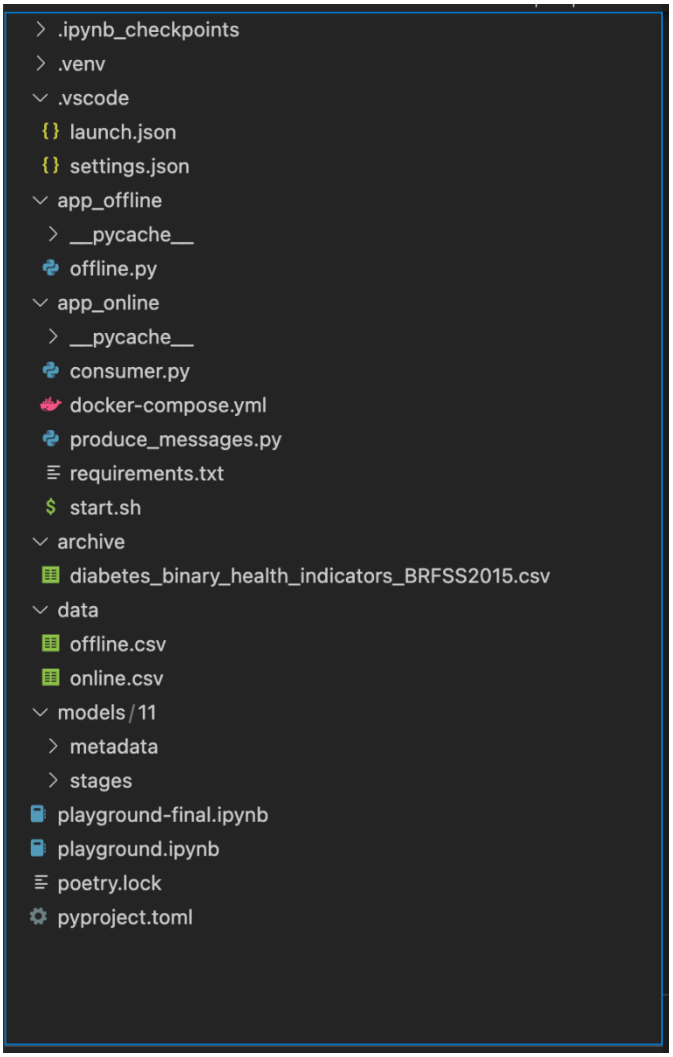
Домашна задача 3 – Spark

Милан Тасевски, 196001



Структурата на решението е во две апликации, app_offline и app_online. Во app_offline е кодот за првиот дел од домашната, во app_offline за вториот.

Во кодот има коментари за секој процес од решавањето. Како резимирање, app_offline има main метод во кој се креира pipeline-от за тренирање, од вчитување на податоците па се до зачувување на најдобриот модел. Има посебна фукнција transform_df, која е одговорна за трансформирање на датасетот (скалирање и вектор асемблер), која е дел од pipeline-от. Потоа, се додаваат различни модели со менување на

хиперпараметрите. За избор на најдобар, мораше да ги зачувам со pipelineModel.write().overwrite().save("../models/" + str(id)), со тоа што се зачувува секој нов најдобар модел. На крај според id ги бришам сите други. Ова мораше да се изведе на овој начин затоа што подоцна load на моделот сакав да направам преку PipelineModel.load(). Инаку, не е возможно да се зачува цел pipeline, ако ги ставаме во листа па го земаме само моделот со надобар score. Не се работи за редунтантен код, иако не е идеален.

Оваа скрипта треба да се изврши пред сите други во домашната, со што завршуваме со offline фазата.

Понатаму, скриптата produce_messages.py, заедно со docker-compose.yml е одговорна за продуцирање на пораките и праќање порака по порака од датасетот online.csv, после drop-нување на класата на пациентот. KafkaProducer-от ова го испраќа на topic health_data. Со оваа скрипта имав проблем при извршување на mac, заради Kafka библиотеката и M1 чип-от.

consumer.py е скриптата одговорна за пречекување на stream-от, и враќање назад предикција користејќи го моделот трениран во offline фазата. Првин се лоадира моделот, па се креира спарк сесија и се прави subscribe на topic health_data. За датасетот се кастира json-от и се прави предикција со помош на вчитаниот pipeline model со линијата model.transform(df).select("prediction"). Со ова, се извршува целиот Pipeline врз моменталниот dataframe. На крај, назад се враќа целиот stream во кафка формат на topic-от health_data_predicted.

Проблемот со последната скрипта е во тоа што не успеав да ја извршам, на Mac заради тоа што не поминува ни producer-от (немам податоци за вчитување преку stream), а на Windows поради слаба архитектура и проблеми со Hadoop.