Задания для курса “Анализ данных в медицине”

Классификация

Задача построить классификатор, который по 21 признаку относит объект к одному из 2 классов:
0 = no diabetes
1 = diabetes
Обучить и проверить классификатор на приложеной базе данных (файл **diabetes_012_health_indicators_BRFSS2015.csv**).
Алгоритм выполнения задачи:

1. Выбрать наиболее значимые признаки:
   - удаляются признаки у которых процент пропущенных значений больше порогового
   - удаляются признаки, у которых коэффициент корреляции больше порогового
   - удаляются признаки, состоящие из одного значения
   - Для тех, кто захочет попробовать самостоятельно разобраться - удалить признаки, имеющие низкую важность в модели lightgbm (следить за точностью модели)
2. Убрать пропуски в данных (либо заполнить средним, либо выбросить данный вектор)
3. Сбалансировать датасет, убрать ненужный класс
4. Сформировать обучающую выборку с расчетом не скомпенсированных данных (здоровых намного больше диабетиков). Исходить из расчета, что обучающая выборка должна содержать оба класса в примерно одинаковых пропорциях.
5. Выбрать модель. (логистическая регрессия или случайные лес).

Пример кода с похожей базой данных приложен к файлу.

Описание признаков:

1. Diabetes_012: 0 = no diabetes, 1 = prediabetes, 2 = diabetes
2. HighBP: 0 = no high BP 1 = high BP
3. HighChol: 0 = no high cholesterol 1 = high cholesterol
4. CholCheck: 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years

5. BMI: Body Mass Index
6. Smoker: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
7. Stroke: (Ever told) you had a stroke. 0 = no 1 = yes
8. HeartDiseaseorAttack: coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
9. PhysActivity: physical activity in past 30 days - not including job 0 = no 1 = yes
10. Fruits: Consume Fruit 1 or more times per day 0 = no 1 = yes
11. Veggies: Consume Vegetables 1 or more times per day 0 = no 1 = yes
12. HvyAlcoholConsump: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes
13. AnyHealthcare: Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes
14. NoDocbcCost: Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes
15. GenHlth: Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
16. MentHlth: Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? scale 1-30 days
17. PhysHlth: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? scale 1-30 days
18. DiffWalk: Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
19. Sex: 0 = female 1 = male
20. Age: 13-level age category 1 = 18-24 9 = 60-64 13 = 80 or older
21. Education: Education level scale 1-6 1 = Never attended school or only kindergarten 2 = Grades 1 through 8 (Elementary) 3 = Grades 9 through 11 (Some high school) 4 = Grade 12 or GED (High school graduate) 5 = College 1 year to 3 years (Some college or technical school) 6 = College 4 years or more (College graduate)
22. Income: Income scale, scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

Кластеризация

Задача изучить датасет (файл **clusters.xlsx**) с помощью метода кластеризации.

Задача сводится к загрузке датасета и применению к нему одного из методов кластеризации (см. лекцию по задачам и методам машинного обучения). Как результат необходимо предоставить файл (excel или csv), где каждому вектору (строчке) сопоставляется некий найденный класс. Так же для каждого класса необходимо указать его центр.

Описание датасета:

1. Pregnancies: Number of times the patient has been pregnant.
2. Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test.
3. BloodPressure: Diastolic blood pressure (mm Hg).
4. SkinThickness: Triceps skinfold thickness (mm).
5. Insulin: 2-Hour serum insulin (mu U/ml).
6. BMI: Body mass index (weight in kg/(height in m)^2).
7. DiabetesPedigreeFunction: A function that scores likelihood of diabetes based on family history.
8. Age: Age of the patient (years)