

## EXECUTIVE SUMMARY

It is estimated that 9.3% of the population in the United States have diabetes mellitus (DM), 28% of which are undiagnosed. India is often called capital of the world with as many as 50 million people suffering from type-2 diabetes. The WHO also estimates that 80% of diabetes deaths occur in low and middle-income countries and projects that such deaths will double between 2020 and 2030. It has been further estimated that the global burden of type-2 diabetes is expected to increase to 438 million by 2030 from 285 million people (recorded in 2010). Similarly, for India this increase is estimated to be 58%, from 51 million people in 2010 to 87 million in 2030.

Diabetes is a chronic medical condition, that is, it can be curbed at the initial level by introducing lifestyle changes and controlled after its incidence through medicines in early stages and administration of external insulin in advanced stages. The chances of readmission of diabetic patients are quite high. Despite major advances in science and technology, diabetes continues to be a chronic disease, with a thirty-day readmission rate of around 20%, as compared to an average of 12% for the rest of the diseases. Additionally, readmissions cost hospitals a fair amount of money, so the end goal is to identify and reduce the possibility of a readmission. Prevention of patient readmission has been given a greater importance due to large cost involvement.

So in order to create a prediction system for readmission we studied US hospitals dataset. The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes.

The objective of this project is to predict whether the patient will be readmitted to the hospital or not. For predicting it most accurately we have used various prediction models like Logistic Regression, Decision Trees, Neural Networks, Random Forests, SVM and Naïve Bayes and seen their performance in order to get the best results.

The feature selection was done on the basis of PCA and the correlation analysis. Data was converted and pre-processed by removing null values and changing categorical variables.

The EDA gave us some interesting insights as the readmission is more in female people and in the Caucasian race. The overall comparison of methods on the basis of their accuracy sensitivity and specificity was generated as the output.