Биљана Димитрова

Домашна 3

Вовед

Во оваа задача, фокусот беше на развој систем за класификација на дијабетес врз основа на големо податочно множество со околу 236,000 редици. Податочното множество содржи различни карактеристики (features) кои опишуваат здравствена состојба на пациенти, како и класата Diabetes_binary (1 = има дијабетес, 0 = нема дијабетес).

Податочното множество е поделено на два дела:

1. **Offline податоци**: Се користат за тренирање и евалуација на моделите. (80% од податочното множество)
2. **Online податоци**: Се користат за симулирање на реално време обработка на податоци. Овие податоци се испраќаат ред по ред во Apache Kafka, и притоа се подготвуваат за предвидување. (20% од податочното множестов

Овој пристап овозможува:

- **Offline фаза**: Тренирање на модели и избор на најдобар модел.
- **Online фаза**: Реално време предвидување и збогатување на податоците со предвидените класи.

Offline фаза

Во **offline фазата**, фокусот е на тренирање и евалуација на модели за класификација на дијабетес. Оваа фаза се состои од следните чекори:

1. **Вчитување на податоците**: Податочното множество е вчитано од CSV датотека (offline.csv) со помош на Apache Spark. Податоците вклучуваат различни карактеристики (features) и бинарен атрибут Diabetes_binary.
2. **Трансформација на податоците**: Податоците се трансформирани за да бидат подготвени за тренирање на моделите. Ова вклучува создавање на вектор од карактеристики (features) и скалирање на карактеристиките со средна вредност 0 и стандардна девијација 1.
3. **Тренирање на моделите**: Тренирани се три модели: логистичка регресија (Logistic Regression), дрво на одлука (Decision Tree) и случајна шума (Random Forest). За секој модел е извршена вкрстена валидација (Cross-Validation) за оптимизација на хиперпараметрите, а метриката за евалуација е F1 score.

4. **Евалуација и избор на најдобар модел**: Моделите се евалуирани врз основа на F1 score. Најдобриот модел е случајна шума (Random Forest), кој е зачуван за подоцнежна употреба.

Online фаза

Во **online фазата**, фокусот е на обработка на податоци во реално време со користење на Apache Kafka и Spark Structured Streaming. Оваа фаза се состои од следните чекори:

1. **Producer скрипта**: Податоците од offline.csv се испраќаат ред по ред во Kafka topic наречен health_data. Податоците се испраќаат во JSON формат, без колоната Diabetes_binary.
2. **Apache Spark апликација**:
    - Податоците се вчитани од Kafka и парсирани во соодветна структура.
    - Се применуваат истите трансформации како во offline фазата (VectorAssembler и StandardScaler).
    - Се прави предвидување со моделот случајна шума (Random Forest) обучен во offline фазата.
    - Резултатите се збогатени со предвидената класа и испратени во нов Kafka topic health_data_predicted.
3. **Consumer скрипта:** Има за цел да ги прикажува испратените предвидувања во читлив формат.

Заклучок

- Во **offline фазата**, најдобриот модел за класификација на дијабетес е случајна шума (Random Forest), кој е зачуван за подоцнежна употреба.
- Во **online фазата**, податоците се обработуваат во реално време со користење на Apache Kafka и Spark Structured Streaming. Предвидените класи се испраќаат во нов Kafka topic (health_data_predicted), што овозможува реално време мониторинг и анализа.