Project data & Python 2023

Final report:
The readmission of diabetic patients

DOUBLIER Enzo
KHOLEEFA Hanna
HELLEGOUARCH Bérénice

Summary

Contents

Link of dataset and github :

https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008
https://github.com/Berelleg/Python/blob/main/project_final.ipynb

I- Abstract

Our research focuses on analysing readmission rates of diabetic patients, employing a dataset to study various factors impacting readmission. We applied data preprocessing, visualization, and classification techniques to explore correlations and patterns in the data. Our findings offer insights into demographic, medical, and temporal aspects influencing readmission.

II– Introduction

When we wanted to choose our subject, we put in common all of our passions, and we agree to make our project about health. In fact, Bérénice and Hanna want to choose the Health and Biotechnology option in next year in our school, and Enzo is interested in this subject too.

We thought then about what could make our project interesting about something that could help people, and we thought about diabetes, because we all saw this in high school biology class, but not in depth. Moreover, that is something that concern a lot of people as the prevalence rate of diabetes was 10.2% in 2003 and it will be 11.9% in 2025.

We found diverse studies that was giving factors of readmission, so we wanted to find if there were specific factors that was helping more patients of diabetes to have a readmission.

We found a dataset that looks at first sight promising to answer our question, so we decided to concentrate our work on this data.

That leads to our problematic : How to predict readmission on diabetic patients ?

III – Contribution

During the end of November until the 20th of December, we worked together on the code because we have all the same level, so it was easier to think about it. We made several appointments to do the milestone and to complete the code. For the final report and the poster, we divided the different parts fairly. All the member of the group contributed equivalently on the project.

## IV – Methodology and Results

We imported libraries and performed data preprocessing, including handling missing values, and modifying how the variables were segmented. Visualizations such as line plots, count plots, and correlation matrices aimed in understanding the relationships between variables. Moreover, we applied classification models to predict readmission based on patient data.

1. 4 basic correlations

First, we decided to study 4 basic correlations that come to our mind when we thought about diabetes.

- We first drew graphs of people with diabetes according to their age. Which was the primarily thing to do in our opinion.
- After, we wondered if there was a correlation between gender and diabetic people.
- Then, we thought about the number of procedures in laboratory and the number of readmission of patients.
- And we finished with the comparison with the time spent in the hospital and the readmissions.

These four basic studies revealed us interesting things. First, we see that there is a lot of patients in admission at hospital between 50 and 80 years old. That is expected as the primary research we did revealed us that there were two types of diabetes: type 1 diabetes which appear during childhood and affects around 6% of diabetics and type 2 diabetes which affects 92% and appear after 40 years. That is what we find in the two graph below.
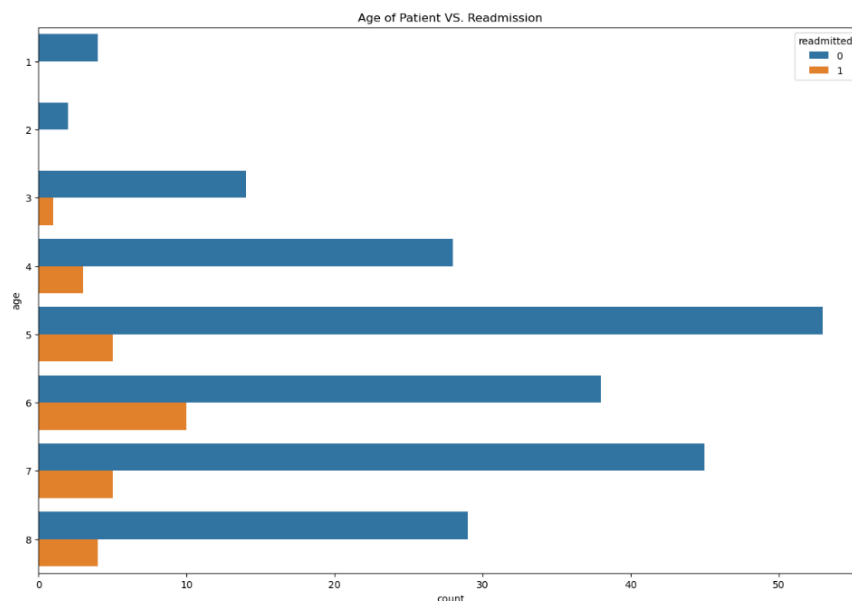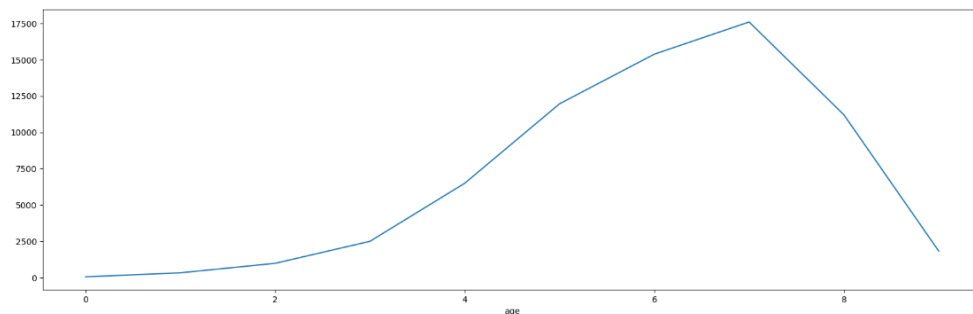
After that we thought interesting to see if diabetes affected men as much as women and we observe that women are more likely to have diabetes. We don't think that we can conclude anything about this because there is only 5 000 of difference which is not enough to tell if women are more likely to have diabetes than men because we sea on our primary research that men are more likely to have diabetes. We suppose we find a different result because of the dataset, we don't have enough people to have precise results.
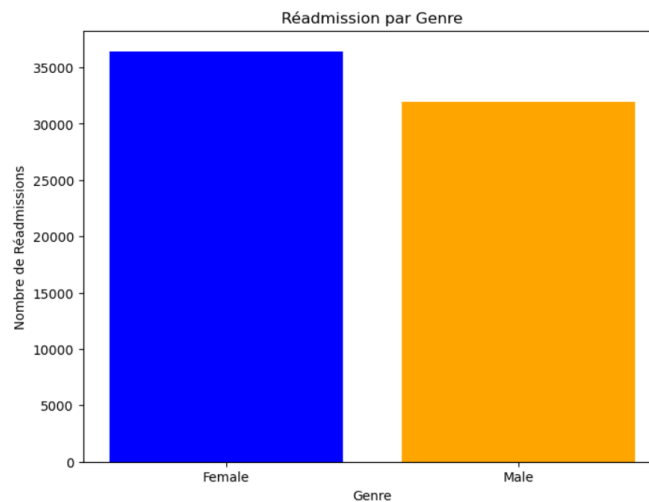
Then we wanted to see if there was a correlation between the number of readmissions and the number of medications and we discover that the graph follows a uniform law, that's means that from 0 to 4000 readmissions, the numbers of medications increase but then the more we have readmissions the more de number of medications decrease. We think that after a certain point, medications are no longer effective enough, that's why we have this graph.

Number of lab procedure VS. Readmission
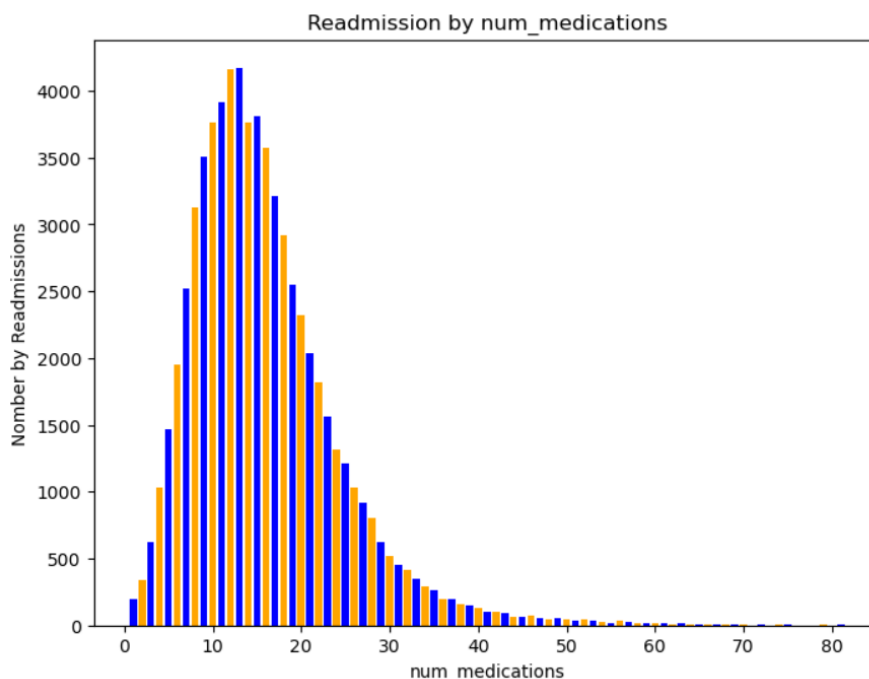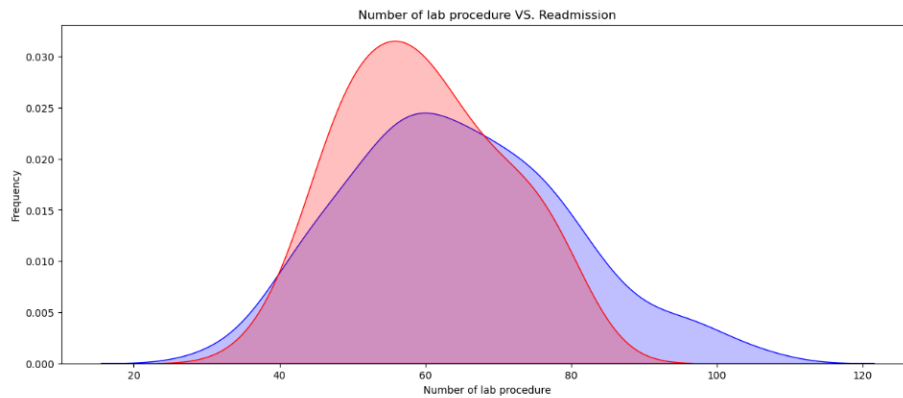
(blue : readmitted, red : non readmitted)

We decided to plot a pie chart of the percentage of people with their staying duration in hospital, and it's showing us that almost 50% of the patients leave the hospital after two days of duration.

With the graphic of the time spent in hospital compared to the readmitted or not people, we saw that the great part of non-readmitted people is not staying more than 6 days to the hospital, with a majority of 4 days, which is shorter than readmitted people in general. That shows us that the time in hospital is a factor to be taken in consideration.


Time in Hospital VS. Readmission
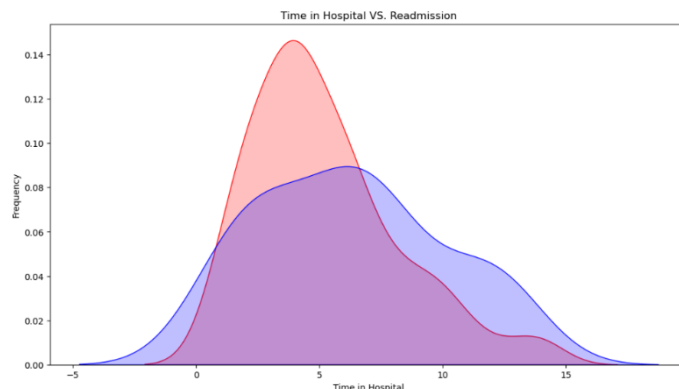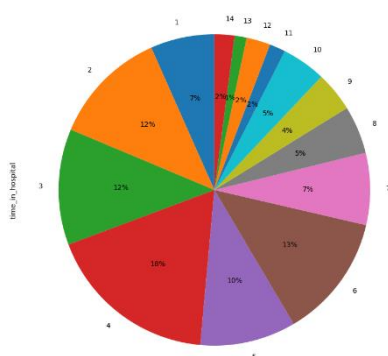
2.   Going deeper

In a second time, we wanted to dig deeper into our subject.

We wanted to think about what could be more interesting and original factors that could help a readmission of a patient :

We thought very interesting to see if there is a correlation with the ethnic, because it's not the first factor that could come to our mind when we thought about it.

 To study this eventual correlation, we decided to plot a graphic that shows the part of patient readmitted in function of their ethnic group appurtenance.
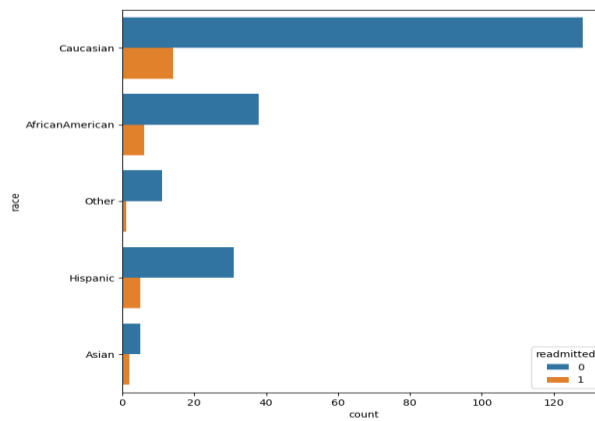
We found surprising that there were differences in readmission in function of ethnicity. We therefore decided to investigate the subject and found the reasons why certain ethnicities have a higher readmission rate. This is certainly due to diet, fast food, living conditions and genes predisposing to diabetes, access to the medical centre...

We wanted after this to plot a correlation matrix, to see if there was an eventual correlation between several factors, and we decided to graph an heatmap, so it was easier to see because of the number of variables.
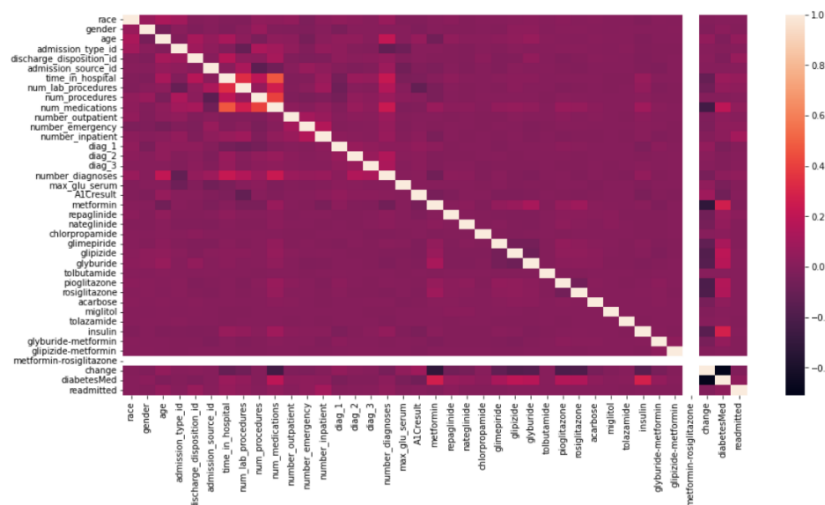
We found a pretty strong correlation between the time spent in the hospital and the number of medications for the patient.

3. Classification

To go further, we decided to draw a classification with two different methods in order to predict if a
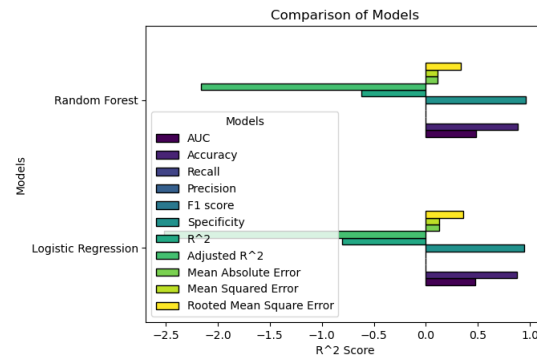
patient is likely to be readmitted or not. We used Logistic Regression and Random Forest. We conclude that Random Forest shows a little bit more better results than the Logistic Regression.

V- Conclusion

In summary, our analysis explored different influential factors in readmission rates for diabetic patients. While age, hospital stay duration, and medical procedures showed strong correlations, we saw that the gender, and the number of medications had minimal impact. Exploring into our data, we surprisingly saw that ethnicity could play a role in the readmission's factors and could be a very interesting factor to explore in new studies.

VI – Sources

https://www.diabetes.org.uk/diabetes-the-basics/differences-between-type-1-and-type-2-diabetes

https://my.clevelandclinic.org/health/diseases/7104-diabetes

https://www.who.int/news-room/fact-sheets/detail/diabetes

https://research-and-innovation.ec.europa.eu/research-area/health/diabetes_en#:~:text=Diabetes%20is%20a%20disease%20characterised,the%20EU%20suffer%20from%20diabetes.