

Практическая работа №3. Исследование древовидных алгоритмов классификации.

В работе исследуются алгоритмы построения классификационных моделей (kNN, дерево решений, случайный лес). Для выполнения работы используются библиотеки языка Python для анализа данных (pandas, sklearn).

Задания:

1. Выбрать набор данных для классификации.
2. Построить модель `DecisionTreeClassifier` и исследовать влияние параметров алгоритма на качество классификации (число уровней, минимальное число объектов узла). Привести оценки качества классификации на обучающем и тестовом наборе для полного и усеченного дерева.
Для выбранной модели построить структуру дерева решений, указать показатели, участвующие в структуре дерева.
3. Построить модель `RandomForest` и исследовать влияние параметров алгоритма на качество классификации (число деревьев, число отбираемых показателей, построение полных\ограниченных деревьев).
Для выбранной модели привести структуру 3 деревьев решений, участвующих в модели.
4. Получить оценки важности показателей с помощью модели `RandomForest`. Для трёх показателей с наибольшей оценкой важности привести диаграммы с группировкой по классу (boxplot или countplot).
5. Построить модели классификации одного типа (kNN, `DecisionTree` или др.) на полном наборе показателей, на 20% показателей с наибольшими оценками важности и на 80% показателей с наименьшими оценками важности. Сопоставить качество классификации для полученных моделей. В случае необходимости повторить эксперимент на большом количестве показателей с наибольшими оценками.
6. *Провести дополнительный анализ коллективных древовидных моделей на базе «бустинга» (последовательного улучшения): `AdaBoostClassifier`, `GradientBoostingClassifier`.