УДК 004.62

ДЕРЕВО РЕШЕНИЙ ДЛЯ КЛАССИФИКАЦИИ СЛУЧАЕВ ИНФИЦИРОВАНИЯ ПАЦИЕНТОВ ИКСОДОВЫМИ КЛЕЩАМИ

В.С. Сафронов, Е.В. Сафронова
Научный руководитель: доцент, к.т.н., С.В. Аксёнов
Национальный исследовательский Томский политехнический университет,
Россия, г. Томск, пр. Ленина, 30, 634050
E-mail: vss75@tpu.ru

DECISION TREE AS A METHOD FOR CLASSIFICATION OF TICK-BORNE INFECTIONS

V.S. Safronov, E.V. Safronova
Scientific Supervisor: Ass. Prof., S.V. Aksenov
Tomsk Polytechnic University, Russia, Tomsk, Lenin str., 30, 634050
E-mail: vss75@tpu.ru

Abstract. *This article describes the process of classifying tick-borne infections using a decision tree. The classification quality was assessed. With the help of the scheme the principle of classification is presented. The most important predictors that have the greatest impact on the diagnosis are given.*

Введение. Многие насекомые являются переносчиками различных заболеваний. Иксодовыми клещами контактным и трансмиссивным путями чаще всего могут быть переданы человеку такие инфекции, как иксодовый клещевой боррелиоз (ИКБ, Болезнь Лайма) и клещевой энцефалит (КЭ), а также их микст. Если в короткие сроки после контакта с клещом не принимаются необходимые меры, то повышается риск заболеть, перенести заболевание в тяжелой форме, а в некоторых случаях возможен летальный исход [1]. Симптомы у ИКБ и КЭ схожи, и при поступлении пациента в медицинское учреждение с подозрением на клещевую инфекцию помимо осмотра врачом, проводится лабораторная диагностика биоматериала, что занимает некоторое время [2, 3]. С клещевыми инфекциями, как и с большинством заболеваний, очень важно как можно раньше поставить диагноз и назначить соответствующее лечение. Методы машинного обучения позволяют выявлять связь, оценивать степень влияния каждого отдельного предиктора на целевую переменную. Такой метод обучения с учителем, как дерево решений достаточно прост в интерпретации результатов разделения набора данных на классы [4]. **Целью данного исследования** является проведение классификации данных пациентов с клещевыми инфекциями с помощью алгоритма дерева решений.

Основная часть. Данные 193 пациентов, обратившихся в медицинские учреждения с клещевыми инфекциями, и проходящими лечение в стационаре медицинских учреждений Томской области, были предоставлены сотрудниками кафедры инфекционных заболеваний Сибирского государственного медицинского университета.

Проведенный при приеме пациентов осмотр протоколируется, результаты фиксируются в истории болезни пациента. В данный документ заносится следующая информация: возраст, пол, рост, вес, характер взаимодействия с переносчиком заболевания, предпринятые меры для снижения риска

заболевания, наличие сопутствующих заболеваний, а также отклонений от нормы в состоянии здоровья (результаты осмотра терапевтом и неврологом), результаты дополнительных обследований (осмотр окулистом, ультразвуковое исследование (УЗИ) и электрокардиограмма (ЭКГ)). После госпитализации с определенной периодичностью проводятся анализы биоматериала, замеры температуры, давления и пульса. Общее количество признаков составило более 150. В результате исключения признаков, напрямую связанных с диагнозом, имеющих большое количество пропущенных значений, а также явно не влияющих на постановку диагноза, предикторов осталось 97.

Разделение данных было проведено в отношении 70 к 30. На рисунке 1 представлено соотношение тестовых и тренировочных данных по различным диагнозам. Разными цветами представлены диагнозы, светлым оттенком – тестовая, темным – обучающая выборки.

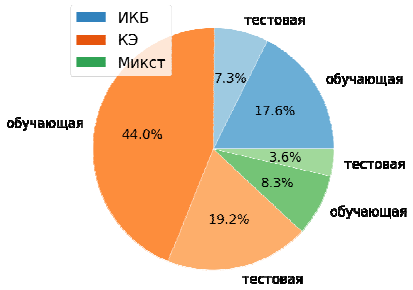

Рис. 1. Процентное соотношение тестовых и тренировочных данных по диагнозам

Прежде чем приступить к обучению модели, была проведена настройка гиперпараметров с помощью перекрестного поиска по сетке. Варьировались значения следующих параметров: критерий расщепления (Энтропия, коэффициент Джини), максимальная глубина дерева (от 2 до 10), пороговое значение критерия расщепления для разделения данных в узле (0,3, 0,2, 0,1). Оптимальной оказалась модель со следующими параметрами: критерий расщепления – Энтропия, максимальная глубина дерева – 3, пороговое значение критерия расщепления – 0,1.

Для оценки качества работы модели классификации были рассчитаны такие метрики, как специфичность и чувствительность, их значения составили 0,70 и 0,69, соответственно. В 69% случаев пациенты верно отнесены к каждому конкретному диагнозу и в 70% случаев определено, что пациенты к данному диагнозу не относятся.

Результат классификации, а также важности предикторов представлены на рисунке 2. Наибольшее влияние на определение диагноза оказывают наличие или отсутствие такого отклонения, как заторможенность и процент содержания в крови нейтрофилов (NEU%). На первом шаге значительная доля пациентов с наличием заторможенности была отнесена к классу КЭ. Оставшиеся случаи при проценте нейтрофилов больше или равном 60,4% алгоритмом отнесены к КЭ, однако сюда же были отнесены 3 и 5 случаев на самом деле являющихся ИКБ и Микст, соответственно. К классу же ИКБ были отнесены 7 случаев КЭ и 11 – Микст.

Клещевой энцефалит является вирусной инфекцией, в то время как Болезнь Лайма представляет собой инфекцию, вызванную бактериями. В большинстве случаев увеличение числа нейтрофилов в общем анализе крови свидетельствует о бактериальной инфекции, что противоречит принципу разделения данных на классы, представленному выше. Пациенты, содержание нейтрофилов в крови

которых превышает норму, должны были быть отнесены к классу ИКБ. Однако, в острый период, особенно при менингеальной и очаговой формах клещевого энцефалита, наблюдается нейтрофильный лейкоцитоз (т.е. повышенное содержание нейтрофилов) [5].

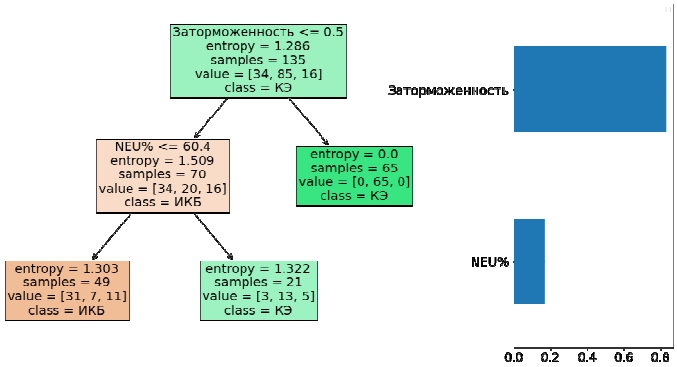

Рис. 2. Визуальное представление классификации деревом решений по диагнозам

Заключение. В рамках данного исследования была проведена классификация данных пациентов, страдающих такими клещевыми инфекциями, как клещевой энцефалит, иксодовый клещевой боррелиоз и микст, методом дерева решений. Доля корректно классифицированных пациентов по диагнозам составила 70%. Наиболее важными признаками для отнесения случая заболевания к тому или иному диагнозу оказались наличие или отсутствие заторможенности, а также процент содержания в крови нейтрофилов. Однако, последний предиктор является достаточно спорным, так как высокое содержание нейтрофилов характерно как для бактериальной болезни Лайма, так и для менингеальной и очаговой форм клещевого энцефалита. Такое заболевание, как микст КЭ и ИКБ деревом решений не было выявлено. Можно сделать вывод, что качество классификации заболеваний по диагнозам клещевых инфекций с помощью дерева решений достаточно невысокое. Для улучшения результатов можно использовать ансамблевые методы, включающие несколько деревьев решений.

СПИСОК ЛИТЕРАТУРЫ

1. Диагностика, лечение и профилактика клещевого энцефалита и иксодового клещевого боррелиоза у военнослужащих МО РФ: методические указания / составители: К. В. Жданов [и др.] – Москва: МО РФ, 2018. – 62 с.
2. Болезнь Лайма: особенности клещевого боррелиоза // СитиЛаб [Электронный ресурс]. – URL: https://citilab.ru/articles/kleshevoi-borrelioz-(bolezn_-laima)/ (дата обращения: 26.01.2022).
3. Клещевой энцефалит // Медицина [Электронный ресурс]. – URL: https://www.medicina.ru/patsientam/zabolevanija/kleshchevoj-encefalit/ (дата обращения: 3.02.2022).
4. Руководство по деревьям принятия решений для машинного обучения и науки о данных // Машинное обучение [Электронный ресурс]. – URL: https://www.machinelearningmastery.ru/a-guide-to-decision-trees-for-machine-learning-and-data-science-fe2607241956/ (дата обращения: 10.01.2022).
5. Ильинских Е.Н. [и др.] Клещевой энцефалит. Методическое пособие для врачей, интернов и клинических ординаторов. – Томск, 2015. – 31 с.