СРАВНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ БИНАРНОЙ КЛАССИФИКАЦИИ МЕДИЦИНСКИХ ДАННЫХ

Г.Х. Ташматов, О.Б. Рузибаев

Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий, Ташкент, Узбекистан

(E-mail: ganijon.tashmatov@gmail.com, o.ruzibaev@tuit.uz)

Аннотация. В работе выполнено сравнение шести алгоритмов машинного обучения для задачи бинарной классификации медицинских наблюдений на наборе Breast Cancer Wisconsin (Diagnostic). Проведены аудит данных, стратифицированное разбиение выборки, построение пайплайнов предобработки, подбор параметров и оценивание моделей по метрикам Accuracy, Precision, Recall, F1 и ROC-AUC. Наилучший ROC-AUC на отложенной тестовой выборке показал Random Forest (0,997), тогда как максимальные Accuracy и F1 были получены для SVM с RBF-ядром. Результаты демонстрируют, что ансамблевые и margin-based методы обеспечивают устойчивое качество на компактном медицинском датасете.

Ключевые слова: машинное обучение, бинарная классификация, медицинские данные, ROC-AUC, Random Forest, SVM, Breast Cancer Wisconsin.

1. Введение

Сравнение классических алгоритмов машинного обучения на одном и *том 19же наборе медицинских признаков является важной практической задачей.* Такой подход позволяет определить, какие модели лучше работают при

умеренном размере выборки, высокой коррелированности признаков и необходимости интерпретации результата. Цель исследования заключается в построении воспроизводимого мини-цикла анализа данных: от аудита и подготовки выборки до сравнения моделей, визуализации результатов и формального вывода о качестве классификации.

В качестве объекта исследования использован открытый диагностический набор Breast Cancer Wisconsin (Diagnostic). Он содержит 569 наблюдений и 30 вещественных признаков, рассчитанных по цифровым изображениям тонкоигольной аспирационной биопсии молочной железы. Задача сформулирована как бинарная классификация наблюдений на доброкачественные и злокачественные случаи.

Параметр	Значение
Набор данных	Breast Cancer Wisconsin (Diagnostic)
Число наблюдений / признаков	569 / 30
Постановка задачи	Бинарная классификация: malignant vs benign
Классы	357 benign и 212 malignant
Источник	UCI Machine Learning Repository; локальный доступ через sklearn.datasets.load_breast_cancer()

Таблица 1. Контекст эксперимента и происхождение данных

2. Методы

Первичный аудит показал отсутствие пропусков и полных дубликатов строк. Несмотря на это, в рабочие пайплайны был включён медианный импьютер, чтобы сделать эксперимент устойчивым к возможным изменениям структуры данных. Для моделей, чувствительных к масштабу признаков,

применялась стандартизация. Данные были разделены на обучающую и тестовую части в пропорции 80/20 со стратификацией по целевой переменной.

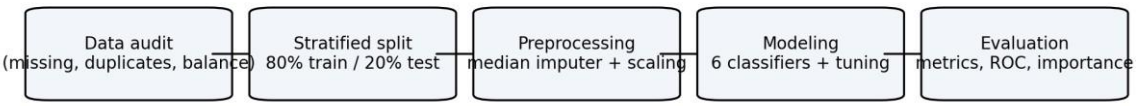


Рис. 1. Последовательность экспериментального пайплайна

На этапе настройки использовался поиск по сетке на обучающей части с 5-кратная кросс-валидация и целевой метрикой ROC-AUC. После подбора параметров каждая модель дообучалась на всей обучающей части и проверялась на отложенном тесте. В исследование включены Logistic Regression, SVM с RBF-ядром, k-NN, Gaussian Naive Bayes, Random Forest и Gradient Boosting.

Для оценки качества использовались стандартные показатели классификации, основанные на количестве истинно положительных, истинно отрицательных, ложноположительных и ложноотрицательных решений:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{F1} = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall})$$

Алгоритм	Назначение	Итоговые параметры
Logistic Regression	Линейная вероятностная модель; базовая интерпретируемость	C=1

SVM (RBF)	Нелинейная граница решения; устойчивость к сложным разделениям	C=2; gamma=scale
k-NN	Локальный instance-based подход; зависит от масштаба признаков	n_neighbors=3; weights=distance
Gaussian Naive Bayes	Быстрый вероятностный базовый метод	var_smoothing=1e-11
Random Forest	Ансамбль деревьев; устойчивость к нелинейностям	n_estimators=200; min_samples_split=5; class_weight=balanced
Gradient Boosting	Последовательный ансамбль слабых моделей	n_estimators=200; learning_rate=0.1; max_depth=1; subsample=0.8

Таблица 2. Сравниваемые алгоритмы и итоговые настройки

3. Результаты

Визуальный анализ показал умеренный дисбаланс классов и высокую коррелированность части признаков. Наиболее сильную связь с целевой переменной имеют признаки, связанные с concave points, perimeter, radius и area. Это указывает на то, что различия между классами в данном датасете во многом определяются морфометрическими характеристиками клеточного ядра.

Количественное сравнение показало, что все шесть моделей достигли высокого качества. Однако между ними имеются различия: Random Forest дал максимальный ROC-AUC на тесте, а SVM обеспечил лучшую Accuracy и F1. k-NN и Gaussian Naive Bayes ожидаемо уступили ансамблевым и marginbased методам, хотя сохранили качество выше 0,92 по Accuracy.

Модель	CV ROCAUC	Accuracy	Precision	Recall	F1	Test ROCAUC
Random Forest	0,989	0,965	1,000	0,905	0,950	0,997
Logistic Regression	0,996	0,974	0,976	0,952	0,964	0,996
SVM (RBF)	0,995	0,982	1,000	0,952	0,976	0,995
Gradient Boosting	0,994	0,965	1,000	0,905	0,950	0,991
Gaussian Naive Bayes	0,987	0,921	0,923	0,857	0,889	0,989
k-NN	0,989	0,939	0,973	0,857	0,911	0,982

Таблица 3. Сравнение качества классификаторов на отложенном тесте

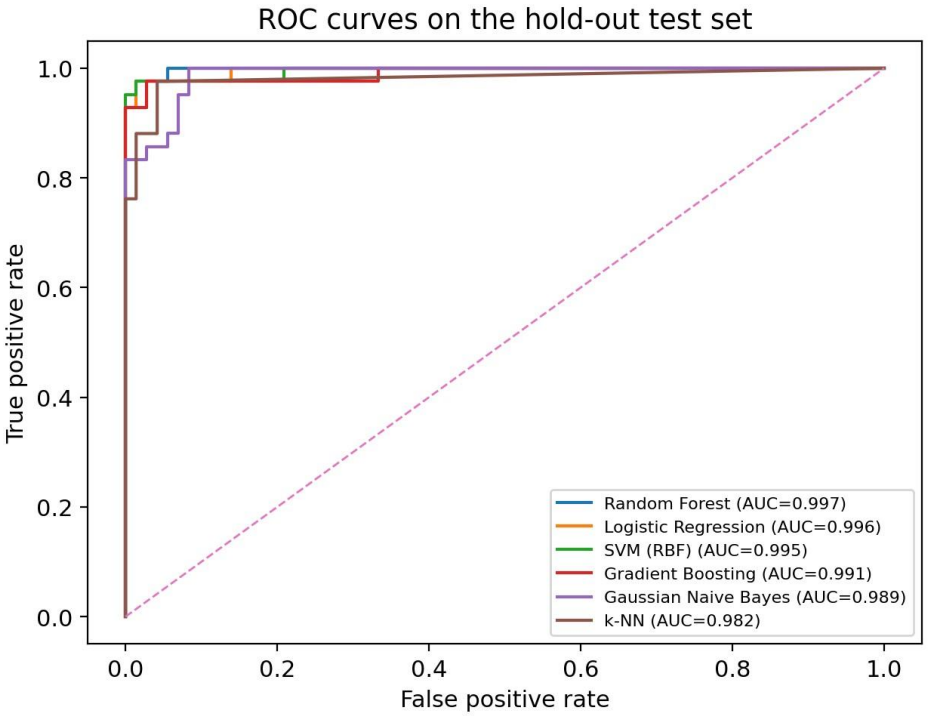
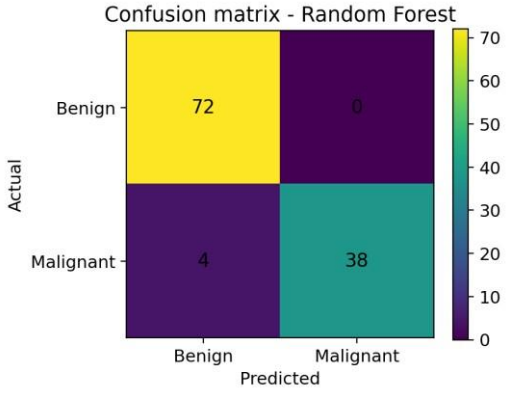


Рис. 2. ROC-кривые сравниваемых моделей на тестовой выборке

Если основной целевой метрикой считать ROC-AUC, то лучшей моделью становится Random Forest. Для неё доверительный интервал бутстреп-оценки Accuracy составил [0,930; 0,991], а для ROC-AUC — [0,992; 1,000]. Матрица ошибок показывает малое количество ошибок обоих типов, а анализ permutation importance подтверждает высокую значимость признаков worst concave points, worst perimeter, worst radius и worst area.



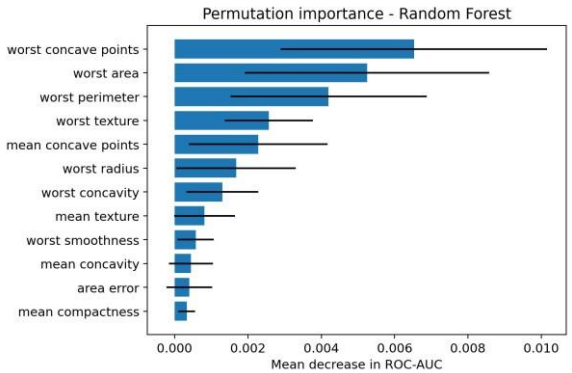


Рис. 3. Матрица ошибок и permutation importance для модели Random Forest

4. Обсуждение

Полученные результаты показывают, что даже на сравнительно компактном медицинском наборе данных классические модели машинного обучения могут обеспечивать почти экспертное разделение классов. Это объясняется высокой информативностью исходных морфометрических признаков и качеством разметки. С практической точки зрения выбор между Random Forest и SVM зависит от приоритета: если важна вероятностная ранжировка и устойчивость по ROC-AUC, целесообразен Random Forest; если важнее итоговая точность и F1 при текущем пороге решения, уместен SVM.

Следует учитывать ограничения эксперимента. Набор данных относительно небольшой и хорошо изученный, поэтому значения метрик могут быть оптимистичными по сравнению с реальной клинической практикой. Кроме того, эксперимент выполнялся на открытом датасете, а не на вновь собранных клинических данных. В дальнейшем целесообразно добавить внешние датасеты, nested cross-validation, калибровку вероятностей
и методы объяснимого ИИ, например SHAP или локальные контрфактуальные объяснения.

5. Заключение

Выполнен полный демонстрационный цикл исследования: аудит данных, подготовка пайплайнов, обучение шести алгоритмов, сравнение метрик, построение визуализаций и оформление результата в формате научной статьи. На данном наборе данных лидером по ROC-AUC стала модель Random Forest (0,997), тогда как SVM продемонстрировал наилучшие Accuracy и F1. Таким образом, поставленная задача была решена, а эксперимент показал, что

в компактной вычислительной среде можно сформировать воспроизводимый научный отчёт по классической задаче классификации.

Список литературы

1. Wolberg W., Mangasarian O., Street N., Street W. Breast Cancer Wisconsin (Diagnostic) [Dataset]. UCI Machine Learning Repository, 1993. DOI: 10.24432/C5DW2B.

2. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.

3. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45, No. 1. P. 5–32.

4. Cortes C., Vapnik V. Support-vector networks // Machine Learning. 1995. Vol.
20. P. 273–297.

5. Cover T., Hart P. Nearest neighbor pattern classification // IEEE Transactions on Information Theory. 1967. Vol. 13, No. 1. P. 21–27.

6. Friedman J. Greedy function approximation: A gradient boosting machine // Annals of Statistics. 2001. Vol. 29, No. 5. P. 1189–1232.