**DTSC701/CSCI 636**

**INTRODUCTION TO BIG DATA**

**Group project: "Analyzing Diabetes Trends: A Comprehensive Study of Hospital Data (1999-2008) for Readmission Prediction".**

**A close-up of a doctor's hands holding a glucometer

Description automatically generated**

**Group members: Professor:**

* **Oksana Dura 1316268 Liangwen Wu**
* **Chinasa Okonkwo 1319774**
* **Akena Hutchinson 1269747**
* **Krishnarjuna Reddy Acholu 1319435**

**Table of Contents**

[***Abstract: 3***](#_heading=h.gjdgxs)

[**Snippet of the dataset: 4**](#_heading=h.30j0zll)

[**Data Dictionary 4**](#_heading=h.1fob9te)

[***Development of the project (Setting up the environment): 7***](#_heading=h.3znysh7)

[**Data Lake vs. Data warehouse 7**](#_heading=h.2et92p0)

[**After choosing Data Lake the next steps require setting up EMR and S3. To successfully execute the program please follow the steps carefully: 8**](#_heading=h.tyjcwt)

[**Create an EMR Cluster: 9**](#_heading=h.3dy6vkm)

[**Go to the AWS S3 service. Create a bucket. 13**](#_heading=h.1t3h5sf)

[***Preprocessing the data with a spark application: 13***](#_heading=h.4d34og8)

[***Decision tree model 17***](#_heading=h.2s8eyo1)

# Abstract:

This group project focuses on the comprehensive analysis of a dataset spanning the years 1999 to 2008, encompassing information on diabetes across 130 hospitals. The dataset comprises 47 features, 101,766 instances, and presents challenges with missing values. The primary objective of our study is to uncover insightful patterns within the dataset. To achieve this, we engage in meticulous data preprocessing and subsequently employ a classification algorithm. Our specific aim is to predict the likelihood of patient readmission based on various data attributes, including but not limited to age, race, and medical history. Through this investigation, we seek to contribute valuable insights to the understanding of factors influencing diabetes-related hospital readmissions.

## Snippet of the dataset:

A screenshot of a table

Description automatically generated

Original source of the data set [https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Farchive.ics.uci.edu%2Fml%2Fdatasets%2FDiabetes%2B130-US%2Bhospitals%2Bfor%2Byears%2B1999-2008)

## Data Dictionary

* encounter\_id: Unique identifier of an encounter.
* patient\_nbr: Unique identifier of a patient.
* race: Race. Values are Caucasian, Asian, African American, Hispanic, and other.
* gender: Gender. Values: male, female, and unknown/invalid.
* age: Age Grouped in 10-year intervals: [0, 10), [10, 20),..., [90, 100).
* weight: Weight in pounds.
* admission\_type\_id: Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available.
* discharge\_disposition\_id: Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available.
* admission\_source\_id: Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital.
* time\_in\_hospital: Integer number of days between admission and discharge.
* payer\_code: Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay.
* medical\_specialty: Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon.
* num\_lab\_procedures: Number of lab tests performed during the encounter.
* num\_procedures: Number of procedures (other than lab tests) performed during the encounter.
* num\_medications: Number of distinct generic names administered during the encounter.
* number\_outpatient: Number of outpatient visits of the patient in the year preceding the encounter.
* number\_emergency: Number of emergency visits of the patient in the year preceding the encounter.
* number\_inpatient: Number of inpatient visits of the patient in the year preceding the encounter.
* diag\_1: The primary diagnosis (coded as first three digits of ICD9); 848 distinct values.
* diag\_2: Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values.
* diag\_3: Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values.
* number\_diagnoses: Number of diagnoses entered to the system.
* max\_glu\_serum. Indicates the range of the result or if the test was not taken. Values: >200, >300, normal, and none if not measured.
* A1Cresult: Indicates the range of the result or if the test was not taken. Values: >8 if the result was greater than 8%, >7 if the result was greater than 7% but less than 8%, normal if the result was less than 7%, and none if not measured.
* metformin: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* repaglinide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* nateglinide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* chlorpropamide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* glimepiride: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* acetohexamide. The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* glipizide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* glyburide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* tolbutamide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* pioglitazone: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* rosiglitazone: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* acarbose: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* miglitol:The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* troglitazone: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* tolazamide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* examide: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* citoglipton: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* insulin: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* glyburide-metformin: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* glipizide-metformin: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* glimepiride-pioglitazone: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* metformin-rosiglitazone. The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* metformin-pioglitazone: The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: up if the dosage was increased during the encounter, down if the dosage was decreased, steady if the dosage did not change, and no if the drug was not prescribed.
* change: Indicates if there was a change in diabetic medications (either dosage or generic name). Values: change and no change.
* diabetesMed: Indicates if there was any diabetic medication prescribed. Values: yes and no.
* readmitted: Days to inpatient readmission. Values: <30 if the patient was readmitted in less than 30 days, >30 if the patient was readmitted in more than 30 days, and No for no record of readmission

In this database, we have 3 different outputs:

* No readmission.
* A readmission in less than 30 days.
* A readmission in more than 30 days.

The purpose of this study is to find out indicators that a patient will be readmitted or not.

# Development of the project (Setting up the environment):

## Data Lake vs. Data warehouse

In our project focused on diabetes data analysis across 130 hospitals from 1999 to 2008, the choice of a data lake over a data warehouse is deliberate and strategic. Here's why:

* Data Diversity: Our dataset encompasses varied data types and formats. A data lake's schema flexibility allows us to store raw, unprocessed data without rigid structures.
* Scalability: Dealing with extensive datasets spanning multiple years and hospitals demands scalability. Data lakes, built on distributed storage, offer a cost-efficient solution for handling large volumes of data.
* Cost Efficiency: Storing raw data in its native format reduces preprocessing requirements, potentially lowering storage costs compared to traditional data warehouses.
* Advanced Analytics: Enabling direct access to raw data, data lakes facilitate advanced analytics and machine learning, crucial for our goal of uncovering patterns and predicting readmissions.

## After choosing Data Lake the next steps require setting up EMR and S3. To successfully execute the program please follow the steps carefully:

**AWS Account:**

* Ensure you have an AWS account with the necessary permissions to create and manage EMR clusters.

A screenshot of a computer

Description automatically generated

## Create an EMR Cluster:

Choose the following configurations:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

In Hardware Configuration for Instance type: m5.xlarge, Number of instance:

3

A screenshot of a computer

Description automatically generated

Decide how long your cluster with run for:

A screenshot of a computer

Description automatically generated

Create SSH key or use one of the options

A screenshot of a computer

Description automatically generated

Choose the service role:

A screenshot of a service

Description automatically generated

Choose the Instance Profile:

A screenshot of a computer

Description automatically generated

Create a cluster:

A screenshot of a computer

Description automatically generated

The cluster was created:

A screenshot of a computer

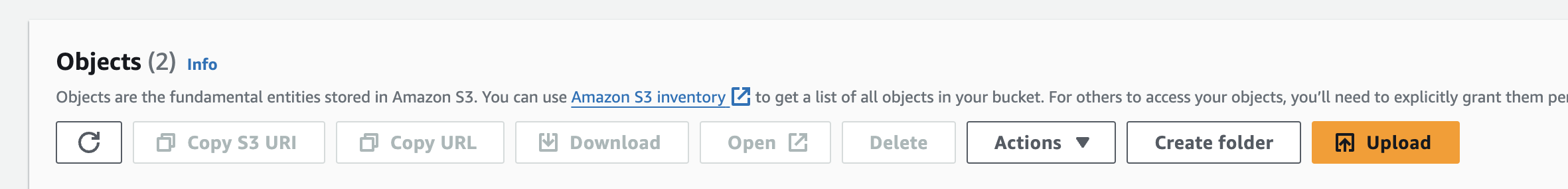
Description automatically generated

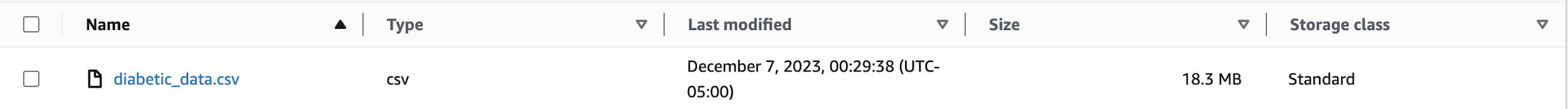
## Go to the AWS S3 service. Create a bucket.

A screenshot of a computer

Description automatically generated

Upload the csv file:



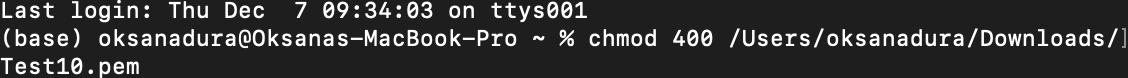


# Preprocessing the data with a spark application:

(The project was worked on a MAC terminal the commands might change for Windows or Linux operating system)

Open the terminal and connect to the cluster through the following commands:

Make sure the key has the right permissions. If key file has restrictive permissions, you might need to run chmod 400



**SSH into the Master Node:**

After creating an EMR cluster, obtain the public DNS or IP address of the master node from the AWS EMR console.

Use SSH to connect to the master node

Then run the following command on the terminal:

vi preprocessing.py

Paste the preprocessing code here. First press I in the keyboard to be able to input data. Then to exit press esc and after :wq enter. Once you are out press the following command on your terminal:

spark-submit preprocessing.py --packages org.apache.hadoop:hadoop-aws:3.3.0

One the program is successful excecated the preprocessing file will be saved on S3 in the output folder.



A screenshot of a computer

Description automatically generated

High-level application history

A screenshot of a computer

Description automatically generated

Preprocessing Spark Application code explained.

This Python script utilizes PySpark, a Python library for Apache Spark, to preprocess and save a dataset related to diabetes.

* **Importing Libraries:**

The script begins by importing necessary PySpark and machine learning libraries.

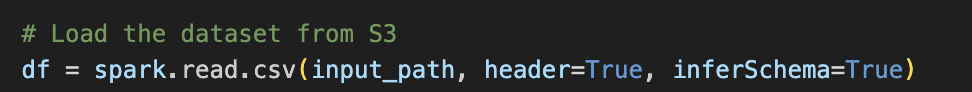
A screen shot of a computer screen

Description automatically generated

* **Spark Session Setup:**

It creates a Spark session named "DatasetPreprocessing" to interact with the Spark cluster.

* **Loading the Dataset:**



* **Data Cleaning:**

A screen shot of a computer program

Description automatically generated

* **Diagnosis Columns Preprocessing:**

A screen shot of a computer code

Description automatically generated

* **Categorical Column Encoding**

A screen shot of a computer program

Description automatically generated

**Pipeline Creation / Execution**

A screen shot of a computer program

Description automatically generated

* **Error Handling:**

A screen shot of a computer program

Description automatically generated

Decision tree model (The model was executed on Google Collab)

Preparing the data for testing and training:

A computer screen shot of a program code

Description automatically generated

# Define the hyperparameters to search through for Decision Tree

A black background with text and numbers

Description automatically generated

# Initialize variables to store the best parameters and accuracy

A black screen with white text

Description automatically generated

# Loop through hyperparameters and train Decision Tree models

A screen shot of a computer

Description automatically generated

# Train the best Decision Tree model

A screen shot of a computer program

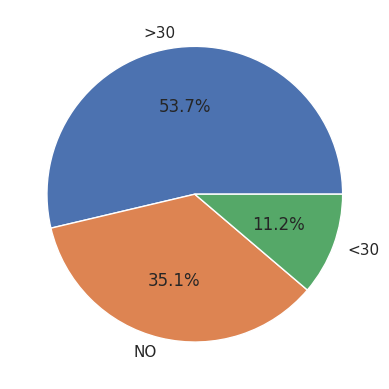
Description automatically generated

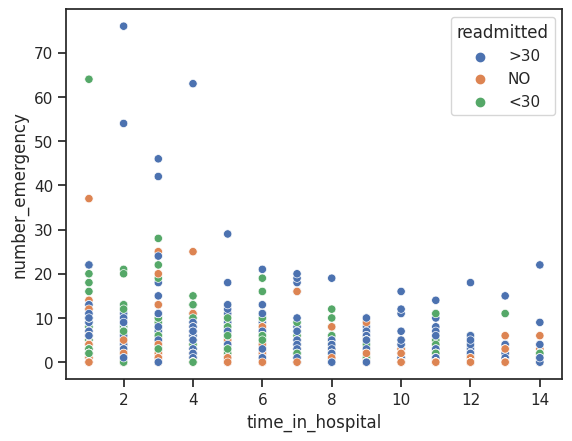
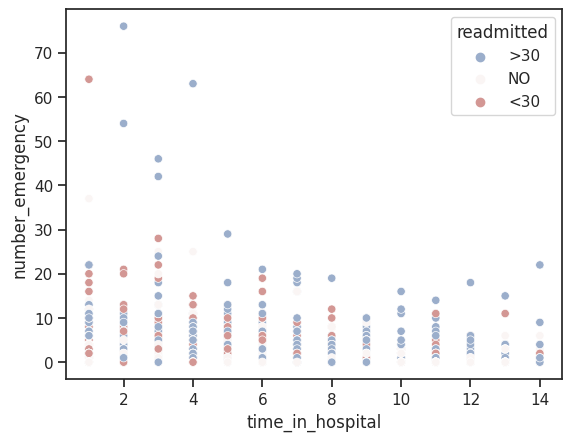
Output:

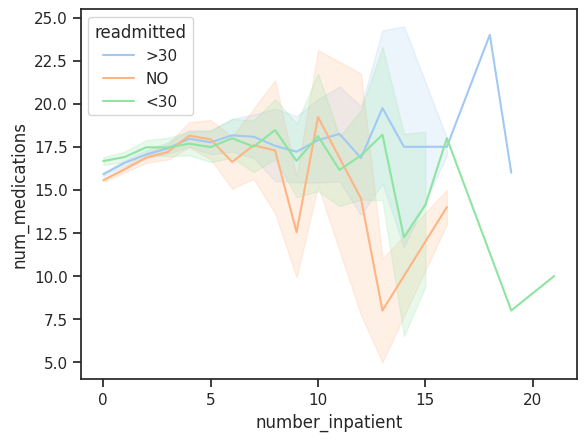
A screenshot of a computer

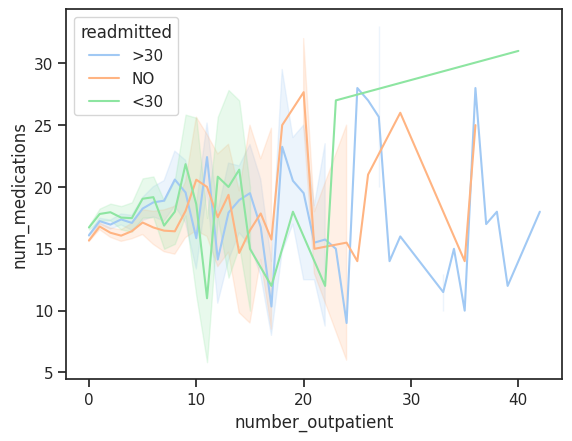
Description automatically generated

# Graphical Analysis:

Percentage of patients who have been readmitted within 30 days or more

Time spent in the hospital by the number of emergency incidents. Each item is further separated by if or when the patient was readmitted to the hospital. 

Number of medications given to inpatients with a separation on when or if they have been readmitted to the hospital

Number of medications given to outpatients with a separation on when or if they have been readmitted to the hospital

# Conclusion:

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion Matrix: | | Actual Values | |
| Positive | Negative |
| Predicted Values | Positive | 5560 | 2578 |
| Negative | 3143 | 3480 |

Accuracy: = = 0.6124 = 61%

Error Rate: = = 0.388 = 39%

Sensitivity: = = 0.638 = 64%

Specificity: = = 0.574 = 57%

Our prediction accuracy is about 61%. With a sensitivity rate of 64% and a specificity rate of 57% we can conclude that the current treatment plan may not be the most effective treatment to be given to the patients. More often than not patients are being readmitted into the hospital. Doctors may either have to consider outside factors such as diet and exercise into treatment as well as medication.