Анализ и прогнозирование индивидуального риска развития онкологических заболеваний на основе данных об образе жизни, окружающей среде и генетике

А.Д. Онучина
Научный руководитель
Д.А. Смирнова
МАОУ СОШ № 57, Калининград, Россия
Email: arinadenisovna388@gmail.com

Работа представляет собой практико-ориентированный проект, направленный на разработку системы поддержки принятия врачебных решений, обеспечивающей прогнозирование индивидуального риска развития онкологических заболеваний на основе факторов образа жизни, окружающей среды и наследственной предрасположенности. Актуальность темы определяется высокой распространенностью и смертностью от онкологических заболеваний в мире, необходимостью ранней профилактики и персонализированной медицины, а также возрастающей доступностью медицинских и поведенческих данных, позволяющих применять методы статистики и машинного обучения для оценки индивидуального риска. Цель работы заключается в создании и сравнительной оценке прогностических моделей, позволяющих по совокупности индивидуальных характеристик определять вероятность развития одного из пяти распространенных типов рака (молочной железы, простаты, кожи, толстой кишки, легких). Для достижения цели решались задачи: предварительная обработка набора данных; статистический анализ значимости факторов; построение и настройка алгоритмов машинного обучения; выбор наилучших моделей и оценка их точности; разработка концепции веб-интерфейса для практического применения.

Исследование выполнено на основе базы данных из 2000 наблюдений (возраст пациентов 25-90 лет), включающей демографические показатели, факторы образа жизни и среды, а также биомаркеры и наследственные характеристики. Статистический анализ показал значимое влияние на целевую переменную факторов питания (потребление кальция, красного мяса, обработанных продуктов, фруктов и овощей), физической формы (ожирение, уровень физической активности), окружающей среды (загрязнение воздуха, профессиональные риски), а также возраст и курение. Анализ с помощью методов машинного обучения подтвердил, что наиболее информативными предикторами типа онкологического заболевания являются факторы питания, биологического пола, а также физической формы.

Далее для разработки системы прогнозирования индивидуального риска развития онкологических заболеваний была исследована эффективность ряда алгоритмов машинного обучения (метод k-ближайших соседей, линейная и гребневая регрессия, дерево решений, случайный лес, градиентный бустинг). При обучении алгоритмов применялся метод кросс-валидации на пять фолдов и оптимизация гиперпараметров.

Главным результатом является разработка пяти прогностических моделей – по одной для каждого типа рака – с наилучшими показателями у линейной и гребневой регрессии: средний корень из среднеквадратичной ошибки (RMSE) равен 0,047 ± 0,005; средний коэффициент детерминации R^2 равен 0,833 ± 0,025. Полученные результаты свидетельствуют о высокой объясняющей способности и практической применимости. Предусмотрена практическая реализация в виде веб-интерфейса с анкетированием пользователя и автоматическим расчетом индивидуального индекса риска. В ходе работы показано, что интеграция статистического анализа факторов и методов машинного обучения позволяет достоверно прогнозировать индивидуальный риск развития различных типов рака и может служить основой персонализированных профилактических рекомендаций.

Возьмите на заметку:
По данным об образе жизни и наследственности можно достаточно точно оценить индивидуальный риск развития пяти распространенных видов рака. При этом, факторы питания и физической формы оказывают наибольшее влияние на прогноз риска развития рака.