LEPL1109 - Statistics and Data Sciences
HACKATHON 2 - Diabetes health indicators

Group n°18 November 29, 2024 Names and Noma of participants:

Part. 1: Decaluwé Maxime - 50802200
Part. 2: Defrenne Simon - 42242200
Part. 3: Mil-Homens Cavaco Mathieu - 38282200
Part. 4: Peiffer Thibaut - 47352200
Part. 5: Roekens Raphaël - 70732200
Part. 6: Starck Robin - 88952200

Lastname	Firstname	Noma
Decaluwé	Maxime	50802200
Defrenne	Simon	42242200
Mil-Homens Cavaco	Mathieu	38282200
Peiffer	Thibaut	47352200
Roekens	Raphaël	70732200
Starck	Robin	88952200

Please, read carefully the following guidelines:

- Answer in English, with complete sentences and correct grammar. Feel free to use grammar checker tools such as LanguageTools free and open-source plugin;
- Do not modify questions, and input all answers inside `\begin{answer}...\end{answer}` environments;
- Each question should be followed by an answer;
- At the end of each question, there is the length of the expected answer. This is for your information but it is not too important if you do not respect these recommendations.
- Clearly cite every source of information (even for pictures!);
- Whenever possible, use the `.pdf` format when you export your images: this usually makes your report look prettier[1];
- Do not forget to also submit your code on Moodle.
- **Reminder:** You need to belong to a group to submit your project on Moodle.

Contents

[1] This is because `.pdf` is a vector format, meaning that it keeps a perfect description of your image, while `.png` and other standard formats use compression. In other words, this means you can zoom as much as you want on your figure without decreasing image resolution. For simple plots, vector formats can also save a lot of memory space. On the other hand, we recommend using `.png` when you are plotting many data points: large scatter plots, heatmap, etc.

1 Description of the project

1.1 Your objective

You work in the diabetology department at **Saint Luc University Hospital**. The head of the department has asked you to find a solution for classifying and predicting **whether patients are at high risk of developing diabetes**. This will enable them to schedule an appointment with these patients to set up prevention tools. To do this, you have a database of patients who have passed through the department in recent years. In addition, the head of the department feels that the poll is too long, and would like to **reduce the number of questions while maintaining the reliability and quality of the results**. The attached `.ipynb` file will guide you in this process.

Your aim is to determine which characteristics are relevant and enable reliable patient classification. **Be careful**, don't let a potential diabetic patient slip through the cracks.

1.2 The dataset

The dataset is a real dataset based on a questionnaire carried out in the USA some ten years ago. It contains around 70 000 entries and is a collection of 22 features individually defined in table **??**.

Features name	Description	Range
`Diabetes`	Diabetes (0:no diabetes; 1:diabetes)	{0, 1}
`HighBP`	High blood pressure (0:no; 1:yes)	{0, 1}
`HighChol`	High cholesterol (0:no; 1:yes)	{0, 1}
`CholCheck`	Cholesterol check in 5 years (0:no; 1:yes)	{0, 1}
`BMI`	Body mass index in [kg/m^2]	/
`Smoker`	Smoked at least 100 cigarettes in your life (0:no; 1:yes)	{0, 1}
`Stroke`	Stroke (0:no; 1:yes)	{0, 1}
`HeartDisease`	Heart disease (0:no; 1:yes)	{0, 1}
`PhysActivity`	Physical activity in past 30 days (0:no; 1:yes)	{0, 1}
`Fruits`	Consume fruit 1 or more times per day (0:no; 1:yes)	{0, 1}
`Veggies`	Consume vegetables 1 or more times per day (0:no; 1:yes)	{0, 1}
`Alcohol`	Heavy alcohol drinkers (0:no; 1:yes)	{0, 1}
`AnyHelathcare`	Health insurance (0:no; 1:yes)	{0, 1}
`NoDocbcCost`	No doctor because of cost (0:no; 1:yes)	{0, 1}
`GenHlth`	General health (1:excellent; 5:poor)	{1, . . . , 5}
`MenHlth`	Number of days out of the last 30 when mental health was poor	{0, . . . , 30}
`PhysHlth`	Number of days out of the last 30 when physical health was poor	{0, . . . , 30}
`DiffWalk`	Serious difficulty for walking (0:no; 1:yes)	{0, 1}
`Sex`	0:female; 1:male	{0, 1}
`Age`	Age category (1:18-24; . . . ; 13:80 or older)	{1, . . . , 13}
`Education`	Education level (1:never; 6:university)	{1, . . . , 6}
`Income`	Income scale (1:less than $10,000; . . . ; 8:$75,000 or more)	{1, . . . , 8}

Table 1: Data set features

2 Questions and answers (4/10)

Question 2.1:

(1/10) What happens to the precision and recall (of any method) when the threshold tends to 0? And when it tends to 1? How can you explain it?
Expected answer length : 8 lines.

Answer to ??:

When the threshold goes to 0, recall approaches its maximum value of 1 because it is the ratio of true positives to the total actual positives, and very few (if any) false negatives occur, since almost everything is classified as positive. However, precision decreases because it is the ratio of true positives to the total predicted positives, and many false positives are included due to the broad classification of positive instances.

When the threshold tends to 1, precision approaches 1, as only the most confident positive predictions are retained, significantly reducing false positives. However, recall decreases because many true positives are missed, resulting in a large number of false negatives.

Question 2.2:

(1/10) Explain which precision/recall trade-off you prefer to have for the specific task asked in this hackathon: don't let a potential diabetic slip through the cracks. How should you adjust the threshold of your model to bring it closer to the desired trade-off? Should it be above or below the default threshold value of 0.5?
Expected answer length : 5 lines.

Answer to ??:

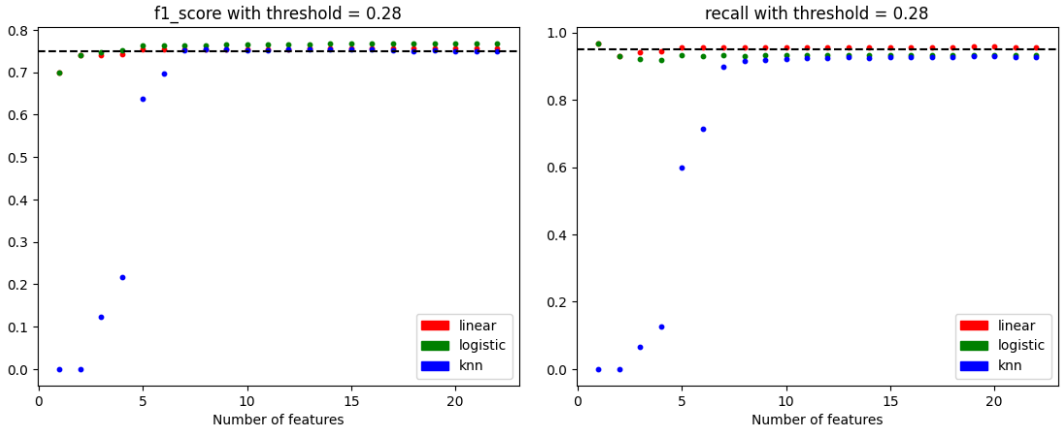
In our situation, we want to reduce as much as possible the amount of false negative, since the worst case is to predict someone is not diabetic while he is positive because that way we would ignore him while him being critically at risk. Telling someone that he is diabetic while he is not is less a problem as we could realize our mistake later on. We then want to maximize the recall of our prediction while maintaining a descent precision through f1 score. Knowing that the higher the threshold, the higher the precision and the lower the recall, for higher recall we need to reduce the thresold : we then need a thresold below 0.5.

Question 2.3:

(1/10) Based on your code, select a final model that you will keep as classifier. **Justify.**
Expected answer length : 5 lines.

Answer to ??:

The model that fits our desired recall and F1-score the best, while requiring the least possible amount of questions is the linear model. Compared to the others, it reaches these values with only 5 features for a threshold of 0.28. The logistic method needs as much questions as the linear for the F1-score, but it gives a lower recall. The KNN method is nearly as good as the logistic for both the recall and the F1 score.



Question 2.4:

(1/10) Could you reduce the length of the questionnaire? If so, how many questions? Which questions? **Justify.**
Expected answer length : 6 lines.

Answer to ??:

If we use linear regression, we can reach the threshold with as few as 5 features, which means we can keep only the 5 questions that are the most correlated with diabetes. These questions are related to: GenHlth, HighBP, BMI, Highchol, and Age. Adding more features does not significantly improve the results for either linear regression or logistic regression.

3 Visualization (2/10)

Question 3.1:

(2/10) To answer this question, we ask you to produce a clear, clean figure expressing a result or giving an overall vision of your work for this hackaton. Please feel free to do as you wish. Be original! The clarity, content and description of your figure will be evaluated.
Expected answer length : 4 lines + 1 figure.

Answer to ??:

This work was articulated into three main parts. First, we analyzed the data and prepared it for further algorithmic processing. Then, we trained several models using the k-folds method. Finally, we chose the best-performing model to predict diabetes patients. Notice from Recall-F1 score curves that at least 5 features are required for logistic and linear regression, and that KNN cannot match those requirements even with all the features.