УДК 796.42

МОДЕЛЬ ОЦЕНКИ ФИЗИОЛОГИЧЕСКОГО СОСТОЯНИЯ СПОРСТМЕНОВ НА ОСНОВЕ АНАЛИЗА ГЕНОТИПОВ

Перетятько Сергей Игоревич, магистрант,
программист отдела мониторинга и оценки качества образовательной деятельности
Курский государственный университет
Peretyatko Sergey, graduate student, Kursk State University, seregivo@gmail.com

Аннотация. В работе рассмотрено использование дерева решений для анализа наборов данных, содержащих медицинские показатели, сведения о генотипах и фенотипах спортсменах. Модель обучена и применена для решения задачи определения состояния здоровья спортсменов.

Ключевые слова: решающее дерево, биохимический анализ, генотип, фенотип, коэффициент джини, межгенные взаимодействия, молекулярная биология.

В современной медицине и молекулярной биологии наиболее широко применяются методы математического анализа различных показателей [1]. Они позволяют выявлять логические закономерности, причинно-следственные связи между биологическими факторами.

Целью настоящей работы является исследование применимости модели дерева решений для определения физиологического состояния спортсменов по анализируемым данным о их генотипах и фенотипах.

Исследование включало три основных этапа:

1). Сбор информации по наборам генов и основным медико-биологическим показателям, подготовка данных для автоматизированной подготовки и подачи на вход модели решающего дерева.

2). Построение решающего дерева с учётом выбора критерия разделения и алгоритма поиска ответов.

3). Визуализация ответов, построенных моделей и их решений, а также результатов анализа. Подведение итогов по оценке качества работы выбранных моделей классификации для решения задачи оценки состояния здоровья людей.

На первом этапе проведения исследования были получены наборы данных, содержащие медицинские, педагогические показатели, сведение о генотипах и фенотипах спортсменов. Эта информация была собрана таблицы, предварительно отредактирована и представлена в более удобном виде для дальнейшей автоматической обработки. С помощью Microsoft Excel таблицы данных были доработаны, текстовые описания формализованы и сохранены в файлах формата .csv. Основные данные по исследуемым генотипам и результатам проведённого биохимического анализа представлены в таблицах 1 и 2.

В таблице 1 по столбцам представлены некоторые гены, кодирующие белки, они являются входными параметрами для модели. В последнем столбце представлен класс, определяющий является ли нормой или отклонением соответствующий набор генов. Ген ACTN3 кодирует белок альфа-актинин-3, который стабилизирует сократительный аппарат скелетных мышц и участвует в метаболизме. Этот ген влияет на скорость и силу сокращения мышц, а также на риск повреждения мышц при физической нагрузке. Белок, который кодирует ген EPAS1, называется эндотелиальным белком с PAS-доменом 1 или гипоксически-индуцируемым фактором-2альфа. Он является членом семейства гипоксически-индуцируемых факторов, которые регулируют экспрессию многих генов, ответственных за адаптацию к низкому уровню кислорода. Например, этот белок стимулирует производство эритропоэтина, который увеличивает количество красных кровяных телец. Также он влияет на ангиогенез, вазодилатацию, глюконеогенез и другие процессы. Ген PPARD кодирует белок, который называется пероксисомным пролифератор-активированным рецептором дельта или PPAR-δ. Это один из трех типов PPAR, которые являются транскрипционными факторами, регулирующими метаболизм жиров, углеводов и белков. Белок PPAR-δ вовлечен в процессы окисления жирных кислот, термогенеза, ангиогенеза и воспаления. Белок PPAR-δ может стимулировать окисление жирных кислот в мышцах, что повышает их энергетическую эффективность и выносливость. Некоторые исследования показали, что полиморфизм гена PPARD, который изменяет активность белка, может быть связан с успехом в спорте, особенно в циклических видах, таких как плавание, бег или велоспорт. Ген PPARGC1A кодирует белок, который называется соактиватором 1 пероксисомного пролифератор-активированного рецептора гамма или PGC-1α. Это другой тип PPAR, который регулирует метаболизм глюкозы, жиров и белков, а также окислительный стресс и термогенез. Белок PGC-1α активирует гены, связанные с митохондриальной функцией и биогенезом, особенно в мышечной ткани. Он также может влиять на спортивную производительность, поскольку он участвует в адаптации мышц к физической нагрузке. Некоторые полиморфизмы гена PPARGC1A могут быть связаны с различными физиологическими показателями, такими как уровень холестерина, содержание жира в теле и выносливость. Ген VDR кодирует белок, который называется рецептором витамина D. Этот белок позволяет организму реагировать на витамин D. Белок VDR также является фактором транскрипции, то есть он регулирует выражение других генов, связанных с различными физиологическими процессами, такими как рост костей, иммунный ответ и обмен кальция и фосфора. Ген VDR имеет несколько полиморфизмов, которые могут влиять на активность белка и на риск развития некоторых заболеваний, таких как остеопороз, ревматоидный артрит или рак. Ген COL1A1 кодирует коллаген типа I. Коллагены это семейство белков, которые укрепляют и поддерживают многие ткани в организме, включая хрящи, кости, сухожилия, кожу и белую часть глаза (склеру). Некоторые варианты гена COL1A1 могут быть связаны с различными физическими характеристиками, такими как масса тела, индекс массы тела (ИМТ), плотность костной ткани и риск развития остеопороза. Также некоторые исследования показали, что ген COL1A1 может влиять на выбор вида спорта, так как он определяет тип мышечных волокон (быстрые или медленные) и скорость восстановления после физической нагрузки.

Таблица 1. – Собранные данные по генам белков, отвечающих за метаболизм

ACTN3	EPAS1	PPARD	PPARGC1A	VDR	COL1A	Вариант разбиения по группам
CT	AA	TT	Gly/Ser	T/C	GT	группа 1 (норма)
CT	AA	TT	Gly/Ser	T/C	GT	группа 1 (норма)
CC	AA	CT	Gly/Ser	T/C	TT	группа 2 (отклонение)
CC	GG	CT	Gly/Ser	C/C	GT	группа 1 (норма)
CC	AA	TT	Ser/Ser	C/C	TT	группа 1 (норма)
CT	AA	CT	Ser/Ser	T/C	TT	группа 1 (норма)
TT	GA	TT	Ser/Ser	T/T	GT	группа 2 (отклонение)
CC	AA	TT	Ser/Ser	T/C	GT	группа 2 (отклонение)
CT	GG	TT	Ser/Ser	T/C	GT	группа 2 (отклонение)
CT	GA	CT	Ser/Ser	T/T	GT	группа 2 (отклонение)
CC	GA	TT	Ser/Ser	T/C	GT	группа 1 (норма)
CT	GA	TT	Ser/Ser	T/C	GT	группа 1 (норма)
CC	GA	TT	Ser/Ser	C/C	GT	группа 1 (норма)
CT	GA	TT	Ser/Ser	T/T	GT	группа 1 (норма)

Было установлено, что проблематично выделить какую-либо градацию по степени влияния гена. Единственный ген, который можно выделить как однозначно оказывающий влияние на исследуемую проблематику - это ген VDR. Для решения проблемы определения степени влияния групп генов на проявление заболевания и следовательно физиологическое состояние спортсменов.

В таблице 2 представлены данные по полу спортсменов, окружающей среде, где чаще всего проводят время исследуемые люди, виды спорта, в которых они являются чемпионами. Также содержатся биохимические показатели, как степени изменения количества магния, кальция и С-телопептидов в организме. Так, магний является важным питательным элементом для работы мышц. Дефицит магния может вызывать мышечные судороги, нарушения сердечного ритма и способствовать нарушению сна, а также быть сопутствующим состоянием при наличии избыточного веса. При занятиях спортом, особенно интенсивных, в организме происходит потеря магния. Особенно это касается тяжелых видов спорта и физических нагрузок, например участия в беговых марафонах или интенсивного плавания. Кальций важен для костных тканей: спортсмены часто испытывают повышенную нагрузку на свои кости, поэтому важно, чтобы они получали достаточное количество кальция для поддержания здоровья своих костей. Кальций помогает укреплять кости и уменьшать риск развития остеопороза.

Эти и другие наборы показателей, собранные в ходе медицинских исследований, были предобработаны и формализованы в виде, пригодном для интеллектуального анализа статистическими методами и вычислительными средствами, в частности, деревом решений.

На следующем шаге было смоделировано дерево решений. Оно является одним из автоматизированных методов многомерного анализа данных, входящих в технологию Data Mining, отличающийся наглядностью и удобством представления закономерностей [2]. Решающее дерево представляет собой древовидный граф – структуру данных, состоящую из узлов принятия решений, соединенных друг с другом ребрами.

Таблица 2. – Результаты биохимического анализа группы спортсменов

N	Gender	Environment	Kind of sport	$Mg_2 - Mg_1$	$Ca_2 - Ca_1$	α-CTx_2-α-CTx_1, пг/мл
1	male	hall	box	0,04	-0,07	-524,6
2	male	hall	box	-0,95	-0,05	-161,0
3	female	hall	judo	-0,06	0,16	-223,5
4	female	hall	athletics sprint	-0,02	-0,13	-375,6
5	female	hall	athletics sprint	0,05	-0,15	-395,9
6	female	street	athletics sprint	0,13	-0,32	-545,7
7	female	street	athletics throwing	-0,04	0,19	-1009,6
8	female	street	athletics throwing	0,00	-0,16	15,8

В дереве имеется один особый узел, именуемый корневым узлом. Другие особые узлы, находящиеся в конце любой цепочки подряд идущих ребер, называют листовыми узлами. Дерево построено на основании обучающей выборки, содержащей информацию о значениях входных переменных, содержащих информацию по генам и медицинским показателям спортсменов, и соответствующих значениях прогнозируемого показателя, относящих его к определённому классу готовности к соревнованиям. Узел принятия решений обеспечивает проверку условия на значение входной переменной, а каждое ребро обозначает один из возможных вариантов. При обучении дерева решений использован индекс Джини как мера качества его обучения. Он рассчитывается по следующей формуле.

$$G = \sum_{i=1}^{C} p_i \times (1 - p_i)$$

где G – индекс Джинни, C – общее количество классов, p_i – вероятность выбора элемента данных с классом i.

В настоящей работе средствами языка программирования python 3 и моделью DecisionTreeClassifier библиотеки sklearn построены и визуализированы деревья классификации с учётом выбранного параметра разделения [3].

Для целевого параметра, определяющего здоров или болен спортсмен, было смоделировано решающее дерево [4], представленное на рисунке 1. Полученная структура имеет 58 вершин (листьев), из которых 30 являются терминальными (листьями). По рисунку можно наглядно увидеть ход алгоритма решения, где на каждом шаге с помощью коэффициента Джини определялся наиболее информативный признак и определялась граница разделения на два класса. В итоге, в конечных листьях визуализированного дерева определены и помечены классы, к которым дерево отнесло анализируемые генотипы из обучающего набора данных.

Для более глубокого анализа также было обучено и выведено решающее дерево, определяющее пол спортсмена по поданному на вход набору гену (см. рисунок 2).

Аналогично были построены деревья решений, относящие людей из анализируемых групп к классу подготовленных к соревнованиям или к классу с отклонениями от норм по полученным результатам биохимического анализа.

После реализации данных моделей классификации, была проведена оценка точности работы этих алгоритмов. Для этого применён метод *score*, определяемый как отношение числа правильно определённых классов к числу всех решений модели [5].

Например, было выявлено, что для дерева, отображённого на рисунке 1, *score* составил 88,33%, для дерева решений на рисунке 2 – 86,67%.

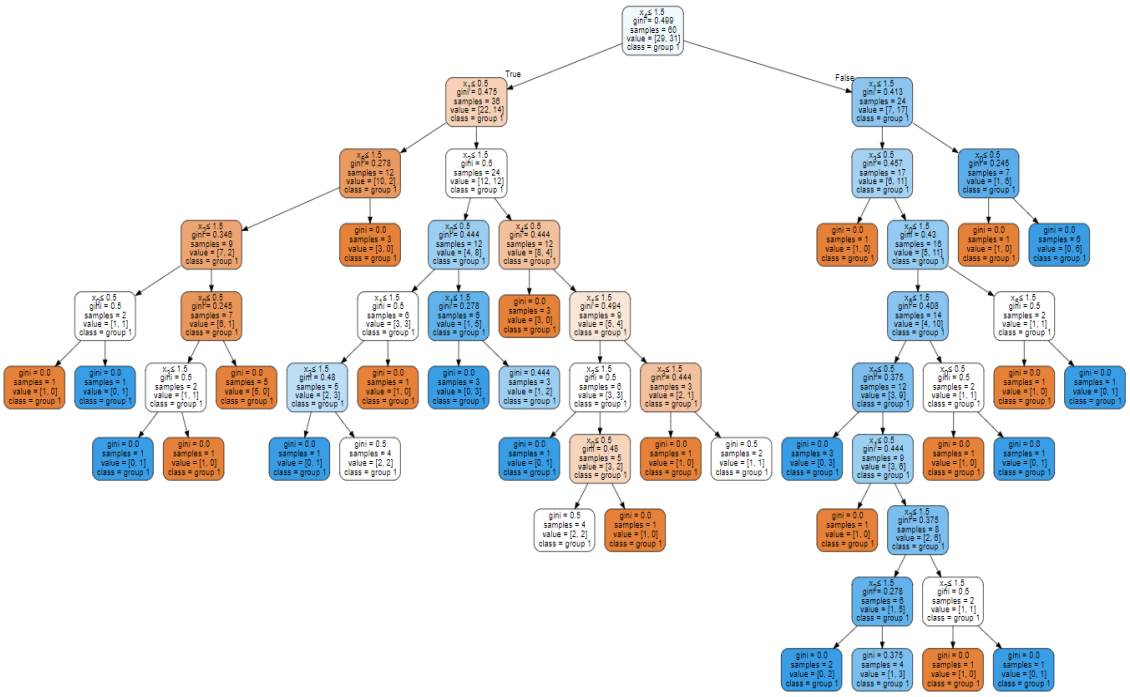

Рисунок 1. – Дерево решений, построенное для анализа генотипов и определения физиологического состояния спортсменов

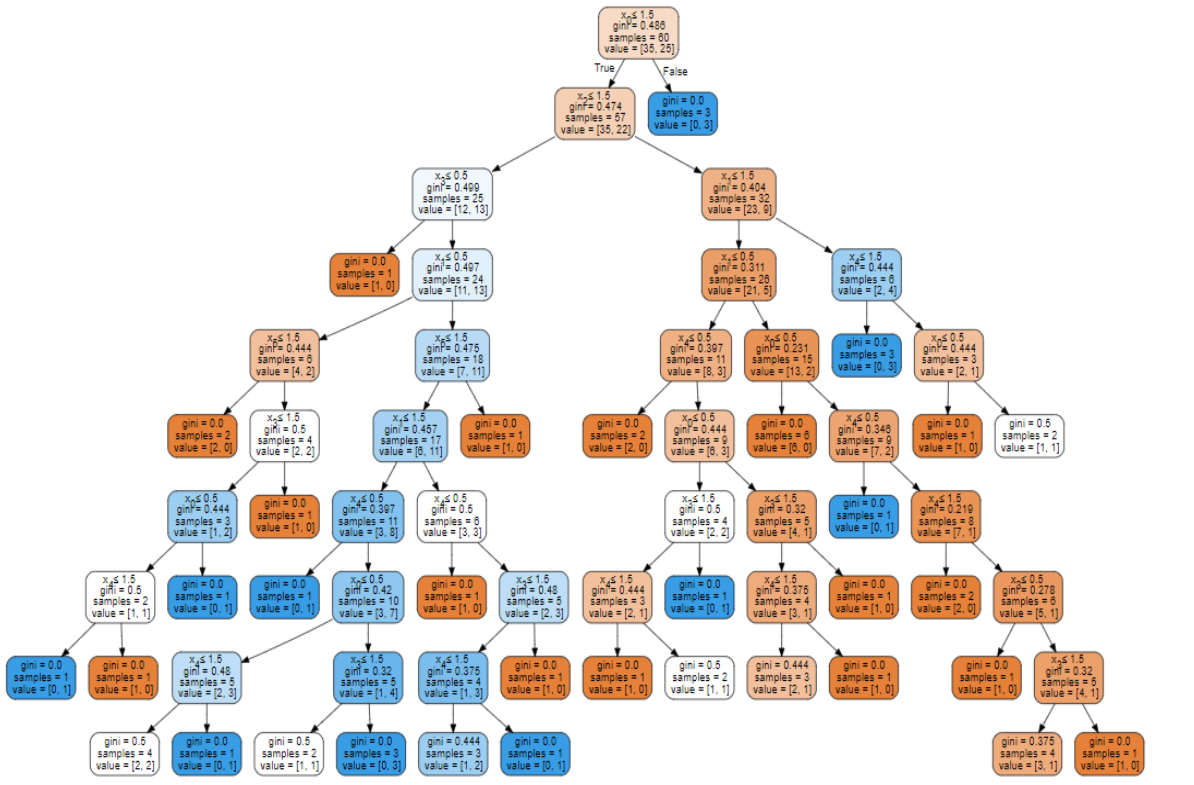

Рисунок 2. – Дерево решений, построенное для определения пола спортсмена по набору генов

Данные значения позволяют сделать заключение, что данные алгоритмы решений машинного обучения позволяют с достаточной уверенностью использовать их для выявления степени подготовленности спортсменов к соревнованиям и тренировкам, определять риски заболеваний, а также исследовать всевозможные межгенные взаимодействия и их влияние на физиологическое состояние людей.

Список использованных источников

1. Гублер Е.В. Вычислительные методы анализа и распознавания патологических процессов. Л. 1978.

2. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян и др. – СПб.: БХВПетербург, 2004.

3. Чубукова, И.А. Data Mining: учебное пособие. / И.А. Чубукова. – М.: Интернет–Университет Информационных Технологий; БИНОМ. Лаборатория знаний, 2006.

4. Кормен, Т., Лейзерсон, Ч., Ривест, Р., Штайн, К. Алгоритмы: построение и анализ, 2-е издание. : Пер. с англ. – М. : Издательский дом “Вильямс”, 2011. – 1296 с. : ил. – Парал. тит. англ.

5. Classification and Regression Trees / L. Breiman et al. – Wadsworth, Belmont, CA, 1984.