ИСПОЛЬЗОВАНИЕ АЛГОРИТМА ДЕРЕВА РЕШЕНИЙ ДЛЯ АНАЛИЗА МНОГОМЕРНЫХ ДАННЫХ НА ПРИМЕРЕ ДАННЫХ ПО ОНКОЛОГИЧЕСКИМ ЗАБОЛЕВАНИЯМ ЛЁГКИХ

Корховая А.Б.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Лапицкая Н.В. – зав. каф. ПОИТ

В данной работе были рассмотрены и определены основные задачи для разработки алгоритмов по обработки медицинских данных онкологий лёгких. Были проанализированы подходы в машинном обучении и выбраны необходимые для данного решения задачи.

В контексте данной работы будут рассматриваться медицинские данные собранные на территории Беларуси по онкологическим заболеваниям лёгких. Вопрос проблемы лечения данного вида заболевания не теряет своей актуальности:

1) Рак является одной из основных причин смерти в мире: в 2018 году произошло 9,6 миллиона случаев смерти от рака;

2) Употребление табака является отдельным самым значительным фактором риска развития рака, который приводит к более чем 20% глобальных случаев смерти от рака и примерно 70% глобальных случаев смерти от рака легких;

3) По прогнозам исследований, число случаев заболевания от рака будет продолжать расти от 16 до 22 миллионов в следующие десятилетия.

С ходом развития технологий появились новые возможности использования знаний для поиска возможных путей предотвращения смертей от данного вида заболевания, например, основываясь на статистике построить линию протекания болезни больного в зависимости от различных факторов. Стоит отметить, что существуют успешные случаи излечение от онкологий лёгких, однако таких случаев очень мало.

Основной целью моего исследования стало определение пациента к одной из возможных категорий в зависимости от предложенных параметров (факторов).

Количество факторов, влияющих на исход заболевания, может быть огромнейшее количество и одной из технических задач становится определение значимых признаков, независящих друг от друга, для определения прогноза исхода заболевания [1].

Передо мной была поставлена задача: на основе предложенных данных по онкологии лёгких выявить подходящие алгоритмы анализа, которые в будущем врачи могли бы использовать для оценки возможных исходов лечения.

В предложенном датасете, для начала, было выделено 11 параметров, для которых будет проводиться анализ и в будущем прогнозироваться исход заболевания:

1) Лучевая терапия (1-была проведена, 0 – не проводилась).

2) Химия терапия (1 – была проведена, 0 – не проводилась).

3) Возраст (числовое значение).

4) Стадия рака лёгких, которая была установлена как диагноз (в диапазоне 1-4).

5) Шифр гистологии рака лёгких.

6) Жалобы (1 – пациент жаловался на плохое состояние, 0 – жалобы отсутствовали).

7) Пол пациента (1 – мужской, 0 - женский).

8) Онкоанамнез (1 – у родственников был рак, 0 – у родственников не встречалось данного заболевания).

9) Статус курения (1 – пациент курит, 0 – пациент не курит).

10) Операция (1 – проводилась, 0 – не проводилась).

11) Тип проводимой операции (зашифровано врачом: 0-3 – различные типы операций, 4 – операция не проводилась).

Параметр, определяющий результат заболевания, – зависимая переменная – также является бинарным: исход заболевания (1 – летальный исход; 0 – выздоровление).

Первым шагом в анализе данных стало определения их нормального распределения [1]. Результат оказался отрицательным: предложенные данные не имеют нормального распределения, а также очень мало количество выздоровевших в сравнении с умершими в результате болезни. График распределения данных представлен на рисунке 1:

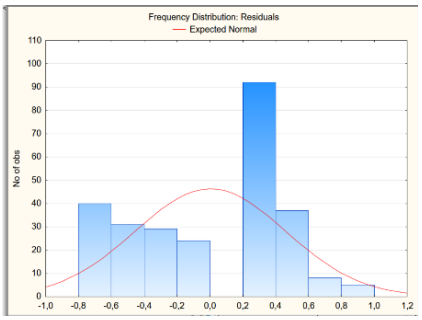

Рис. 1 – График нормального распределения данных

Основываясь на распределении данных, можно сделать вывод о нерациональности использования алгоритмов регрессионного анализа. Сама задача же заключена в определении определённого класса для пациента: выздоровеет или нет. Таким образом можно предположить о возможности применения дерева решений для решения поставленной задачи. Тип дерева решений определяется тем, что необходимо определить и так как зависящая переменная имеет дискретные значения, то деревья классификации станут наиболее подходящим решением [2].

При помощи ПО «STATISTICA» был проведен первичный эксперимент построения дерева решений (классификации) основываясь на модели, которую ПО генерирует сама и автоматически рассчитывает значимые признаки для данной модели. Полученные результаты представлены на рисунке 2 и 3:

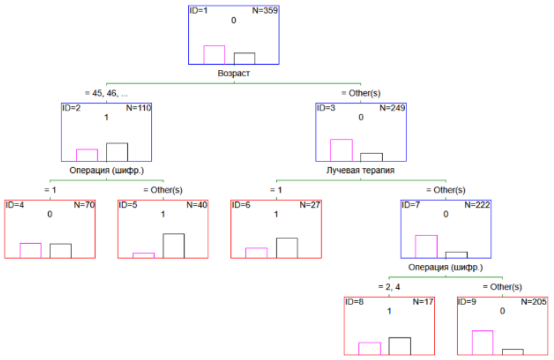

Рис. 2 – Дерево классификации

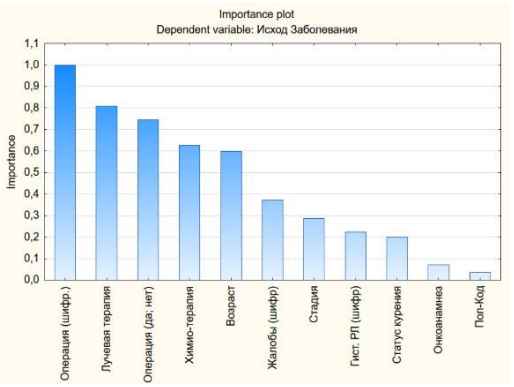

Рис. 3 – График значимости параметров выборки

Одна из сильных сторон дерева классификации является его гибкость, особенно важно для данной выборки его возможность работать с параметрами различных типов [3]: бинарных и числовых. Таким образом построенное дерево решение определила самые важные признаки, такие как возраст пациента, проводилась ли лучевая терапия и операция, и операция какого типа проводилась. Стоит отметить что самыми незначительными параметрами стали: пол, онкоанамнез и статус курения. Исходя из построенного алгоритма можно сказать, что данные параметры не должны влиять на протекание болезни онкологией лёгких.

В конечном счете, цель анализа с помощью деревьев классификации состоит в том, чтобы получить максимально точный прогноз. А максимально точный прогноз – это прогноз с минимальным числом неправильных классификаций [3]. Однако по полученным результатам, можно сказать, что для половины тестируемой выборки – классы распределяются неверно. И, следовательно, необходимы дальнейшие улучшения алгоритма, для решения данной задачи.

Данный этап моей работы на первоначальном этапе определил первые значащие признаки, которые влияют на протекание онкологии лёгких. Также было построено дерево решений, которое на основе данных признаков может отнести пациента к определённому классу.

Основываясь на построенных результатах можно сказать, что дерево решений не является оптимальным решением поставленной задачи, поэтому в дальнейшем моей целью станет применение алгоритма «Лес деревьев решений» и выбора оптимальных параметров для улучшения результатов алгоритма.

Список использованных источников:

6. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001. ISBN 0-387-95284-5.
7. Шахиди А. Деревья решений — общие принципы работы. URL: http://www.basegroup.ru/library/analysis/tree/description/
8. Quinlan, J. R. Induction of Decision Trees // Machine Learning. Kluwer Academic Publishers. 1986. № 1. P. 81–106.