# ANALYSIS ON PATIENT'S ATTENDANCE TO MEDICAL APPOINTMENT

IBNU ILYAS SYABAN NUGRAHA

# Table of Content

# Background

# Background


The data is collected from the Brazilian public health system known as SUS (*Sistema Unico de Saude* / Unified Health System), one of the largest health system in the world.

This institution representing government investment of more than 9% of GDP. However, its operation is not homogeneous and there are distinct perceptions of quality from citizens in different regions of the country, resulting several cancellation to medical appointment.

This analysis is carried out to understand the outline of the Brazilian public health system and build a machine learning model to predict the patient's attendance.

# Dataset

The dataset is from Kaggle with total of **110,527** row of data and **14** columns.

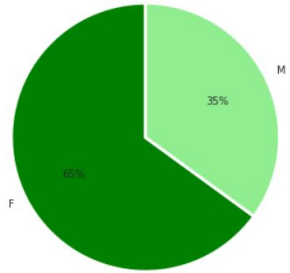
| Data Feature | Description |
|---|---|
| PatientId | Identification of a patient |
| AppointmentID | Identification of each appointment |
| Gender | Male or Female |
| ScheduledDay | The day someone called or registered the appointment, this is before appointment of course |
| AppointmentDay | The day of the actual appointment, when they have to visit the doctor |
| Age | How old is the patient |
| Neighbourhood | Location of the hospital |

| Data Feature | Description |
|---|---|
| Scholarship | Patient is enrolled in Brazilian welfare program or not (equivalent to BPJS) |
| Hipertension | Hypertension status of the patients |
| Diabetes | Diabetes status of the patients |
| Alcoholism | Alcoholism status of the patients |
| Handcap | Handicap status of the patients |
| SMS_received | Messages sent to patients |
| No-show | Attendance status of the patients (Yes means not attend) |

# Data Exploration

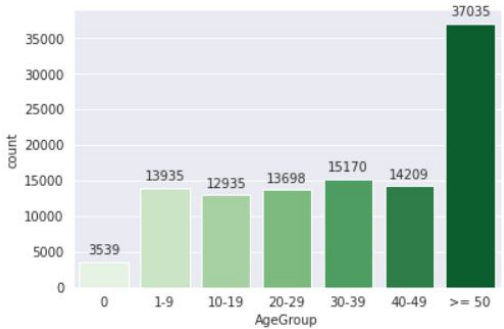
# Patient’s Profile

Gender



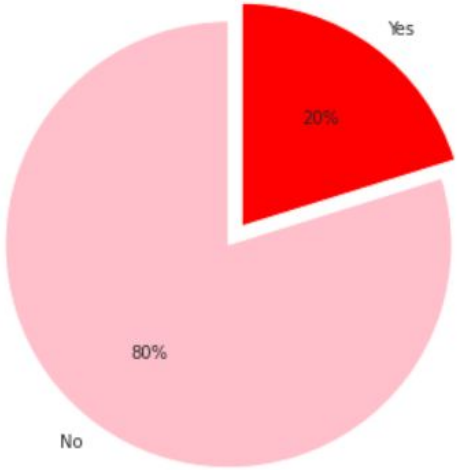

65%
of the patients are female.

Age




Majority of patients are
50 years old or older

No_Show

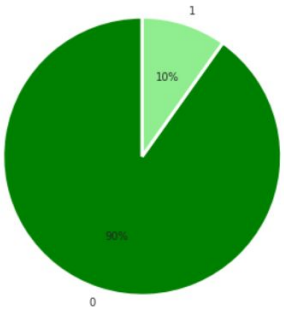



20%

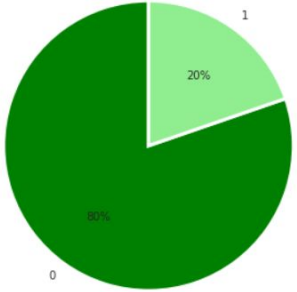
of patients that have scheduled appointment didn’t attend to the hospital.

# Patient’s Profile

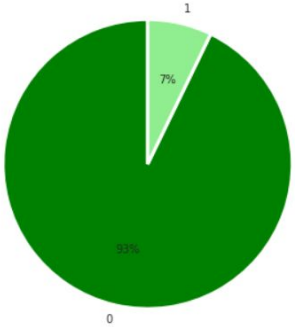
Scholarship



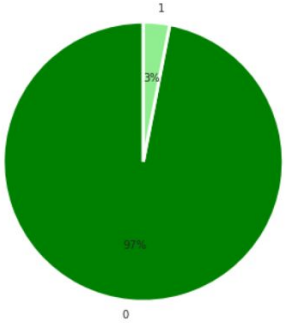

Hypertension



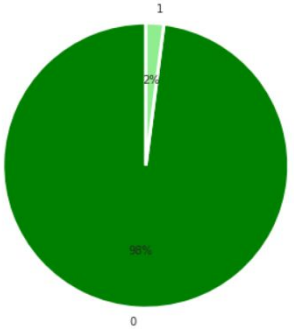

Diabetes




Alcoholism

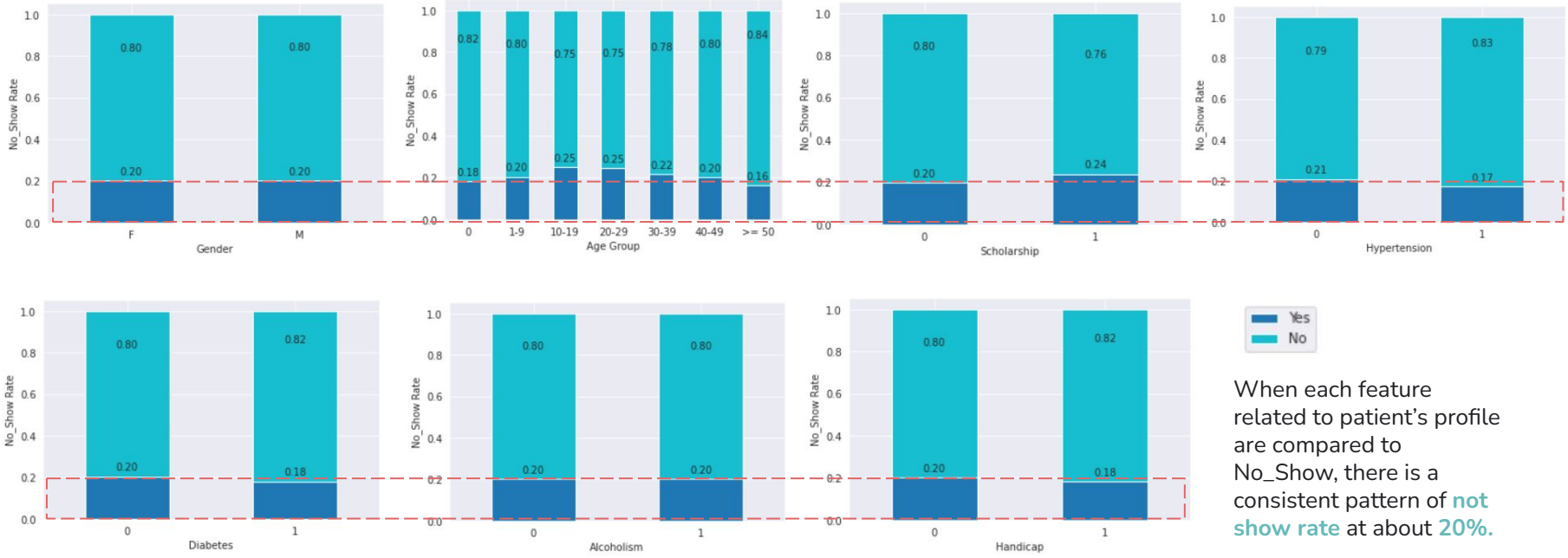



Handicap



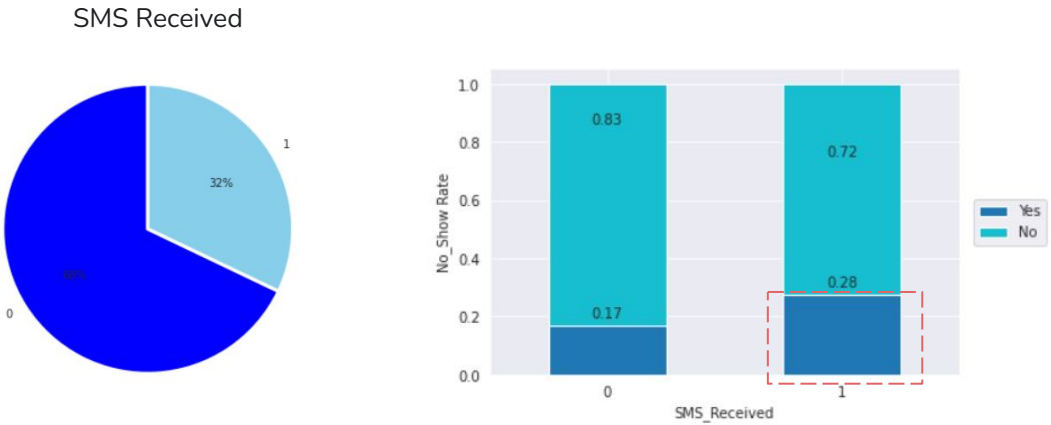

- Only **10%** of the patients have Scholarship.
- Regarding health condition, **80% - 90%** of the patients don’t have major health issues.

# Health Profile



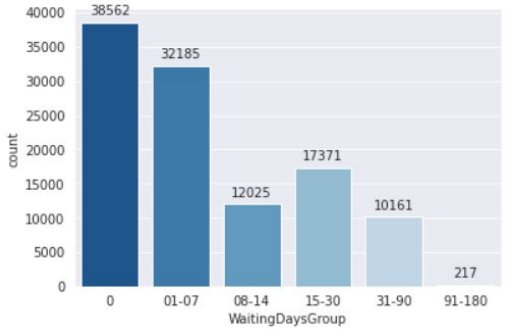

When each feature related to patient's profile are compared to No_Show, there is a consistent pattern of **not show rate** at about **20%.**

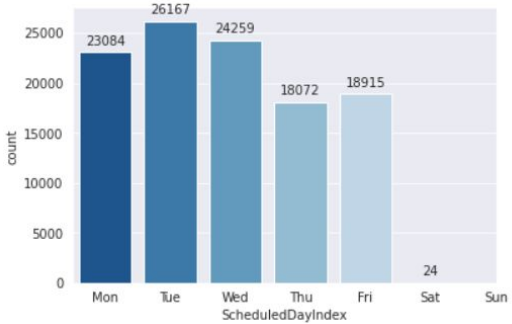
# SMS Received

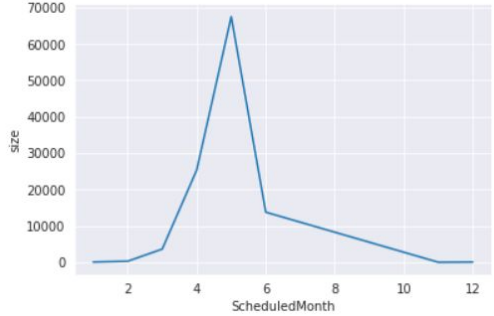


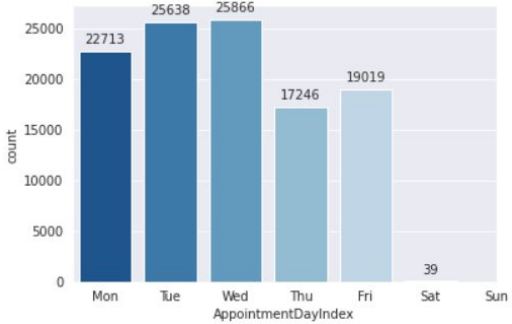

- Only **32%** of the patients received 1 SMS, while the rest don't receive at all.
- Strangely, the No_Show rate is higher on the patients that received SMS.
- There is no detail about the content of the message, but we will assume it's related to the appointment.

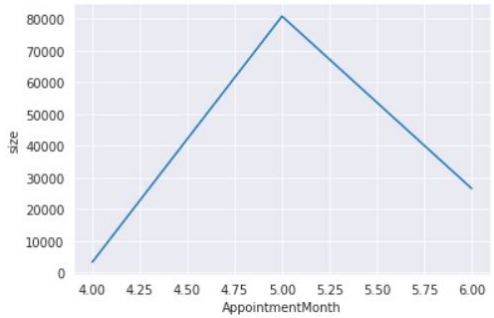
# Scheduling & Appointment



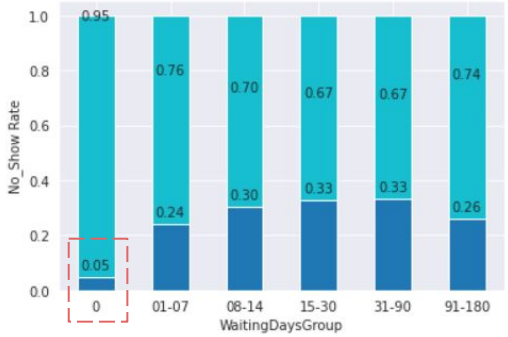



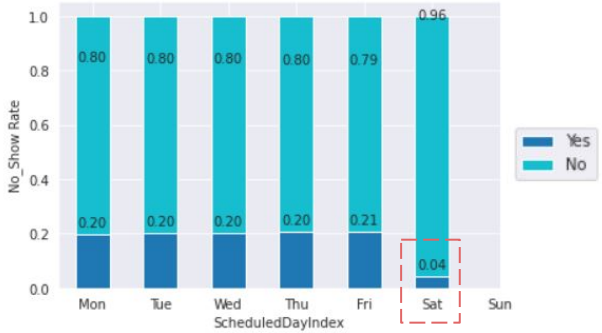



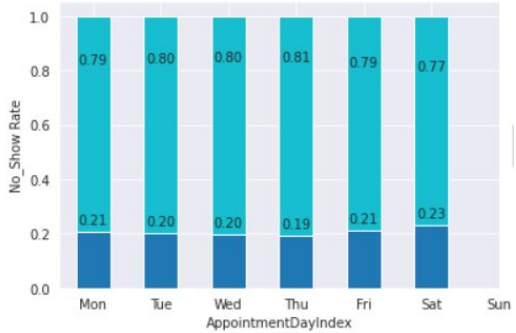




- Most of the appointments are **same day** appointment, followed by **same week** appointment.
- There are several appointments that are **3 - 6 months** long from the scheduling date.

- Almost all scheduling were done on **weekdays**. It's assumed that the hospital doesn't receive / open scheduling on weekends.
- **May** has the highest scheduling made.
- Those facts also happen to the appointment. This could be caused by the majority of same day appointment.




# Scheduling & Appointment

- **Same day** appointment has the smallest no-show rate at **5%,** while others have 245 - 33% No_Show rate.
- The longest waiting days (91-180 days) still has high ratio of attendance.
- Patients scheduled in Saturday has low No_Show rate, but doesn't count for the appointment day.

# Data Modelling

# Data Preprocessing

| Label Encoding | Column Manipulation | One Hot Encoding |
|---|---|---|
| Column:<br>● Gender<br>● No_Show<br><br>Encoded to have 0 or 1 value | Added features:<br>● WaitingDays: the gap between scheduling and appointment day.<br>● ScheduledDayIndex<br>● ScheduledMonth<br>● AppointmentDayIndex<br>● AppointmentMonth<br>● PatientCount: how many times patient makes appointment.<br><br>Dropped features:<br>● PatientID<br>● AppointmentID<br>● ScheduledDay<br>● AppointmentDay | Column:<br>● Neighbourhood<br><br>Total unique values = 81 |

# Modelling

| Model | Method | Model Evaluation |
| --- | --- | --- |
| • Logistic Regression<br>• Decision Tree<br>• Random Forest<br>• XGBoost<br><br>Those four models are chosen to be able to do classification (Yes / No) based on the features to predict No_Show. | • Baseline<br>• Undersampling<br>• Oversampling<br>• Oversampling SMOTE<br><br>Because the data is imbalanced (20:80), imbalanced handling should be done. | • Recall<br>• F1 Score<br><br>Imbalanced nature of the data and to focus on False Negative. |

# Model Evaluation

| index | Model | Method | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| 0 | Logistic Regression | baseline | 0.793802 | 0.331818 | 0.01628 | 0.031037 |
| 1 | Decision Tree | baseline | 0.734268 | 0.338297 | 0.324264 | 0.331132 |
| 2 | Random Forest | baseline | 0.796155 | 0.492848 | 0.169045 | 0.251744 |
| 3 | XGBoost | baseline | 0.79914 | 0.53481 | 0.075379 | 0.132134 |
| 0 | Logistic Regression | undersampling | 0.653291 | 0.309718 | 0.577163 | 0.403115 |
| 1 | Decision Tree | undersampling | 0.6171 | 0.292102 | 0.62355 | 0.397837 |
| 2 | Random Forest | undersampling | 0.624565 | 0.318799 | 0.748439 | 0.447139 |
| 3 | XGBoost | undersampling | 0.613255 | 0.319349 | 0.801293 | 0.456689 |
| 0 | Logistic Regression | oversampling | 0.653834 | 0.309935 | 0.576048 | 0.403027 |
| 1 | Decision Tree | oversampling | 0.723411 | 0.333877 | 0.365299 | 0.348882 |
| 2 | Random Forest | oversampling | 0.774214 | 0.420006 | 0.296833 | 0.347837 |
| 3 | XGBoost | oversampling | 0.631577 | 0.326967 | 0.771186 | 0.45923 |
| 0 | Logistic Regression | oversampling SMOTE | 0.781724 | 0.351351 | 0.089875 | 0.143136 |
| 1 | Decision Tree | oversampling SMOTE | 0.705225 | 0.314327 | 0.383586 | 0.34552 |
| 2 | Random Forest | oversampling SMOTE | 0.757747 | 0.381303 | 0.311998 | 0.343187 |
| 3 | XGBoost | oversampling SMOTE | 0.751821 | 0.37654 | 0.340767 | 0.357762 |
| 0 | Logistic Regression | undersampling - scaled | 0.660122 | 0.31198 | 0.560437 | 0.400829 |
| 1 | Decision Tree | undersampling - scaled | 0.615834 | 0.288026 | 0.60727 | 0.39073 |
| 2 | Random Forest | undersampling - scaled | 0.634427 | 0.322405 | 0.728145 | 0.446924 |
| 3 | XGBoost | undersampling - scaled | 0.62461 | 0.323459 | 0.779215 | 0.45715 |

The best model is **XGBoost** with **undersampling**, with **Recall Score of 80%**.

This means, out of positive value (No_Show = Yes), the model will predict 80% to be true.

# Model Evaluation

| | Model | Method | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|---|
| 3 | XGBoost | undersampling | 0.613255 | 0.319349 | 0.801293 | 0.456689 |
| 2 | Random Forest | undersampling | 0.624565 | 0.318799 | 0.748439 | 0.447139 |
| 0 | Logistic Regression | undersampling | 0.653291 | 0.309718 | 0.577163 | 0.403115 |
| 1 | Decision Tree | undersampling | 0.617100 | 0.292102 | 0.623550 | 0.397837 |

The best model is **XGBoost** with **undersampling**, with **Recall Score of 80%**.

This means, out of positive value (No_Show = Yes), the model will predict 80% to be true.

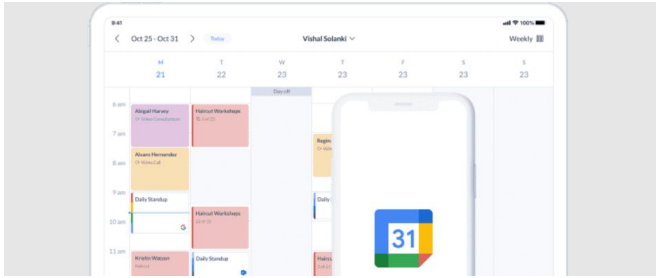
# Conclusion

# Conclusion

- Overall, the **No_Show** rate of patients is **20%**. This number is consistent if we look closely at each features related to patient's profile.
- The time of **scheduling**, **appointment**, and **waiting days** has higher chance in affecting patients to not attend.
- Beside same day, patient has waiting days ranging from 1 - 180 days. We should focus on these patients.

# Recommendation


Push notification regarding Appointment Schedule that gives reminder on timely basis.




Integrate appointment into Calendar App


Open for scheduling and appointment on weekends. But this should be analysed further with the calculation of additional labor and operational cost.