

Домашна работа 3

Ева Смилеска 221053

1. Опис на проблемот

Целта на домашната задача е имплементација на machine learning pipeline за предвидување на дијабетес со offline обука и online streaming предвидување. Користено е податочното множество Diabetes Health Indicators Dataset од BRFSS 2015, кое содржи здравствени индикатори за пациенти со различни нивоа на дијабетес.

Проблемот е поделен на две фази:

- **Offline фаза:** Обука на модели за класификација со Apache Spark
- **Online фаза:** Real-time предвидување преку Kafka streaming

2. Подготовка на податоци

2.1. Поделба на податочното множество

Користен е скриптот split_data.py за поделба на оригиналното множество:

```
offline, online = train_test_split(  
    *arrays: df,  
    test_size=0.2,  
    random_state=42,  
    stratify=df[target_col]  
)
```

Резултати од поделбата:

- **offline.csv:** 80% од податоците (за обука)
- **online.csv:** 20% од податоците (за streaming)
- Балансирани класи со stratify параметар

3. Offline фаза - Обука на модели

3.1. Трансформации на податоци

Користен е Spark ML Pipeline со следните трансформации:

```
assembler = VectorAssembler(inputCols=feature_cols, outputCol="features_raw")
scaler = StandardScaler(inputCol="features_raw", outputCol="features", withMean=False, withStd=True)
```

- **VectorAssembler:** Комбинирање на сите features во еден вектор
- **StandardScaler:** Стандардизација на податоците

3.2. Обучени модели

Обучени се 3 модели со различни хиперпараметри:

1. Logistic Regression

- regParam: [0.0, 0.01, 0.1]
- elasticNetParam: [0.0, 0.5, 1.0]
- Вкупно: 9 комбинации

2. Random Forest

- numTrees: [50, 100]
- maxDepth: [5, 10]
- Вкупно: 4 комбинации

3. Decision Tree

- maxDepth: [5, 10, 15]
- maxBins: [32, 64]
- Вкупно: 6 комбинации

3.3. Избор на најдобар модел

Користев 3-fold Cross-Validation со метрика F1 score.

Резултати:

LR] best CV F1 = 0.8081

[RF] best CV F1 = 0.8005

[DT] best CV F1 = 0.8076

BEST MODEL = LR with CV F1 = 0.8081

Test F1 (offline holdout) = 0.8101

```
[LR] best CV F1 = 0.8081
[RF] best CV F1 = 0.8005
[DT] best CV F1 = 0.8076

BEST MODEL = LR with CV F1 = 0.8081
Test F1 (offline holdout) = 0.8101
```

Најдобриот модел е зачуван во директориумот `models/best_model/`.

4. Online фаза - Streaming предвидување

4.1. Kafka Producer

Скриптата `producer_health.py` ги праќа податоците од `online.csv` во Kafka:

Клучни карактеристики:

- Topic: `health_data`
- Format: JSON
- Лабелата `Diabetes_012` не се праќа (само features)
- Batch processing: секои 50 пораки со 0.1s пауза

Резултати:

Done. Sent 14200 messages to topic 'health_data'.

4.2. Spark Streaming апликација

Скриптата `stream_predict.py` врши real-time предвидување:

Чекори:

1. **Вчитување на модел:** Се вчитува претходно обучениот модел
2. **Читање од Kafka:** Се consume-ат пораки од `health_data` topic
3. **Парсирање:** JSON → DataFrame
4. **Трансформација:** Автоматски се применуваат истите трансформации од pipeline-от
5. **Предвидување:** Секој запис се предвидува со моделот

6. **Збогатување:** Се додава predicted_class поле

7. **Испраќање:** Резултатите се испраќаат на health_data_predicted topic

```
pred = model.transform(parsed)

enriched = pred.withColumn("predicted_class", col("prediction").cast("int"))
```