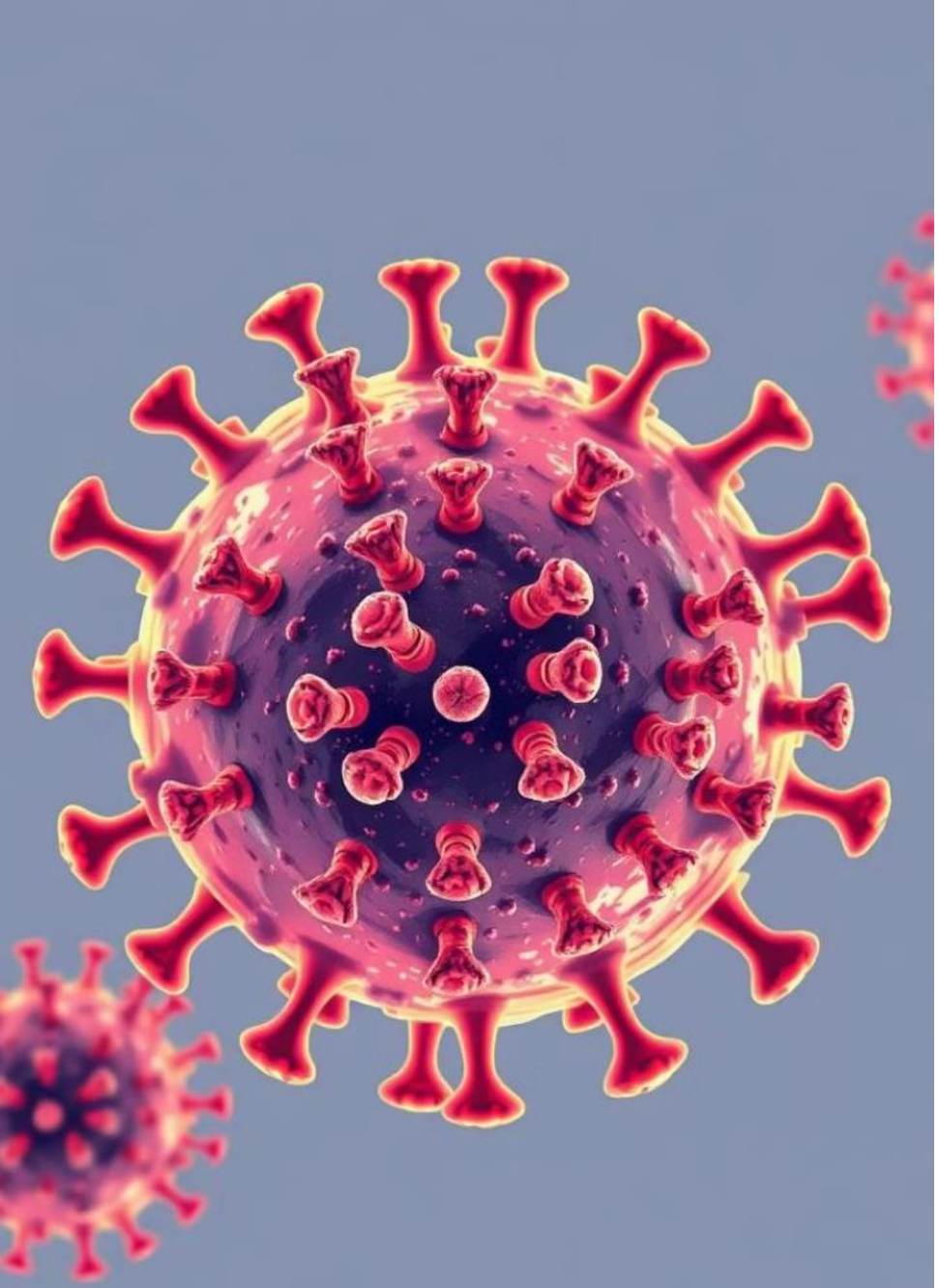


Использование методов машинного обучения для идентификации типа вируса

В наше время, когда респираторные вирусы являются большой угрозой, точная идентификация их типа становится все более важной. В этой презентации мы исследуем применение ансамблевых методов машинного обучения и статистических тестов для достижения этой цели.

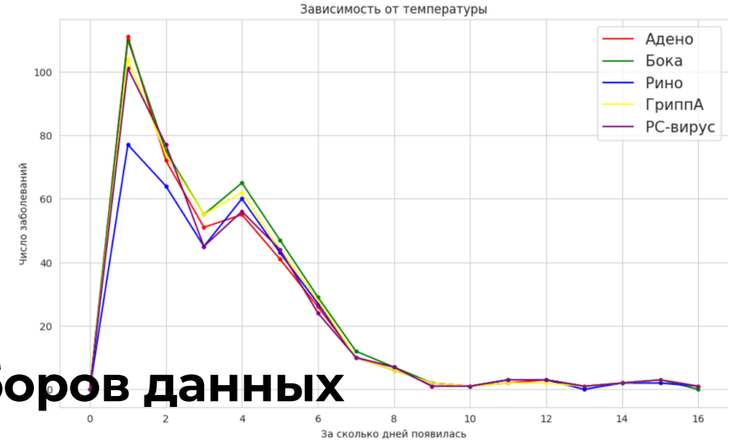
Елисова Анастасия Алексеевна

13 декабря 2024г

Актуальность проблемы и цели исследования

Точная и своевременная диагностика респираторных инфекций жизненно необходима для своевременного лечения, предотвращения распространения и снижения рисков. В этой презентации мы рассмотрим как различные тесты и методы в машинном обучении могут улучшить процесс идентификации типа вируса, обеспечивая точные и эффективные диагностические решения.

Описание наборов данных симптомов



	Количество пулей	Количество единиц
Адено	108	1425
Бока	62	1471
Рино	313	1220
ГриппА	157	1376
РС-вирус	211	1322

Описание наборов данных

Для обучения модели машинного обучения мы будем использовать наборы данных, содержащие информацию о симптомах, лабораторных анализах и других факторах, связанных с вирусными инфекциями. Набор данных был получен из ММКЦ “Коммунарка”. Он включает в себя информацию о 1533 пациентах, у которых были выявлены пять основных типов вирусов: гриппА, риновирус, аденовирус, респираторно-синцитиальный вирус и бокавирус

Описание данных симптомов

Симптомы респираторных заболеваний, такие как кашель, насморк, боль в горле, головная боль, лихорадка, температура за несколько дней до поступления и другие будут использоваться для обучения модели. Будут рассмотрены различные комбинации симптомов для определения характерных признаков каждого типа вируса.

Обзор типов респираторных вирусов

1 Грипп А

Вирус гриппа А является одной из наиболее распространенных причин сезонных эпидемий гриппа. Симптомы включают лихорадку, кашель, боль в горле, головную боль, мышечные боли, усталость и заложенность носа.

2 Риновирус

Риновирус является наиболее частой причиной обычной простуды. Симптомы обычно включают насморк, заложенность носа, чихание, кашель, боль в горле и головную боль.

3 Аденовирус

Аденовирус может вызывать различные респираторные инфекции, включая простуду, бронхит, пневмонию и конъюнктивит. Симптомы включают кашель, насморк, боль в горле, лихорадку, головную боль и боль в мышцах.

4 Респираторно-синцитиальный вирус (РС-вирус)

РС-вирус является распространенной причиной респираторных инфекций у детей, особенно у младенцев. Симптомы включают кашель, насморк, затрудненное дыхание, свистящее дыхание и лихорадку.

5 Бокавирус

Бокавирус является распространенным вирусом, который может вызывать респираторные инфекции у детей. Симптомы обычно включают кашель, насморк, лихорадку и боль в горле.

Методология: Тест Фишера, Тест Манна-Уитни и Перестановочный тест с формульным ообъяснением

Тест Фишера

Тест Фишера - это тест, который используется для сравнения двух групп. Он помогает определить, существуют ли статистически значимые различия в распределении данных в группах. Формула:

$$p_{value} = \frac{(a + b)!\,(c + d)!\,(a + c)!\,(b + d)!}{a!\,b!\,c!\,d!\,n!}$$

	Group 1	Group 2	Row Total
Category 1	a	b	a+b
Category 2	c	d	c+d
Column Total	a+c	b+d	a+b+c+d = n

Тест Манна-Уитни

Тест Манна-Уитни используется для проверки гипотезы о том, что две выборки взяты из распределений с одинаковыми медианами, или что одно распределение доминирует над другим, то есть, значения из одной группы чаще больше значений из другой группы. Это особенно полезно, когда данные не соответствуют предположению о нормальности, и, следовательно, не могут быть проанализированы с помощью стандартного t-теста.

$$U_1 = R_1 - \frac{m(m+1)}{2} \qquad U_2 = R_2 - \frac{n(n+1)}{2}$$

$$\mathrm{U} = \min(U_1, U_2)$$

Перестановочный тест

Перестановочный тест - это статистический метод, используемый для определения достигнутого уровня значимости теста, основанный на алгоритме с перестановками внутри выборок.

```
P-value = (количество перестановок, которые дают более экстремальное значение статистики, чем наблюдаемое) / (общее количество возможных перестановок)
```

Метод статистически взвешенных синдромов

Метод статистически взвешенных синдромов основан на принятии коллективных решений по системам синдромов двумерных областей признакового пространства, в котором преобладают объекты одного из распознаваемых классов. Области задаются с помощью границ, которые находятся через оптимальные разбиения интервалов значений признаков

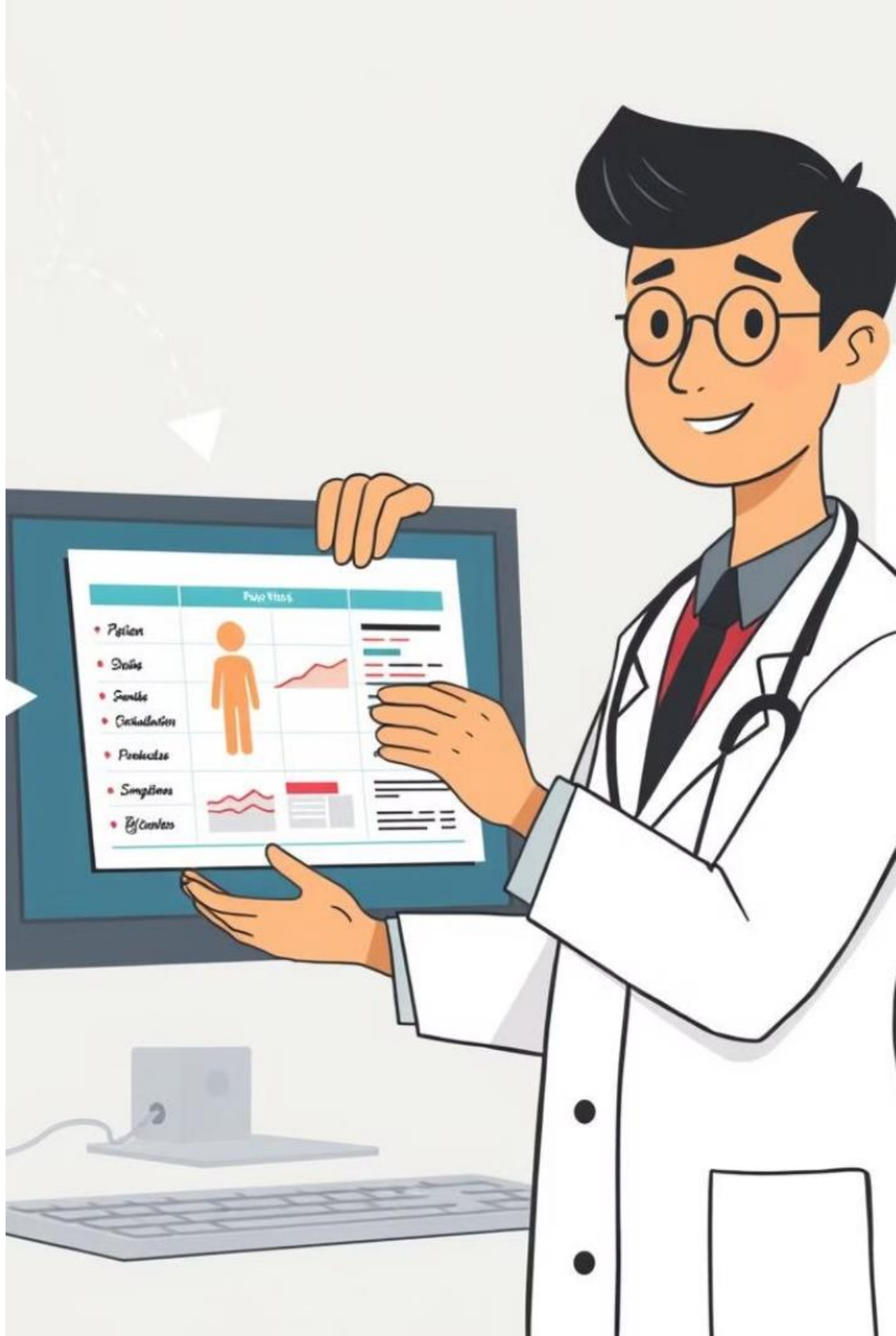
$$F_s(K_j, S_t, R) = \sum_{i=1}^{t} \left(v_0^j - v_i^j\right)^2 m_i$$

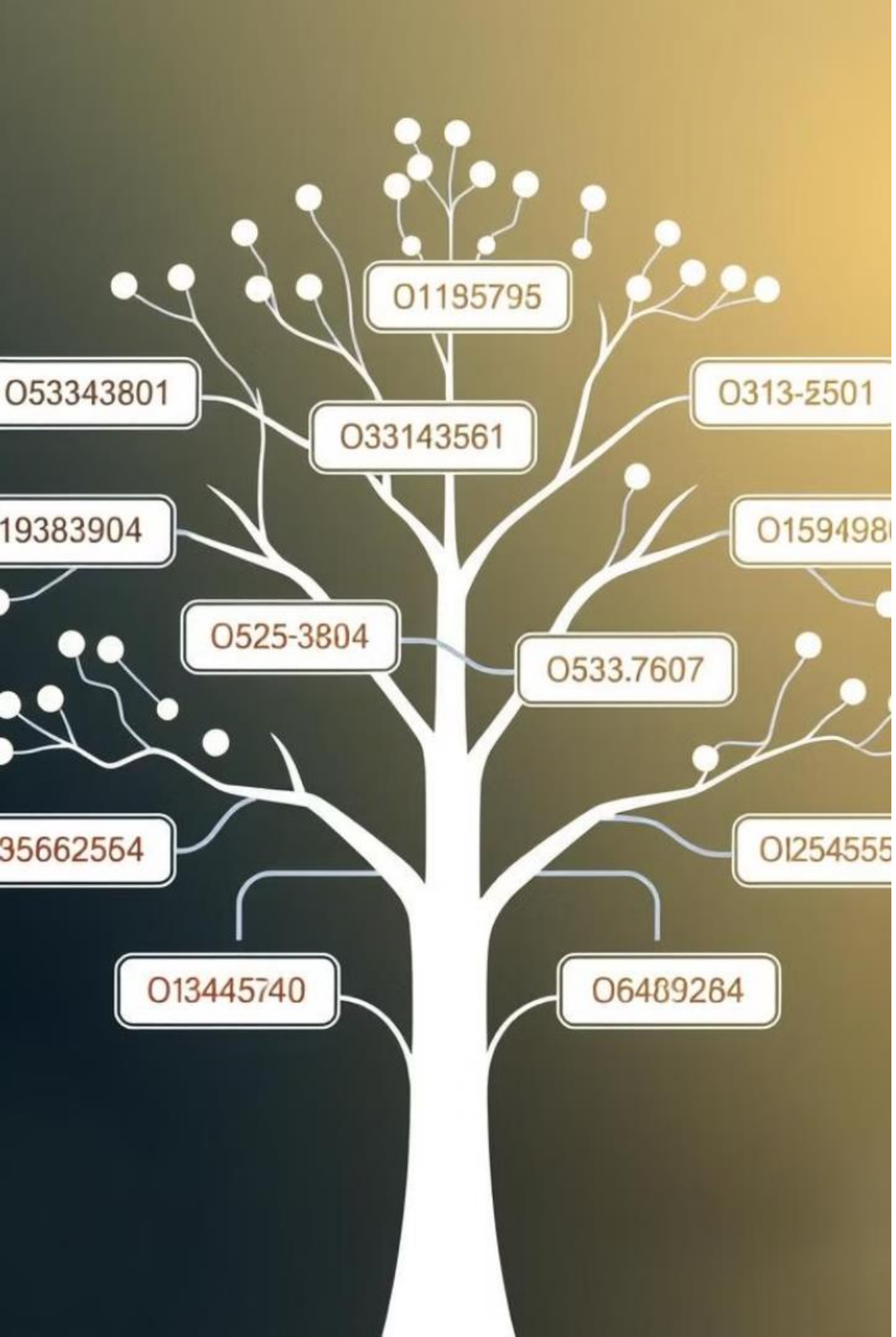
$$F_L(K_j, S_t, R) = \max_{i \in \{1..,t\}} \left(v_0^j - v_i^j\right)^2 m_i$$

$$\Gamma_j(s) = \frac{\sum_{i=1}^{r} w_i v_i^j}{\sum_{i=1}^{r} w_i}$$

$$w_i = \frac{m_i}{m_i + 1} \frac{1}{v_i^j (1 - v_i^j)}$$

Где F_s - интегральный, F_L локальный функционалы, r – элементы разбиения, v – доля объектов классов, m – число объектов из описания r.





Применение ансамблевых методов для классификации типов вирусов

Случайный лес

Случайный лес - это ансамбль из множества деревьев решений, которые обучаются на случайных подмножествах данных. Он может справляться с высокоразмерными данными и может иметь высокую точность.

Градиентный бустинг

Градиентный бустинг - это алгоритм, который объединяет несколько слабых моделей в более сильную. Он часто демонстрирует высокую точность и устойчивость к переобучению.

Бэггинг

Бэггинг - это метод, который использует повторную выборку из обучающего набора для создания множества моделей. Он может улучшить точность и устойчивость к переобучению.

Анализ результатов:

0.93

Точность

Точность показывает процент правильных предсказаний среди всех предсказаний.

0.718

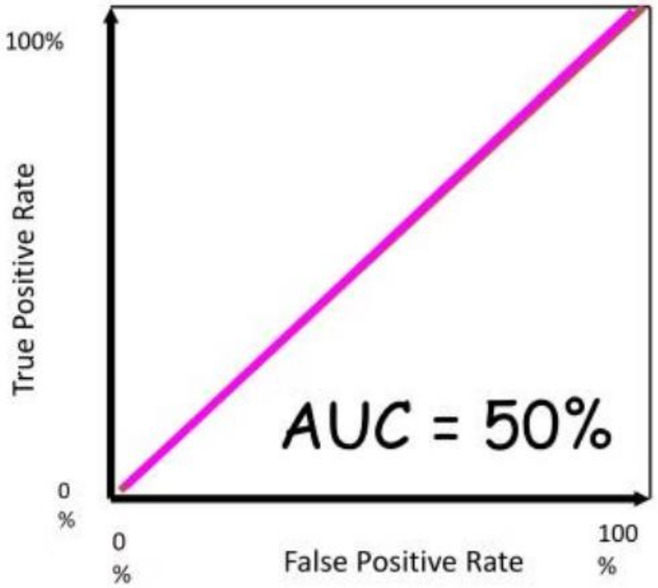
Roc-auc

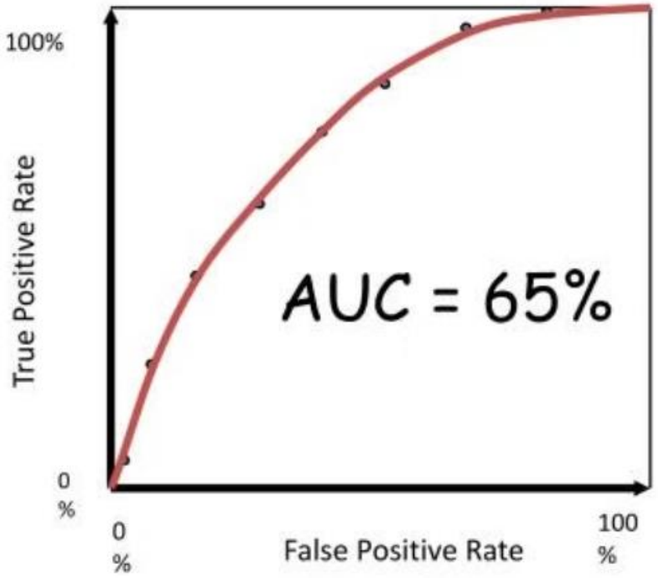
Значение roc-auc для лучшего метода Catboost с использованием статистических тестов

0.73

Roc-auc

Значение, полученное с использованием метода статистически взвешенных синдромов





Подведение итогов

Представленные подходы на основе машинного обучения открывают новые возможности для повышения качества и доступности диагностики вирусных инфекций. Дальнейшее развитие этих методов, их внедрение в реальную клиническую практику и интеграция с другими передовыми технологиями имеют большой потенциал для улучшения системы здравоохранения.