
LEPL1109 - Statistics and Data Sciences

HACKATHON 2 - Diabetes health indicators

Group n°XX

November 29, 2024

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 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 prettier¹;
- 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	Description of the project	2
1.1	Your objective	2
1.2	The dataset	2
2	Questions and answers (4/10)	3
2.1	3
2.2	3
2.3	3
2.4	3
3	Visualization (2/10)	4
3.1	4

¹This 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 1.

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 2.1:

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 2.2:

Maxime

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 2.3:

Maxime

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 2.4:

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 3.1: