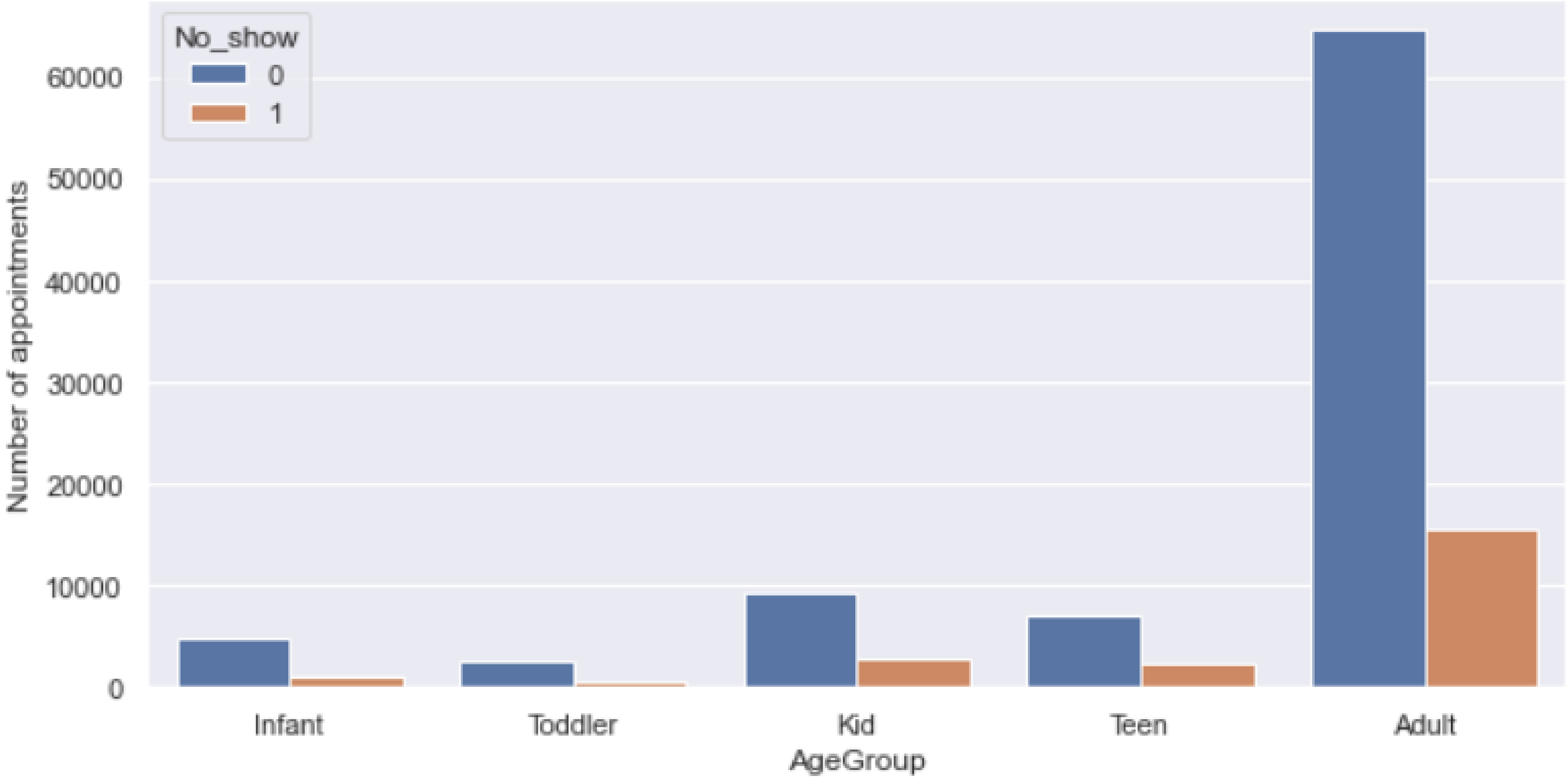
DATA ANALYST



DATA ANALYST

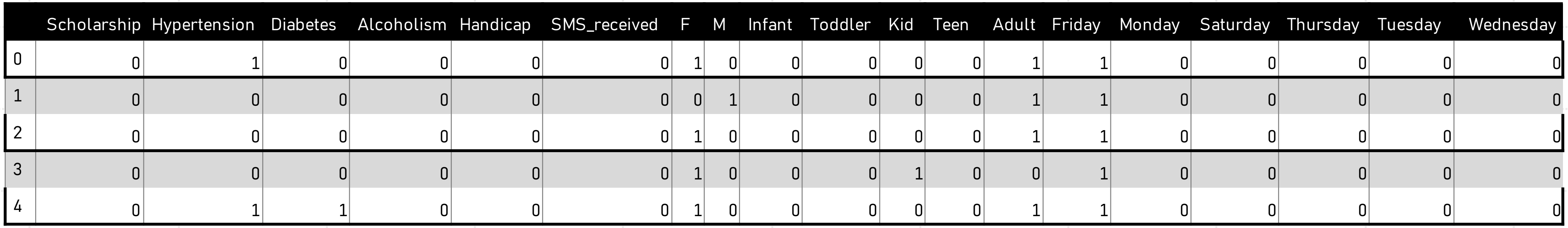


DATA ANALYST



DATA MODLING

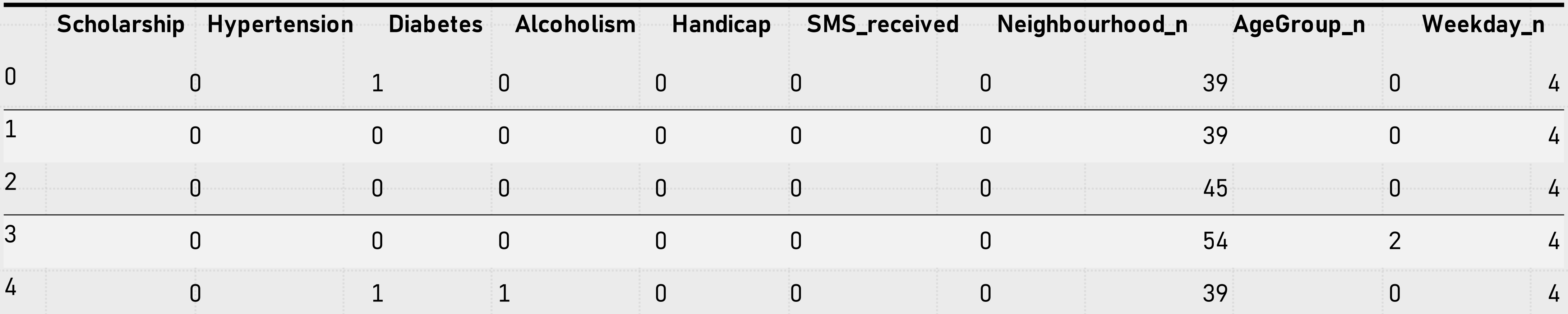
- **Logistic regression**, the categorical columns are extracted to binary by
  the get_dummies function. and the entire training dataset of 100,000 records was split into 80/20 train vs. Test. And the score is **0.80**



	Scholarship	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	F	M	Infant	Toddler	Kid	Teen	Adult	Friday	Monday	Saturday	Thursday	Tuesday	Wednesday
0	0	1	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0
1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0
2	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0
4	0	1	1	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0

DATA MODLING

- **Decision tree**
- we include neighborhood columns to enhance the score by using the LabelEncoder() function, and the score increased by 1%. And the score is **0.81**



	Scholarship	Hypertension	Diabetes	Alcoholism	Handicap	SMS_received	Neighbourhood_n	AgeGroup_n	Weekday_n
0	0	1	0	0	0	0	39	0	4
1	0	0	0	0	0	0	39	0	4
2	0	0	0	0	0	0	45	0	4
3	0	0	0	0	0	0	54	2	4
4	0	1	1	0	0	0	39	0	4

CONCLUSION

In conclusion, And after testing the models It seems that our models are always predicting that the patient will attend the appointment. Furthermore, the data was gathered in a **short time span**. The model could be improved if we added more features such as :

- **Forecast** factors like weather and temperature
- **Social factors** such as marital status and employment status
- Hospitals **location**
- The **clinic** name

Thanks