

Рударење на масивни податоци – Домашна бр. 3

Симона Ристовска 221003

1. Offline фаза

Податочно множество

Податоците од множеството **Diabetes Health Indicators Dataset (BRFSS 2015)**.
се **небалансириани**, при што класата за предијабетес е значително помалку застапена
во споредба со останатите класи.

Поделба на податоците

Со цел да се задржи соодносот на класите, при сите поделби го користев **stratified sampling**.

1. Offline / Online поделба – специфицирано во задачата

- **80% offline податоци**, користени за обучување и валидација на моделите
- **20% online податоци**, наменети за online фазата

Поделбата беше извршена со stratified sampling според целната променлива Diabetes_012.

2. Train / Validation поделба (offline податоци)

Offline податоците дополнително беа поделени на:

- **60% податоци за тренирање**
- **20% податоци за валидација**

И кај оваа поделба беше задржан истиот сооднос на класите.

Preprocessing на податоците

Сите влезни променливи се нумерички.

За моделите кои пресметуваат растојанија (во овој случај KNN) беше применет **StandardScaler** со цел истите да се стандардизираат. Ова го направив преку **Pipeline функцијата**, со што се овозможува иста предобработка при тренинг и при понатамошна употреба на моделот.

Обучување и избор на модел

Користев 3 три различни модели за класификација: **Decision Tree**, **K-Nearest Neighbors (KNN)**, **Gradient Boosting**. За секој модел правев **GridSearchCV** со цел да се најдат најдобрите хиперпараметри и **3-fold cross-validation**. Поради мултикласните и небалансираните податоци, како метрика за евалуација беше користен **F1-macro**, кој им

дава еднаква важност на сите класи. Моделот со највисока **F1-macro вредност на validation подмножеството** го земав како најдобар, а тоа е **Decision Tree**.

2. Online фаза

Во рамки на online фазата најпрво е вклучен python producer-от (producer.py) кој ги чита податоците од датотеката online.csv и ги испраќа ред по ред во JSON формат кон Kafka topic-от health_data. При ова, колоната која означува дали пациентот има дијабетес (Diabetes_012) е отстранета, согласно барањата на задачата. Податоците се испраќаат со доцнење се со цел да се симулира реален поток на податоци.

```
{'HighBP': 0.0, 'HighChol': 1.0, 'CholCheck': 1.0, 'BMI': 30.0, 'Smoker': 0.0, 'Stroke': 0.0, 'HeartDiseaseorAttack': 0.0, 'PhysActivity': 0.0, 'Fruits': 0.0, 'Veggies': 0.0, 'HvyAlcoholConsump': 0.0, 'AnyHealthcare': 1.0, 'NoDocbcCost': 0.0, 'GenHlth': 2.0, 'MentHlth': 0.0, 'PhysHlth': 0.0, 'DiffWalk': 1.0, 'Sex': 1.0, 'Age': 7.0, 'Education': 4.0, 'Income': 7.0}
Sleeping for 1.21778584750507 seconds

{'HighBP': 1.0, 'HighChol': 1.0, 'CholCheck': 1.0, 'BMI': 26.0, 'Smoker': 1.0, 'Stroke': 1.0, 'HeartDiseaseorAttack': 0.0, 'PhysActivity': 0.0, 'Fruits': 1.0, 'Veggies': 1.0, 'HvyAlcoholConsump': 0.0, 'AnyHealthcare': 1.0, 'NoDocbcCost': 0.0, 'GenHlth': 4.0, 'MentHlth': 0.0, 'PhysHlth': 30.0, 'DiffWalk': 1.0, 'Sex': 0.0, 'Age': 13.0, 'Education': 4.0, 'Income': 2.0}
Sleeping for 1.381934571385663 seconds

{'HighBP': 1.0, 'HighChol': 0.0, 'CholCheck': 1.0, 'BMI': 41.0, 'Smoker': 0.0, 'Stroke': 0.0, 'HeartDiseaseorAttack': 0.0, 'PhysActivity': 1.0, 'Fruits': 1.0, 'Veggies': 1.0, 'HvyAlcoholConsump': 0.0, 'AnyHealthcare': 1.0, 'NoDocbcCost': 1.0, 'GenHlth': 4.0, 'MentHlth': 10.0, 'PhysHlth': 0.0, 'DiffWalk': 1.0, 'Sex': 1.0, 'Age': 4.0, 'Education': 4.0, 'Income': 5.0}
Sleeping for 0.8091140274976476 seconds

{'HighBP': 1.0, 'HighChol': 1.0, 'CholCheck': 1.0, 'BMI': 29.0, 'Smoker': 1.0, 'Stroke': 0.0, 'HeartDiseaseorAttack': 0.0, 'PhysActivity': 0.0, 'Fruits': 0.0, 'Veggies': 0.0, 'HvyAlcoholConsump': 0.0, 'AnyHealthcare': 1.0, 'NoDocbcCost': 0.0, 'GenHlth': 2.0, 'MentHlth': 0.0, 'PhysHlth': 10.0, 'DiffWalk': 1.0, 'Sex': 1.0, 'Age': 13.0, 'Education': 4.0, 'Income': 6.0}
Sleeping for 0.5067371577394852 seconds

{'HighBP': 0.0, 'HighChol': 1.0, 'CholCheck': 1.0, 'BMI': 35.0, 'Smoker': 1.0, 'Stroke': 0.0, 'HeartDiseaseorAttack': 0.0, 'PhysActivity': 0.0, 'Fruits': 0.0, 'Veggies': 1.0, 'HvyAlcoholConsump': 0.0, 'AnyHealthcare': 1.0, 'NoDocbcCost': 1.0, 'GenHlth': 5.0, 'MentHlth': 8.0, 'PhysHlth': 30.0, 'DiffWalk': 1.0, 'Sex': 0.0, 'Age': 9.0, 'Education': 4.0, 'Income': 6.0}
Sleeping for 1.2318946331382534 seconds
```

Потоа е Apache Spark апликацијата (spark.py) со помош на Spark Structured Streaming го вчитува потокот на податоци од Kafka topic-от health_data. Читаните JSON пораки се парсираат со претходно дефинирана schema. Секој запис кој Spark го прочитал од Kafka topic-от се предава на претходно обучениот Decision Tree модел (best_diabetes_model.pkl) со цел да се изврши предикција. Добиената предвидена класа се додава како ново поле (diabetes_pred) во записот и потоа се испраќаат во JSON формат кон нов Kafka topic со име health_data_predicted.

```
+-----+
|{"HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":26.0,"Smoker":1.0,"Stroke":1.0,"HeartDiseaseorAttack":0.0,"PhysActivity":0.0,"Fruits":1.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":0.0,"GenHlth":4.0,"MentHlth":0.0,"PhysHlth":39.0,"DiffWalk":1.0,"Sex":0.0,"Age":13.0,"Education":4.0,"Income":2.0,"diabetes_pred":0}|
+-----+
|{"HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":41.0,"Smoker":0.0,"Stroke":1.0,"HeartDiseaseorAttack":10.0,"PhysActivity":0.0,"Fruits":1.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":0.0,"GenHlth":4.0,"MentHlth":10.0,"PhysHlth":1.0,"DiffWalk":1.0,"Sex":1.0,"Age":4.0,"Education":4.0,"Income":5.0,"diabetes_pred":1}|
+-----+
|{"HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":29.0,"Smoker":1.0,"Stroke":1.0,"HeartDiseaseorAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":0.0,"GenHlth":2.0,"MentHlth":0.0,"PhysHlth":18.0,"DiffWalk":1.0,"Sex":1.0,"Age":13.0,"Education":4.0,"Income":6.0,"diabetes_pred":0}|
+-----+
|{"HighBP":0.0,"HighChol":1.0,"CholCheck":1.0,"BMI":35.0,"Smoker":1.0,"Stroke":0.0,"HeartDiseaseorAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":1.0,"GenHlth":5.0,"MentHlth":8.0,"PhysHlth":38.0,"DiffWalk":1.0,"Sex":0.0,"Age":9.0,"Education":4.0,"Income":6.0,"diabetes_pred":0}|
+-----+
|value
+-----+
|{"HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":25.0,"Smoker":1.0,"Stroke":0.0,"HeartDiseaseorAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":1.0,"GenHlth":5.0,"MentHlth":18.0,"PhysHlth":38.0,"DiffWalk":1.0,"Sex":0.0,"Age":9.0,"Education":4.0,"Income":6.0,"diabetes_pred":0}|
+-----+
|value
+-----+
|{"HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":34.0,"Smoker":1.0,"Stroke":1.0,"HeartDiseaseorAttack":0.0,"PhysActivity":1.0,"Fruits":1.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":0.0,"GenHlth":5.0,"MentHlth":18.0,"PhysHlth":38.0,"DiffWalk":1.0,"Sex":1.0,"Age":14.0,"Education":4.0,"Income":6.0,"diabetes_pred":2}|
+-----+
|{"HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":30.0,"Smoker":1.0,"Stroke":1.0,"HeartDiseaseorAttack":1.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":0.0,"GenHlth":4.0,"MentHlth":0.0,"PhysHlth":38.0,"DiffWalk":1.0,"Sex":1.0,"Age":13.0,"Education":4.0,"Income":6.0,"diabetes_pred":2}|
+-----+
|{"HighBP":0.0,"HighChol":1.0,"CholCheck":1.0,"BMI":27.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaseorAttack":0.0,"PhysActivity":1.0,"Fruits":1.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDocbcCost":0.0,"GenHlth":4.0,"MentHlth":14.0,"PhysHlth":30.0,"DiffWalk":0.0,"Sex":0.0,"Age":6.0,"Education":6.0,"Income":8.0,"diabetes_pred":0}|
+-----+
```

Дополнително, изработен е и **Kafka consumer** кој служи за верификација на резултатите и ги чита пораките од topic-от health_data_predicted, прикажувајќи ги предвидувањата во

КОНЗОЛА .

```
("HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":35.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":1.0,"MenHlth":0.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":0.0,"Age":18.0,"Education":0.0,"Income":4.0,"diabetes_pred":0)  
("HighBP":0.0,"HighChol":1.0,"CholCheck":1.0,"BMI":30.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":2.0,"MenHlth":3.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":1.0,"Age":18.0,"Education":0.0,"Income":4.0,"diabetes_pred":0)  
("HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":23.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":2.0,"MenHlth":1.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":0.0,"Age":18.0,"Education":0.0,"Income":4.0,"diabetes_pred":0)  
("HighBP":0.0,"HighChol":1.0,"CholCheck":1.0,"BMI":23.0,"Smoker":1.0,"Stroke":0.0,"HeartDiseaserAttack":1.0,"PhysActivity":1.0,"Fruits":0.0,"Veggies":0.0,"HvyAlcoholConsump":1.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":3.0,"MenHlth":1.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":0.0,"Age":18.0,"Education":4.0,"Income":4.0,"diabetes_pred":0)  
("HighBP":0.0,"HighChol":0.0,"CholCheck":1.0,"BMI":24.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":1.0,"Fruits":1.0,"Veggies":1.0,"HvyAlcoholConsump":1.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":1.0,"MenHlth":0.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":0.0,"Age":18.0,"Education":6.0,"Income":8.0,"diabetes_pred":0)  
("HighBP":0.0,"HighChol":0.0,"CholCheck":1.0,"BMI":24.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":1.0,"Fruits":1.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":2.0,"MenHlth":0.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":0.0,"Age":18.0,"Education":4.0,"Income":4.0,"diabetes_pred":0)  
("HighBP":0.0,"HighChol":1.0,"CholCheck":1.0,"BMI":30.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":2.0,"MenHlth":0.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":0.0,"Age":18.0,"Education":0.0,"Income":7.0,"diabetes_pred":0)  
("HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":26.0,"Smoker":1.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":1.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":4.0,"MenHlth":0.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":1.0,"Age":17.0,"Education":4.0,"Income":7.0,"diabetes_pred":0)  
("HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":30.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":4.0,"MenHlth":0.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":1.0,"Age":17.0,"Education":4.0,"Income":12.0,"diabetes_pred":0)  
("HighBP":0.0,"HighChol":1.0,"CholCheck":1.0,"BMI":30.0,"Smoker":0.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":5.0,"MenHlth":1.0,"PhysHlth":0.0,"DiffWalk":0.0,"Sex":1.0,"Age":17.0,"Education":4.0,"Income":12.0,"diabetes_pred":0)  
("HighBP":1.0,"HighChol":1.0,"CholCheck":1.0,"BMI":29.0,"Smoker":1.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":0.0,"Veggies":0.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":0.0,"GenHlth":2.0,"MenHlth":0.0,"PhysHlth":0.0,"DiffWalk":1.0,"Sex":1.0,"Age":17.0,"Education":4.0,"Income":6.0,"diabetes_pred":0)  
("HighBP":0.0,"HighChol":1.0,"CholCheck":1.0,"BMI":35.0,"Smoker":1.0,"Stroke":0.0,"HeartDiseaserAttack":0.0,"PhysActivity":0.0,"Fruits":1.0,"Veggies":1.0,"HvyAlcoholConsump":0.0,"AnyHealthcare":1.0,"NoDo  
cbcCost":1.0,"GenHlth":5.0,"MenHlth":0.0,"PhysHlth":38.0,"DiffWalk":1.0,"Sex":0.0,"Age":19.0,"Education":4.0,"Income":4.0,"diabetes_pred":0)
```