Заключение:

В работе были рассмотрено использование фильтров и встроенных методов в задаче отбора признаков. К достоинствам фильтров можно отнести низкую стоимость вычислений (линейно зависит от количества признаков) и интерпретируемость. К недостаткам – то, что они рассматривают каждый признак изолировано, поэтому не могут выявить более сложные зависимости в данных, например, зависимость от нескольких предикторов. Эти методы хорошо подойдут, если в данных большое количество признаков, но малое количество объектов (что встречается, например, в медицинских, или биологических исследованиях).

Также в работе были рассмотрены модель решающего дерева и случайного леса. К достоинствам решающего дерева можно отнести порождение четких правил классификации и быстрые процессы обучения и прогнозирования, а к достоинствам случайного леса - существование методов оценивания значимости отдельных признаков в модели и способность эффективно обрабатывать данные с большим числом признаков и классов. К недостаткам деревьев относят чувствительность к шумам во входных данных и необходимость отсекать ветви дерева (pruning) или устанавливать минимальное число элементов в листьях дерева или максимальную глубину дерева для борьбы с переобучением. К недостаткам случайного леса – увеличенную сложность интерпретации по сравнению с решающим деревом.