**Medical Appointments No Show**

**Karan Shah and Manali Killedar**

**(Team 12)**

1. **Domain & Motivation**

The medical/healthcare industry is moving rapidly towards analytics. Additionally, both of us have been biology student at some point; so we thought to explore a problem in this industry. People make appointments with doctors, follow all the instructions, receive messages reminding them of their appointments, and yet they still end up missing their appointments. What if it was possible to predict no-shows? If we are able to predict the show ups, we can optimize the usage of the resources and schedule maximum appointments in a day.

The domain of the problem is a pure binary classification task (supervised learning). To make it a little more interesting, we want to see how unsupervised learning can be implemented within a supervised learning task.

**II**. **Type of Learning**

There are three ways we can look at this. First, we can just split our dataset into train and test sets using K-fold cross validation and train our classifiers on the train sets and predict on the test sets. Second, we can cluster the dataset into two homogenous subgroups representing individuals that will show and individuals that will not, and we can create new labels for both the train and test sets. We can then train our classification models on only the cluster labels for the train sets and predict on only the cluster labels for the test sets. In this case, we will be training and predicting our classification model on the cluster labels rather than the dataset itself. Finally, instead of using cluster labels to train our classification model, we can add the cluster labels as features into our dataset, and train/predict our classification models on this dataset.

More specifically, some of the clustering algorithms we can explore include K-Means clustering, Agglomerative Hierarchical clustering, and Expectation Maximization (EM) clustering using Gaussian mixture models (GMM). Some of the classification algorithms we can explore include logistic regression, support vector machines, and random forests.

III. **Explanation of Inequality**

In terms of the inequality P(T, E+ΔE) > P(T,E), the task (T) is to use to determine if we can accurately predict individuals that miss their appointments. Additionally, we hope to identify common characteristics and features of such individuals as well. Experience (E) in performing this task is related directly to the attributes in the medical appointment no shows dataset. The gain in experience (ΔE) results from the clustering techniques and classification techniques which will help provide a boost in performance of the task. Performance (P) will the error rate in our classification task. Therefore, we plan to assess our work by obtaining the misclassification rate between the predicted label and the truth label. Moreover, since this is a classification task, we plan to create a confusion matrix along with a ROC curve to assess our different classifiers.

IV. **Data**

We found our dataset from kaggle datasets (<https://www.kaggle.com/joniarroba/noshowappointments>) . The data is for the year 2016 with about 300K records of patients scheduling appointments from Brazil. The data has some basic information of the person with his/her medical history at the time of booking an appointment. The various features are the appointment date, age, gender, neighborhood (location), whether the person is diabetic, whether he/she suffers from hypertension, whether the person is alcoholic, if he/she is handicapped. There is a feature which tells whether the person is entitled to the **Bolsa Família** scholarship which is given by the Government of Brazil to families with low-income. It also has the information whether the person received SMSes and if the person finally showed up for the scheduled appointment or he/she didn’t. So the data basically gives lot of relevant information of the person to classify whether a  future appointment will be completed.