PYTHON FOR DATA ANALYSIS

Diabetes 130-US hospitals for years 1999-2008

Oscar Pastural - Richard Goudelin - Capucine Boudin

DEFINITION OF THE DATASET

The dataset provided consists of hospital records from 130 US hospitals and integrated delivery networks, covering a period of ten years from 1999 to 2008. The records pertain to patients who were diagnosed with diabetes and include information on laboratory tests, medications, and hospital stays of up to 14 days..

Content

Ten years (1999-2008) of clinical care at 130 US hospital

Patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days.

Shape

- Feature type : Multivariate
- Instances : 101766
- Features : 47

SUMMARY

PROBLEM DEFINITION

How can we predict the readmission status of a patient ?

How does the problem fit in our context ?

•The total annual cost of diabetes in 2022 is $412.9 billion

•9.3% of the population in the United States have diabetes

Predicting the readmission status involves identifying the key parameters that play a significant role and reduce complications associated with diabetes.

DATA PREPROCESSING

Data cleaning

•Removing :
- Features with a high percentage of missing values ('weight', 'medical_specialty', 'payer_code','max_glu_serum','A1Cresult')
-Redudancies (29423 patient_nbr)
-Useless columns ('encounter_id', 'patient_nbr', 'payer_code')
- 3 missing values in the gender column
-Drugs with zero variance àfeature selection based on a variance threshold at 0.05. (11 features are kept, 15 are deleted)

Encoding data

•Transformed non-numerical columns into a data_encoded dataframe :
['gender', 'admission_type_id', 'discharge_disposition_id', 'admission_source_id', 'diag_1', 'diag_2', 'diag_3', 'metformin', 'repaglinide', 'glimepiride', 'glipizide', 'glyburide', 'pioglitazone', 'rosiglitazone', 'insulin', 'change', 'diabetesMed']

Mapping the data

•Variables mapped by following the IDS mapping providedwith the dataset :
-admission_type_id
-admission_source_id
-discharge_disposition_id

•Variables mapped by information foundallong the dataset :
-diag_1, diag_2 and diag_3

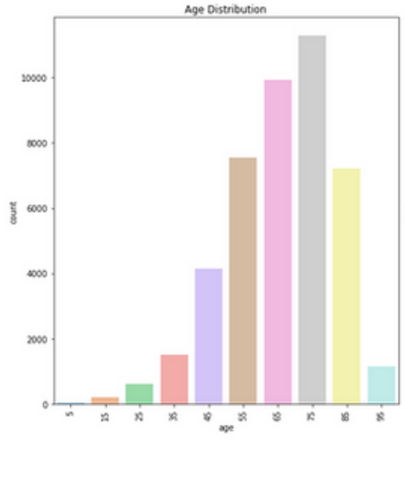
Figure 3

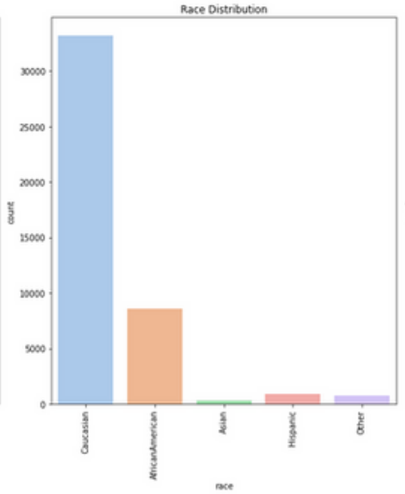
Predicted (adjusted for covariates) readmission rates by the primary diagnosis and HbA1c measurement. Readmission rates were predicted on the reference values of other predictors and the mean value of time in hospital (Table 3). The error bars represent the 95% confidence intervals for the predicted values. The following abbreviations are used for particular icd9 codes: "circulatory" for icd9: 390–459, 785, "digestive" for icd9: 520–579, 787, "genitourinary" for icd9: 580–629, 788, "diabetes" for icd9: 250.xx, "injury" for icd9: 800–999, "musculoskeletal" for icd9: 710–739, "neoplasms" for icd9: 140–239, "respiratory'' for icd9: 460–519, 786, and "other" for otherwise.

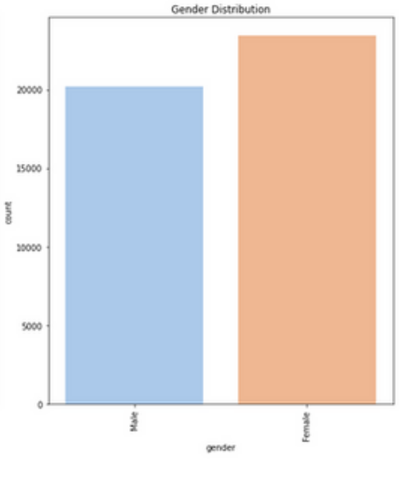
```
replaceDict = {'[0-10)' : 5,
'[10-20)' : 15,
'[20-30)' : 25,
'[30-40)' : 35,
'[40-50)' : 45,
'[50-60)' : 55,
'[60-70)' : 65,
'[70-80)' : 75,
'[80-90)' : 85,
'[90-100)' : 95}

data['age'] = data['age'].apply(lambda x : replaceDict[x])
```

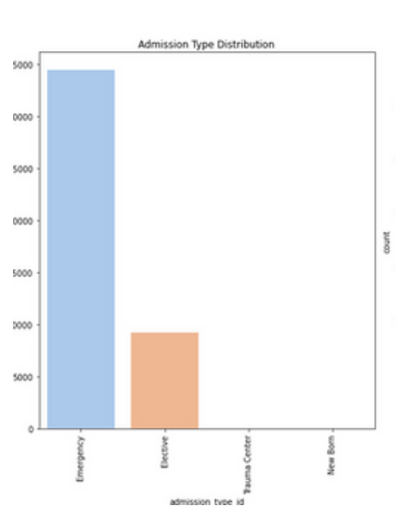
Re-Writing the data

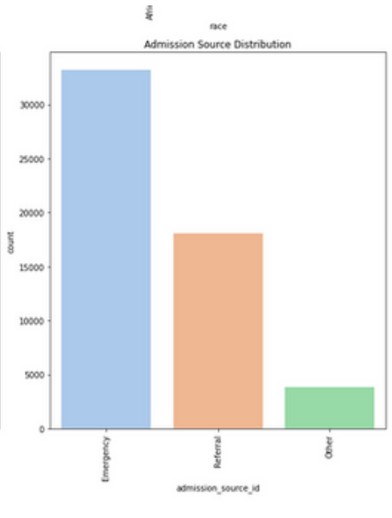
•The age was given as classes such as [0–10),[10–20),[20–30) etc…
•Weconvertedthiscolumn to a numericfeature as the following :

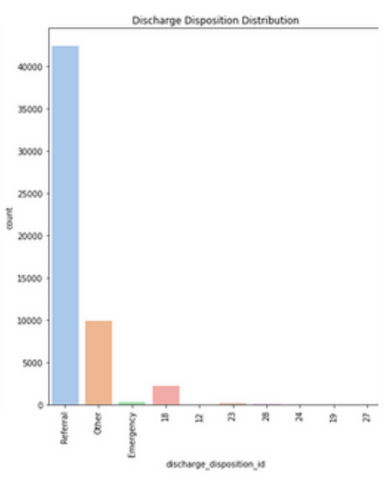


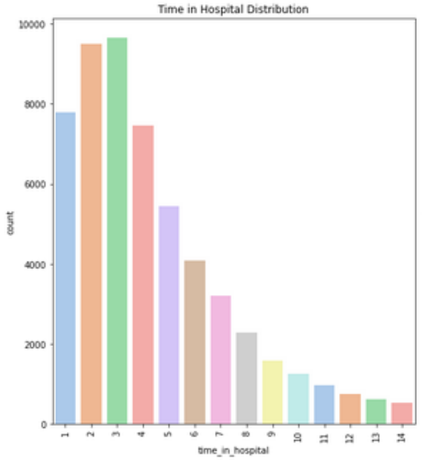


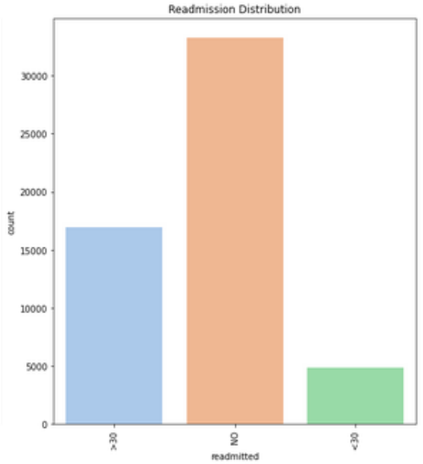


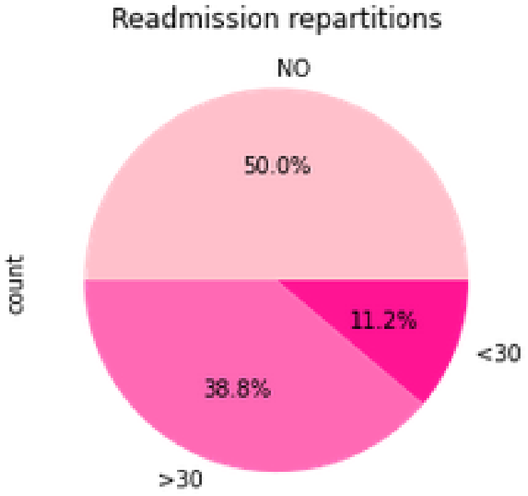


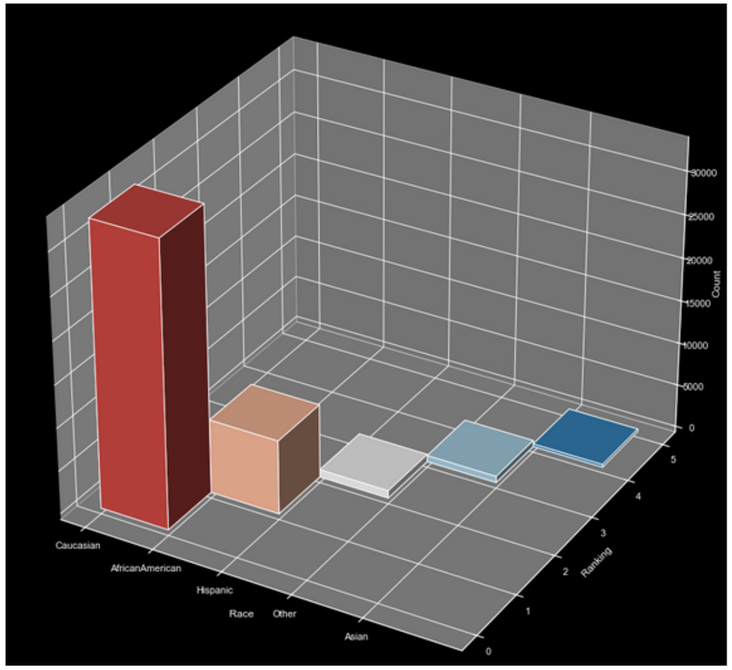






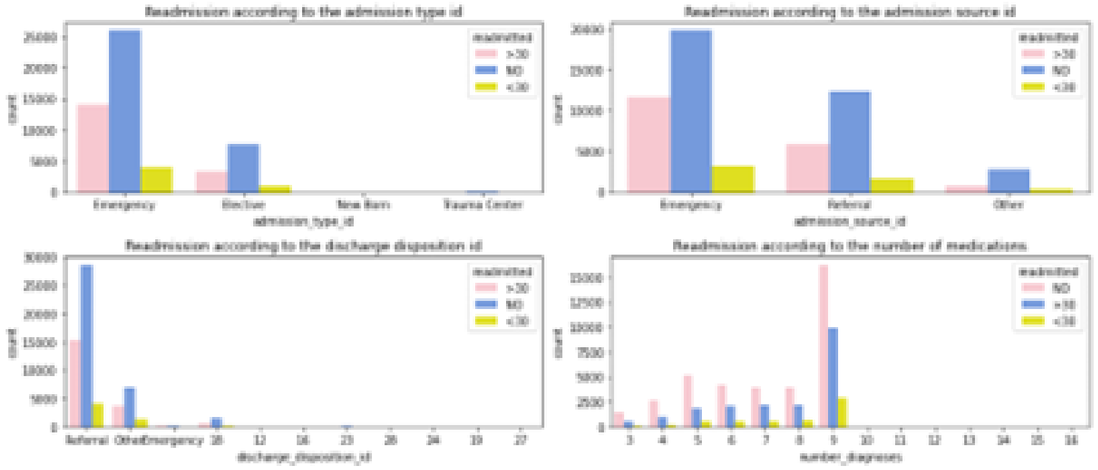






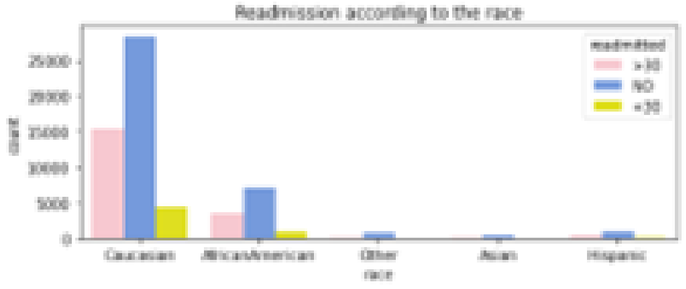
EXPLORING PLOTS

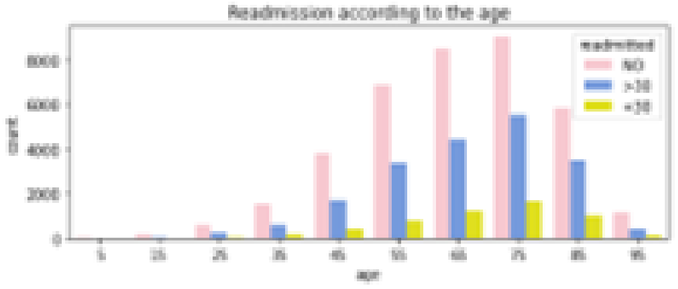
Parameters overview

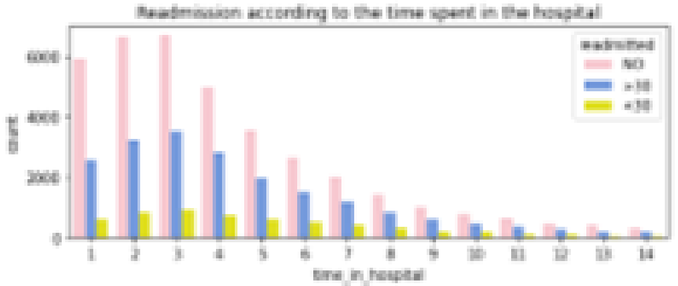


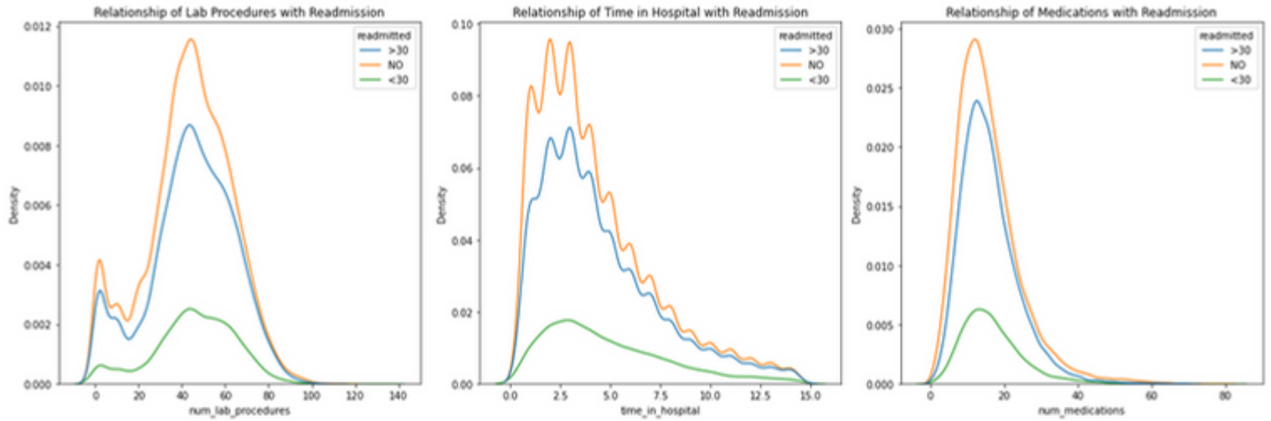
CORRELATIONSWITH THE READMISSION DISTRIBUTION IN TIME

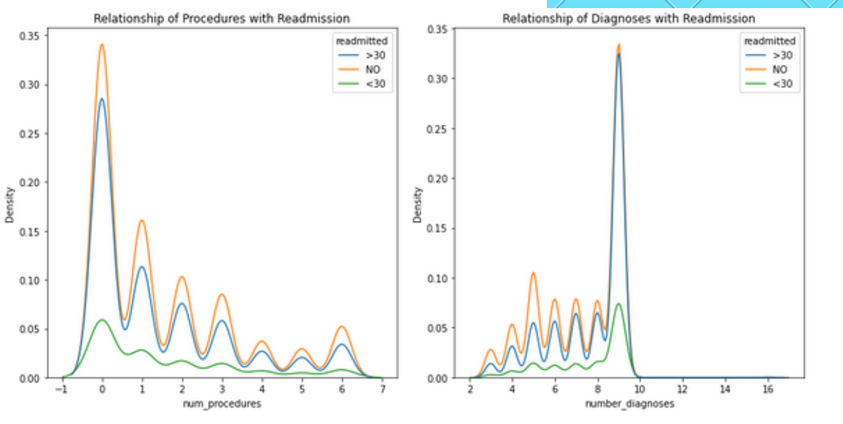
•We now consider the parameters in correlation with the readmission within more or less than 30 days or no readmission at all

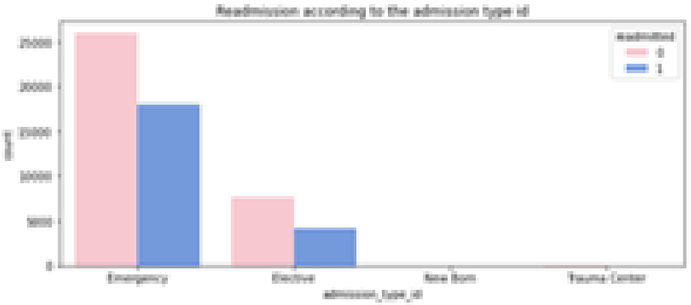


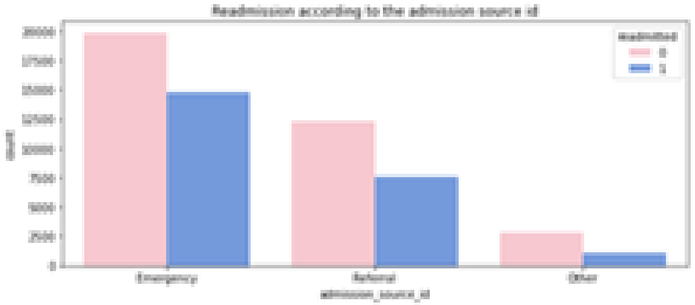


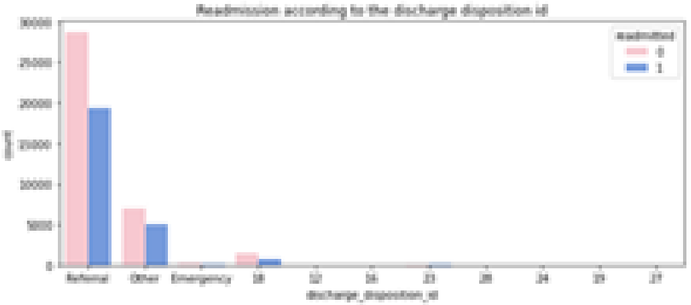


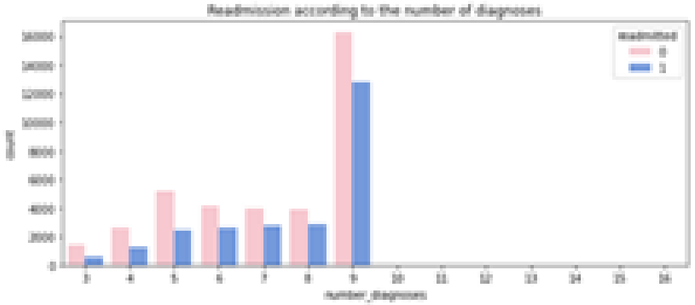






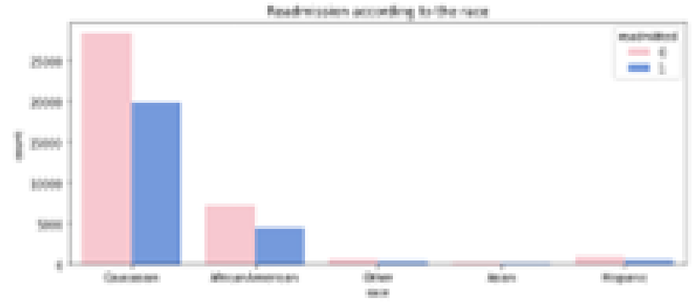


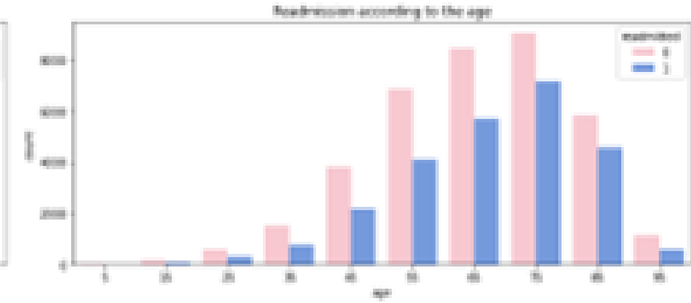


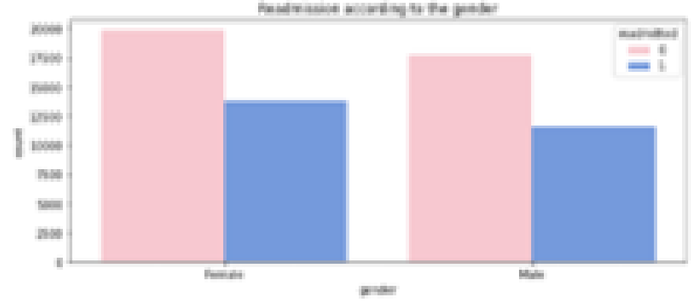


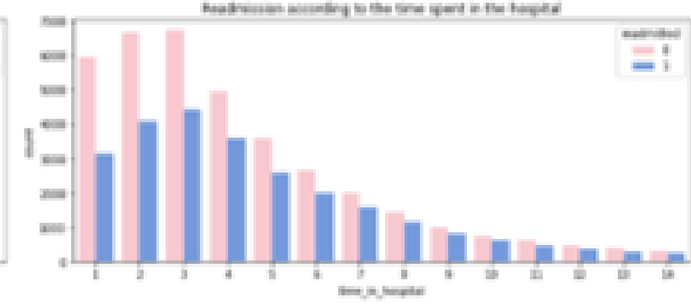
•We now consider the parameters in correlationwith the readmissionstatus. The patient isnoweitherreadmitted or not readmittedsothat the problemis more balanced

CORRELATIONS WITH THE READMISSION STATUS









STUDYING MACHINE LEARNING MODELS

Thus this far, we have fathomed what machine learning models we could apply to our dataset: RandomForest, DecisionTree, Linear Regression and DecisionTree classifiers. We will try to predict our values and evaluate our models.

Linear Regression

- Simplicity and Interpretability: It models the relationship between a dependent variable and one or more independent variables.
- Baseline Model: Can be used as a starting point for regression problems to establish a baseline performance

Random Forest

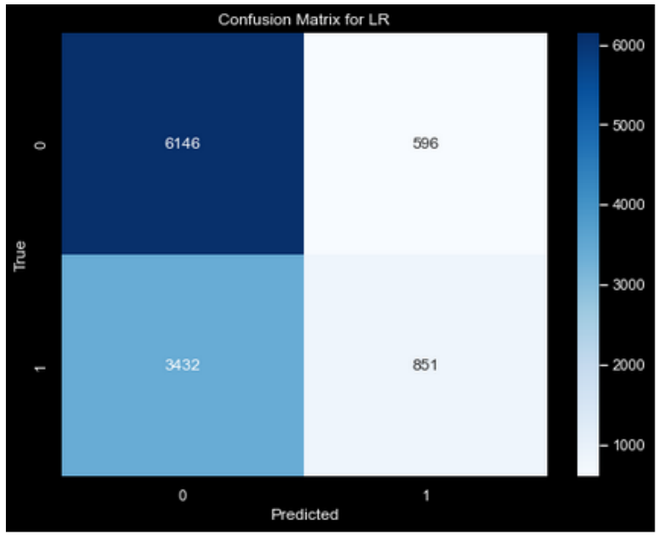
- Non-Linearity: Random Forest can handle non-linear data effectively. It combines multiple decision trees to improve the model's accuracy and prevent overfitting.
- Feature Importance: Provides insights into which features are most important in making predictions.
- Versatility: Works well for both classification and regression tasks and can handle large datasets with higher dimensionality

Decision Tree

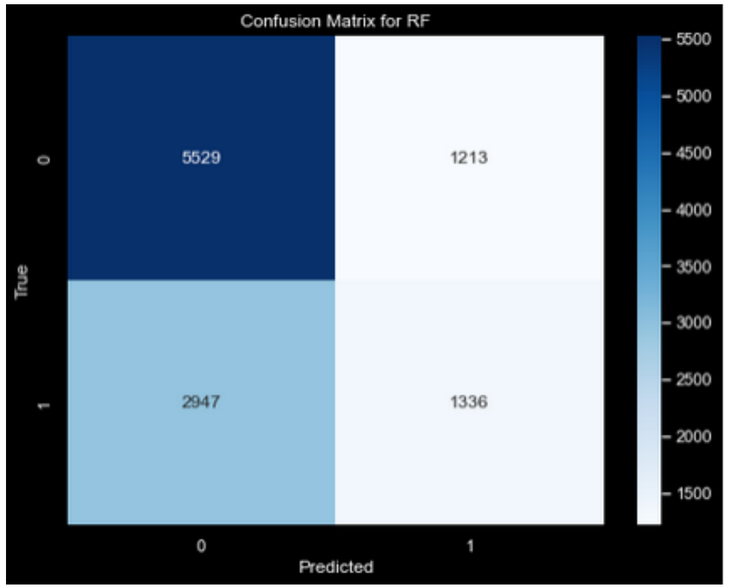
- Easy to Interpret: Decision Trees can be visualized easily, making them highly interpretable.
- Non-Parametric: Does not require much data preprocessing, such as normalization or scaling. They can handle both numerical and categorical data.

CONFUSION MATRIX

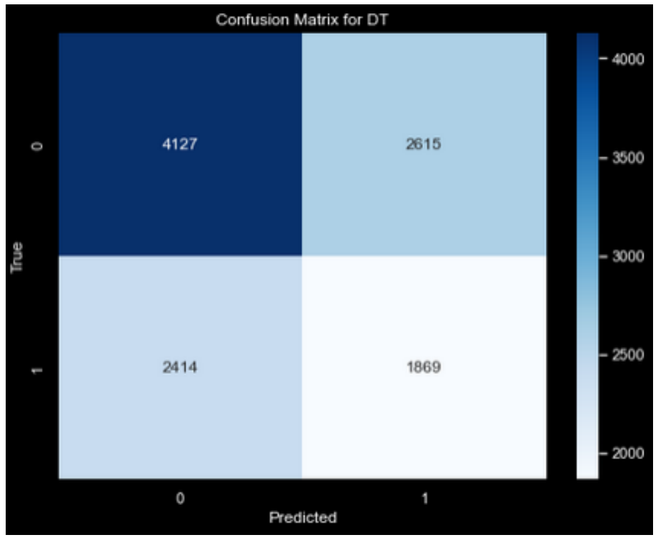
Linear Regression



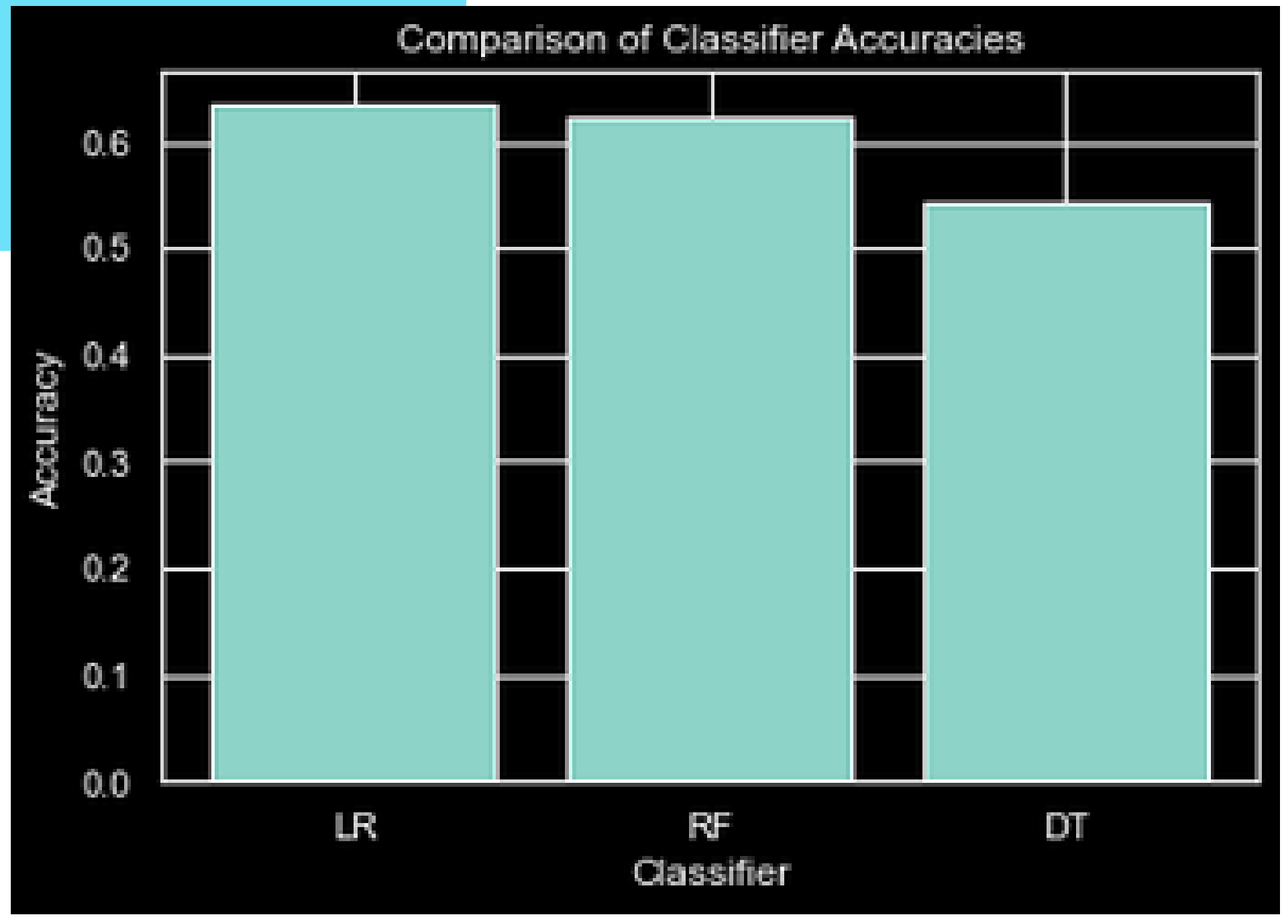
Random Forest



Doctor appointment



COMPARISON OF CLASSIFIER ACCURACIES

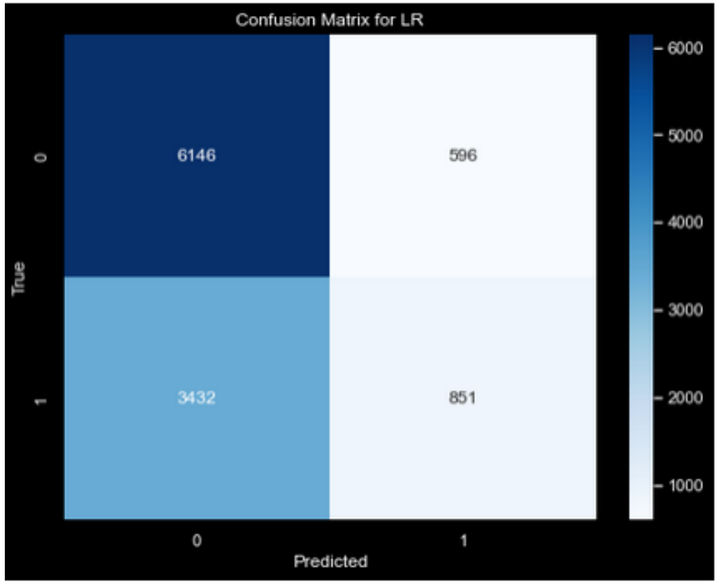


Observation

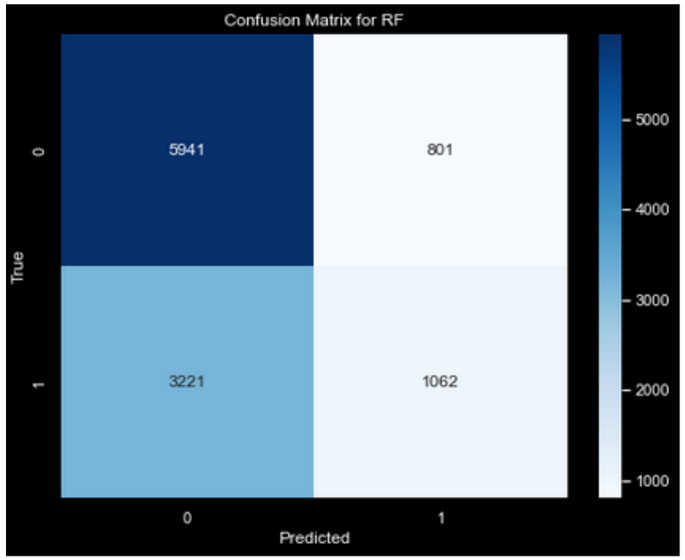
Our first results show us promising results in bout Linear Regression and Random Forest Classifiers, but Decision Tree seems to lack a few parameters to be completely operational.

RE-EVALUATING THE MODEL WITH THE BEST PARAMETERS

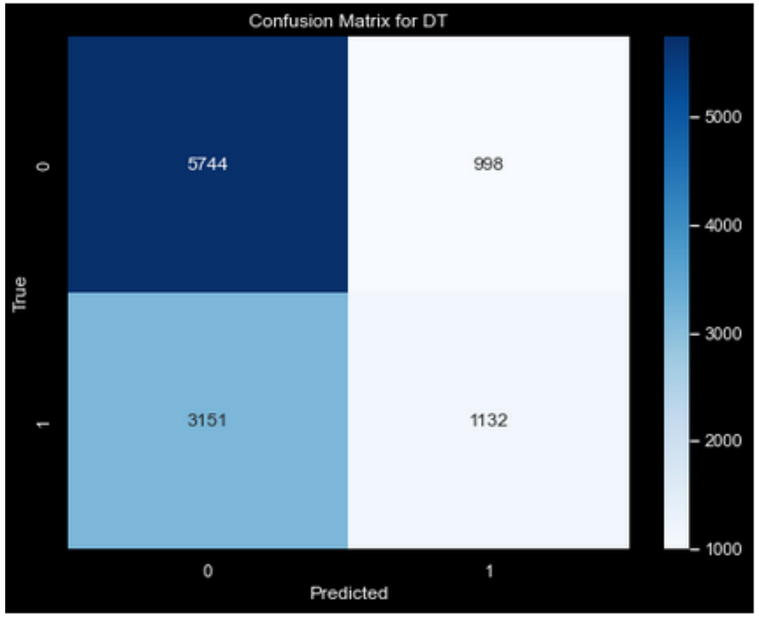
Linear Regression

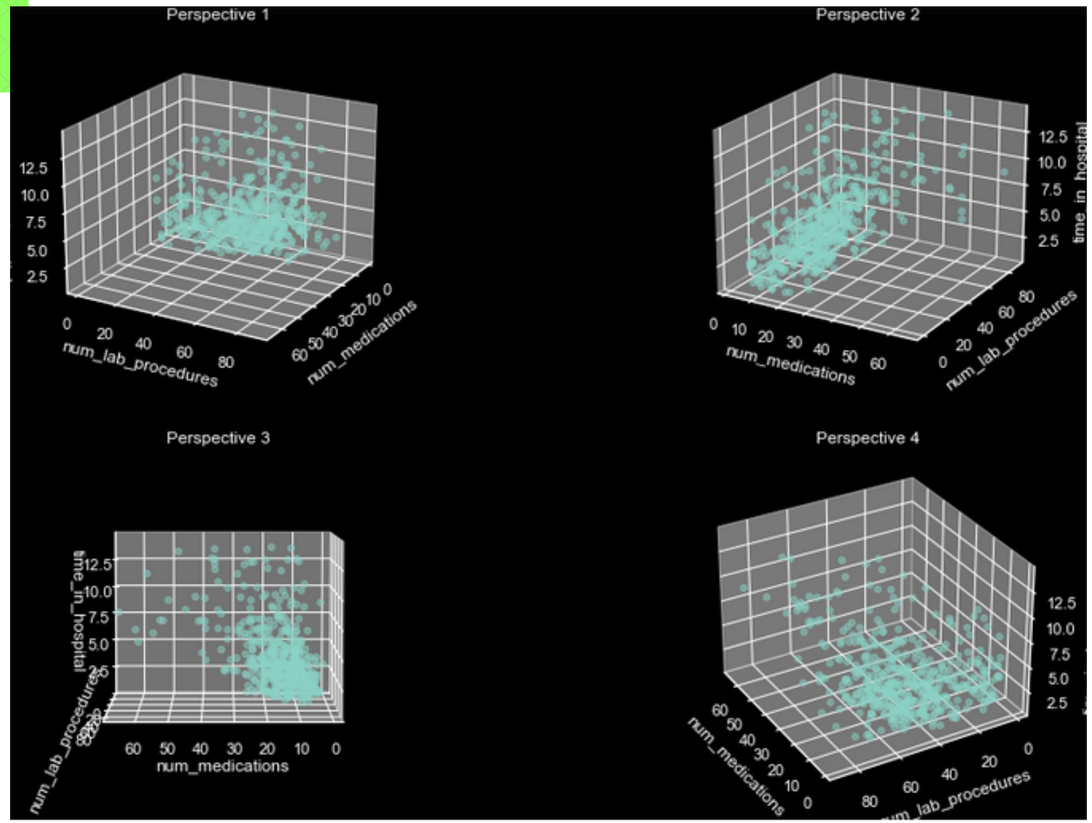


Random Forest



Doctor appointment



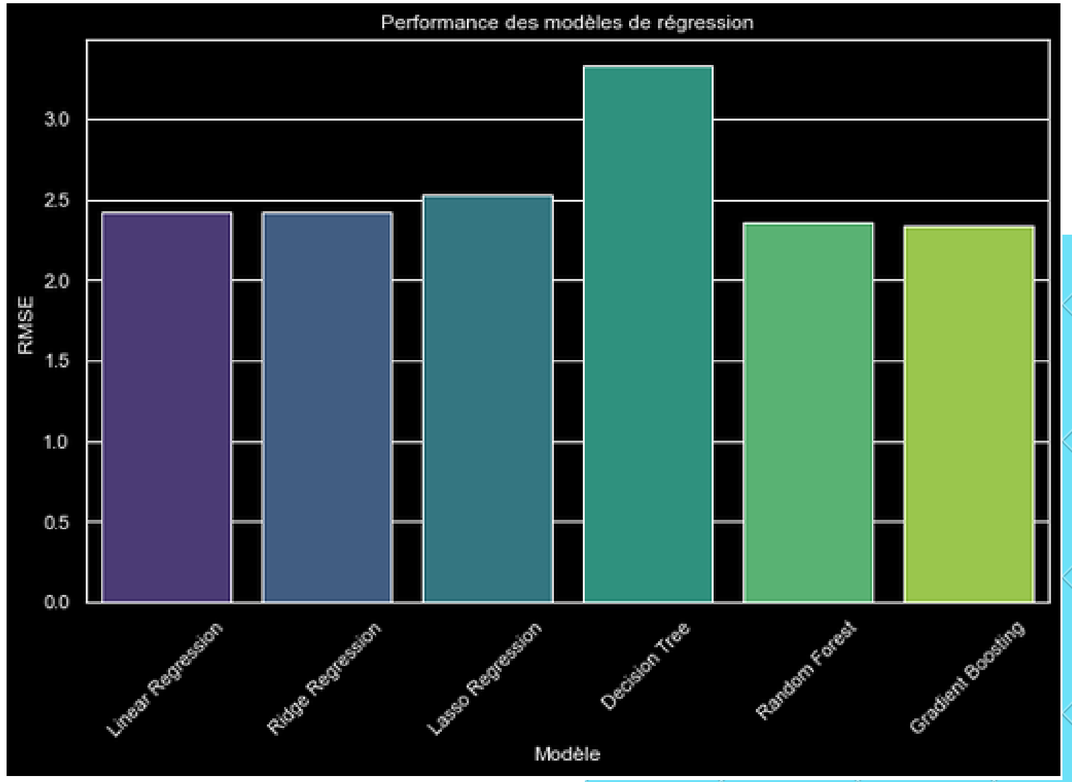


Data explained

Our approach involved a straightforward process: excluding health parameters. Health, in this context, is an aggregate of components considering the number of days spent in the hospital. The solution to our inquiry was inherent in the question itself. Additionally, we had to convert discrete variables into a binary format using dummy variables, ensuring compatibility with the sklearn library.

REGRESSION STUDY

we observe that the Gradient Boosting approach appears to be the most effective, although it is closely trailed by the other methods, with the exception of the decision tree, which performs notably worse. Regarding the most influential parameters, we can confidently assert that they are 'num_medications' and 'num_lab_procedures.' This correlation makes perfect sense, as a higher count of lab procedures is associated with a prolonged hospital stay. However, it comes as a surprise to us that the number of medications is not ranked higher; this finding is unexpected



CONCLUSION

We have now studied this dataset from the Machine Learning Models point of view, and can conclude to good results with the information we have extracted

Reflexions

With the obtained results, we can already make some previsions that lean more towards preventive measures than pinpoint predictions. This shift from a purely reactive to a more preventive stance in healthcare could significantly change how we approach chronic diseases like diabetes. It emphasizes the importance of early intervention and continuous, personalized care to manage the disease more effectively.

In conclusion, while our study has made considerable progress in predicting patient readmission, it also highlights the need for further research. This research should aim to uncover the less visible variables affecting diabetes, improve measurement methods, and ultimately, guide us towards more effective and personalized healthcare solutions

THANK YOU