В качестве проекта для итоговой аттестации по курсу Продвинутые методы машинного обучения я выбрал комплексную задачу.

Комплексная задача включает в себя задачу бинарной классификации. Необходимо построить алгоритм и решить задачу классификации на полученном в задании датасете.

Актуальность работы состоит в том, что с развитием технологий и увеличением объема данных, бинарная классификация становится все более важной для различных приложений, поэтому необходимо отрабатывать известные алгоритмы на новых данных для уточнения и модернизации этих алгоритмов. Такие данные могут быть из различных областей жизни: наука, социальная, экономическая сферы и также сфера безопасности.

Целью работы является решение задачи бинарной классификации на полученном в задании датасете. Предлагалось решить задачу двумя алгоритмами: методом k ближайших соседей и решающим деревом. Необходимо определить превысит ли средний заработок человека порог $50k. Каждый объект выборки — человек, для которого известен набор признаков.

Рассмотрим подробнее каждый из методов. Метод k ближайших соседей – это один из наиболее простых и понятных алгоритмов машинного обучения, используемых для задач классификации и регрессии. Основная идея метода заключается в том, что объект классифицируется на основе большинства классов его k ближайших соседей в пространстве признаков. Для задачи классификации новый объект присваивается классу, который является наиболее часто встречающимся среди k ближайших соседей. Метод дерева решения – это один из наиболее популярных и широко используемых алгоритмов машинного обучения для задач классификации и регрессии. Дерево решений представляет собой модель, которая использует деревоподобную структуру для принятия решений на основе признаков объектов. Критерии разделения для классификации, например информационная энтропия или прирост Джини.

Перейдем к решению задачи. Для начала я выбрал один гиперпараметр, который оптимизировал. Для kNN — число соседей, для DecisonTree — глубина дерева. Остальные параметры оставил в значениях по умолчанию.

Для каждого алгоритма подобрал оптимальные значения указанных гиперпараметров. И построил график среднего значения качества по кросс-валидации алгоритма при заданном значении гиперпараметра, на котором также отобразил доверительный интервал. Из графиков видно, что при оптимальных параметрах точность выше у решающего дерева.

Также я обучил те же классификаторы на трансформированных с помощью PCA данных. PCA снижает размерность данных, сохраняя наиболее значимую информацию. Это уменьшает количество признаков, что может упростить модель и уменьшить риск переобучения. По графику видно, что оптимальное число признаков 40.

Подобрал оптимальные значения исследуемых гиперпараметров для новых данных. Построил график среднего значения качества по кросс-валидации алгоритмов при заданном значении гиперпараметра. Из графиков видно, что немного изменилась динамика среднего значения качества по кросс-валидации алгоритма. Получили оптимальное значение гиперпараметра меньше, чем без PCA, что позволяет быстрее обучать модель.

Далее я протестировал все построенные модели (в том числе и без PCA трансформации) на данных, которые были отложены для теста, их осталось 40% от всего датасета. Результаты для точности приведены в таблице. Видно, что лучше всего справился kNN с PCA преобразованием.

Подведем итоги выполненной работы. В ходе работы была решена задача бинарной классификации. Реализованы и исследованы алгоритмы kNN и DecisionTree, применено PCA преобразование. Поставленная задача выполнена с хорошей точностью.

Спасибо за внимание.