DATA MINING PROJECT

Submitted By:

Akhil (171210005)
Ankit (171210009)
Jayprakash Kumar (171210030)

Objective

- **Predicting whether the person would show up on their medical appointment(s).**

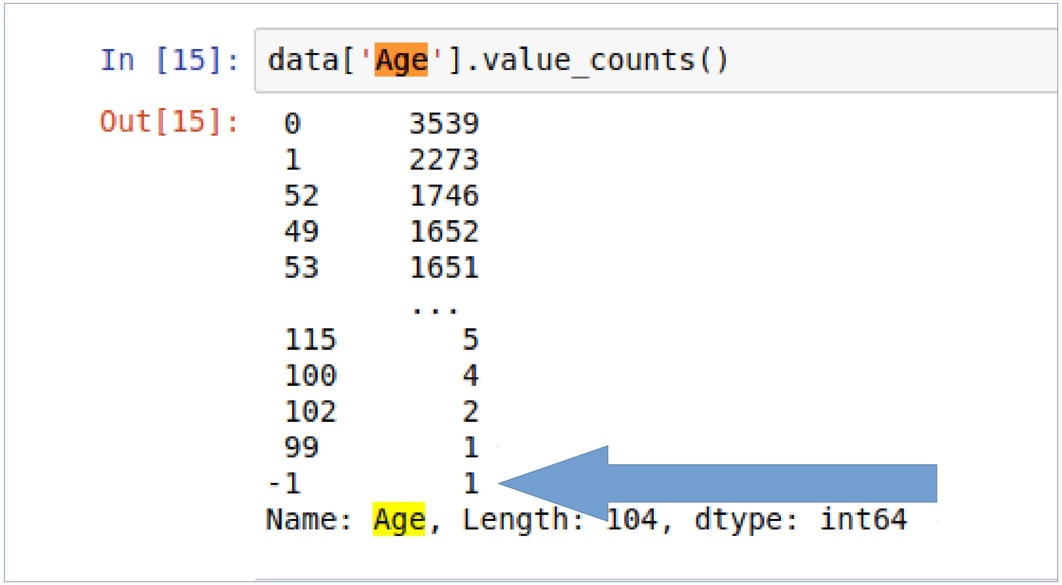
Dataset Used : Medical Appointments No Show

Link : https://www.kaggle.com/joniarroba/noshowappointments

Data Processing

1. Removing inconsistent values

- Negative values in age column
- Removing rows having “Appointment Day” before “Scheduled Day”



```
In [15]: data['Age'].value_counts()

Out[15]:  0      3539
          1      2273
          52     1746
          49     1652
          53     1651
                 ...
          115       5
          100       4
          102       2
          99        1
         -1         1
         Name: Age, Length: 104, dtype: int64
```

<u>Negative values in “Age” Column</u>

2. Encoding Genders, Neighbourhood and Days of the Week, No-Show

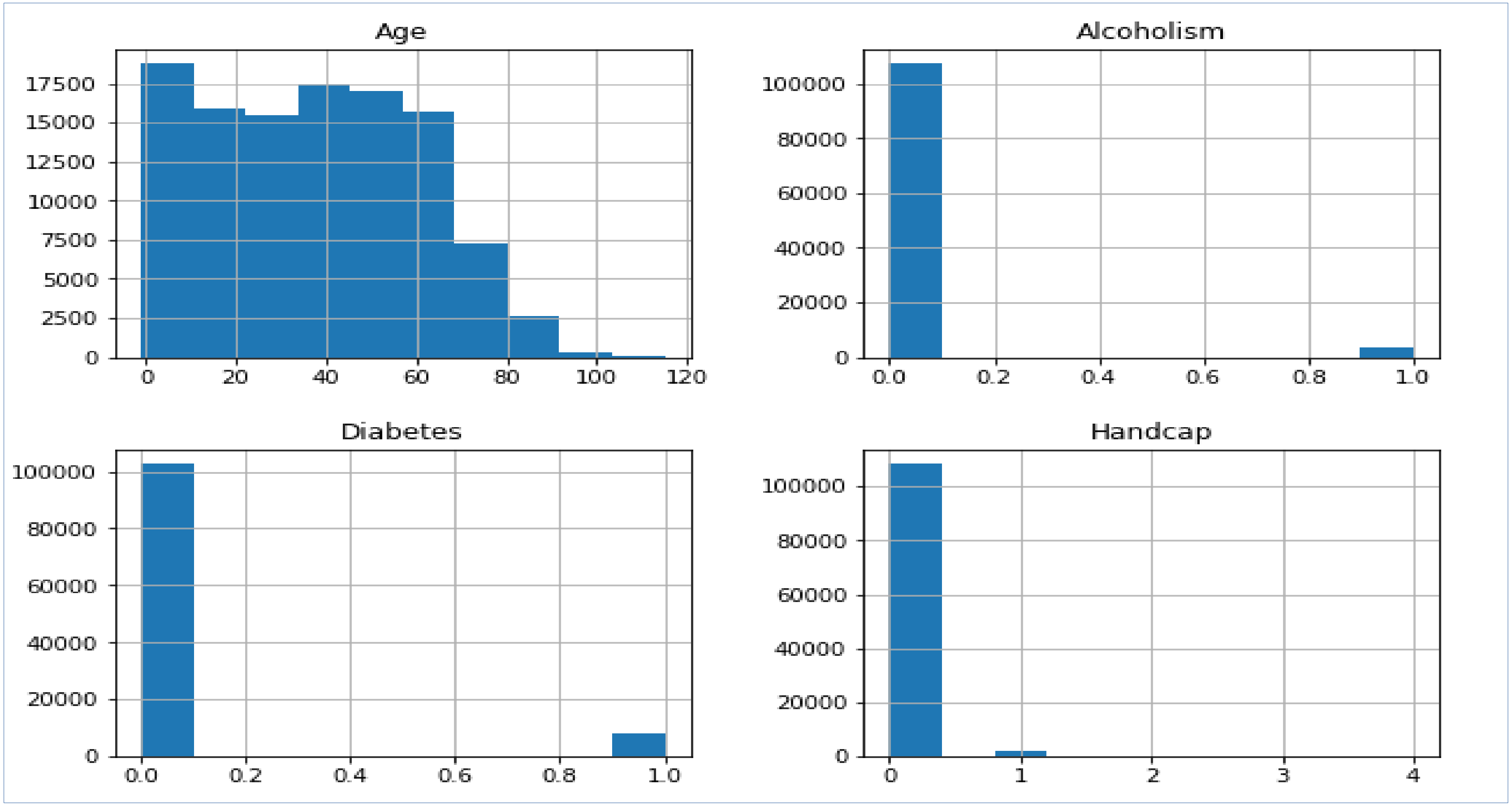
3. Adding a column Waiting Time(in days)

where Waiting Time = Appointement Day – Scheduled Day

4. Removed unnecessary columns like “Patient ID”, “Appointment ID”, “Scheduled Day” and “Appointment Day”

Visualising Data

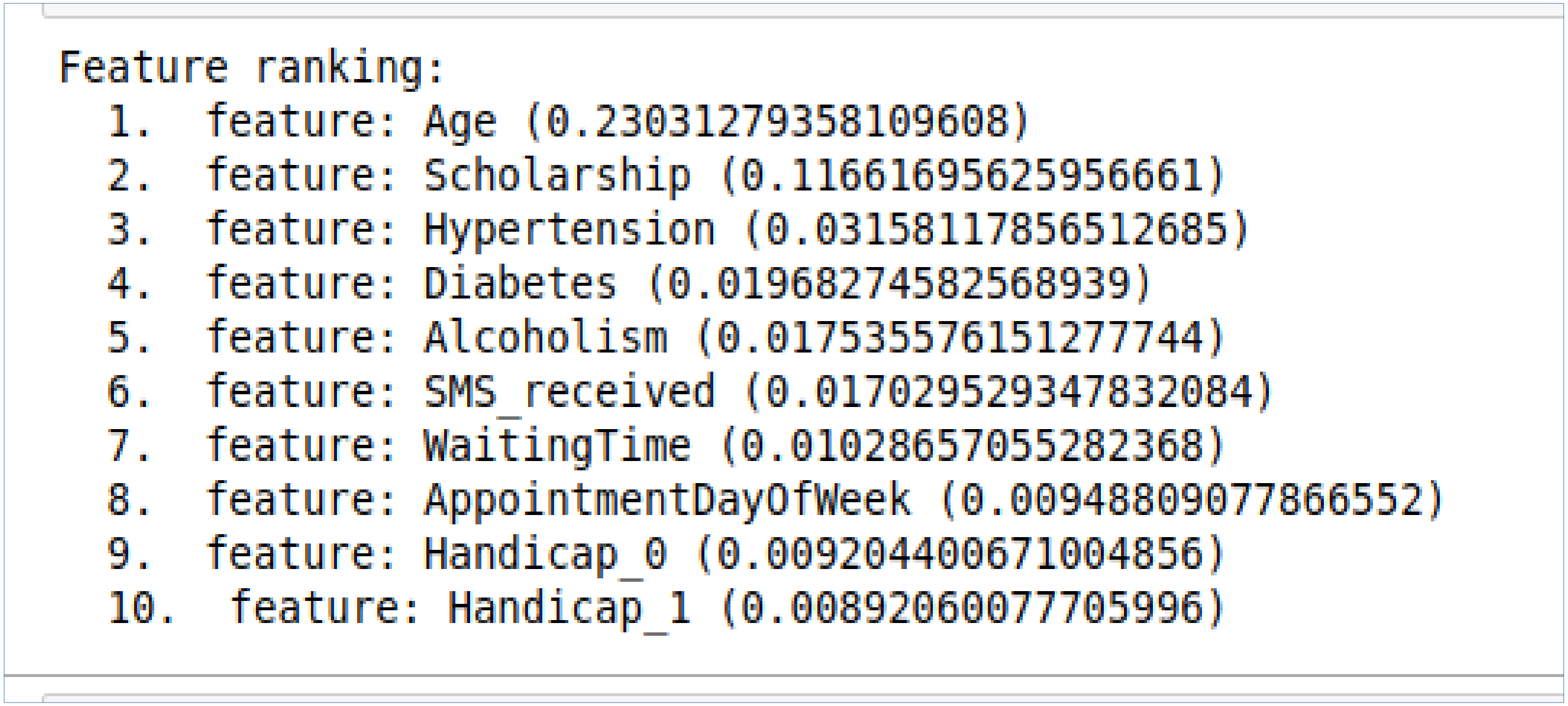
To know the dependency of target variable on different features



Histograms for different features

Choosing Optimal Algorithm

- As this was a Yes/No based classification problem
- It was best suited for a Decision Tree based classification.



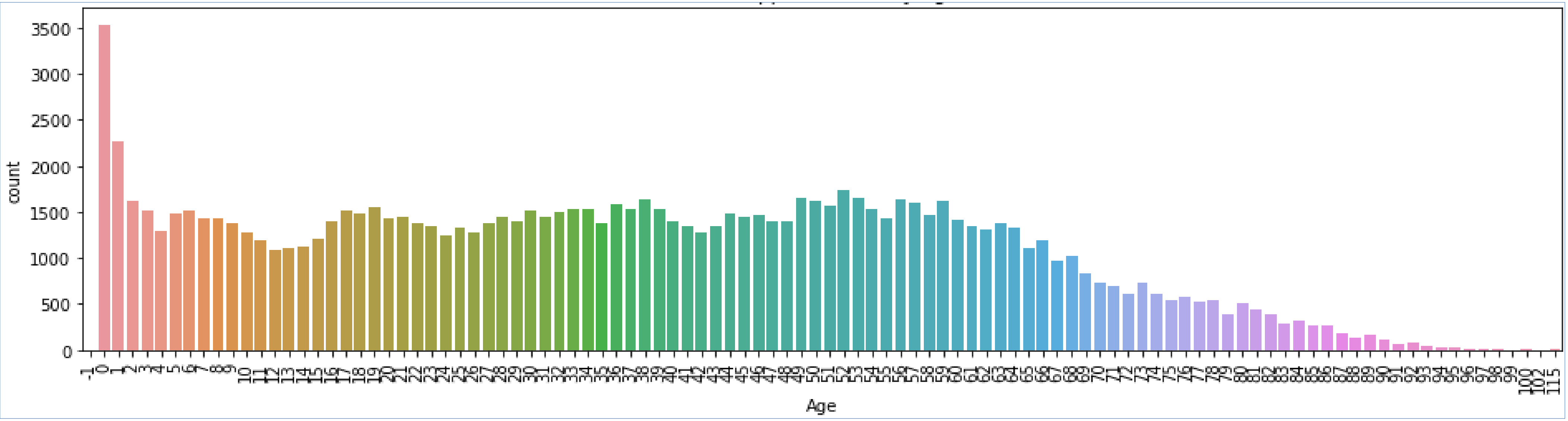
```
Feature ranking:
  1.  feature: Age (0.23031279358109608)
  2.  feature: Scholarship (0.11661695625956661)
  3.  feature: Hypertension (0.03158117856512685)
  4.  feature: Diabetes (0.019682745825689 39)
  5.  feature: Alcoholism (0.01753557615127744)
  6.  feature: SMS_received (0.01702952934783208 4)
  7.  feature: WaitingTime (0.01028657055282368)
  8.  feature: AppointmentDayOfWeek (0.009488090778665 52)
  9.  feature: Handicap_0 (0.00920440067100485 6)
  10.  feature: Handicap_1 (0.00892060077705996)
```

Ranking of features based on their importances

Conclusion

After obtaining the Decision Tree , we came to following conclusions:

- Showing Up of person depends a lot on their “Age”
- And it depends the least on the place where they live i.e Neighbourhood



Histogram Plot of Age of Persons

THANK YOU