

ФАКУЛТЕТ ПО ИЗЧИСЛИТЕЛНА ТЕХНИКА И АВТОМАТИЗИЦИЯ

ИНТЕГРИРАН ПОДХОД ЗА АНАЛИЗ НА СТУДЕНТСКИТЕ АНКЕТИ ЧРЕЗ СТАТИСТИКА И МАШИННО ОБУЧЕНИЕ



Ръководител: доц. д-р инж. Нели Арабаджиева-Калчева — СИТ

Работен колектив: 18 участници — преподаватели, докторанти, инженери и студенти

1. доц. д-р инж. Ивайло Пенев — КНТ
2. доц. д-р инж. Диян Динев — СИТ
3. гл. ас. д-р Мая Тодорова — СИТ
4. гл. ас. д-р инж. Димитричка Николаева — СИТ
5. гл. ас. д-р инж. Евгения Ракитина-Куреши — СИТ
6. гл. ас. д-р Доника Стоянова — СИТ

7. гл. ас. д-р Даниела Петрова — СИТ
8. ас. инж. Геновева Костова — СИТ
9. преп. инж. Велислав Колесниченко — СИТ
10. преп. инж. Стефка Попова — СИТ
11. Дейвид Кирчев — студент, СИТ
12. Диян Димитров — студент, СИТ

13. Ясен Дамьянов — студент, СИТ
14. Александър Илиев — студент, КСТ
15. инж. Георги Соколов
16. Данаил Димитров — студент, КС
17. Росен Василев — студент, СИТ
18. Петър Сталев — студент, СИТ

1 Цел, данни и методика

Цел на изследването

Разработване на интегрирана и възпроизводима рамка за анализ на студентски анкети чрез статистически методи и машинно обучение. Акцентът е върху удовлетвореността, намеренията за продължаване на обучението, професионалната реализация и качеството на анкетните данни.

Използвани данни

- 507 студенти (2017–2024) — намерения за продължаване на обучението.
- 2 000 записа — прогнозиране на професионална реализация.
- 4 432 анкетни отговора — неангажирани и екстремни стилове.
- 87 884 потребителски отзива — многокласов sentiment анализ.
- Публични образователни набори — сравнение и валидация.

Методи и метрики

Използвани са Decision Tree, Random Forest, Gradient Boosting, Histogram-based GB, Logistic Regression, SVM, kNN, MLP/LSTM, Voting и хибридни ансамбли. Оценкаването се базира на Accuracy, Precision, Recall, F1-score, AP, MAE и balanced accuracy.

Научен принос и приложимост

- Интегриран работен процес за анализ на студентски анкети и образователни данни.
- Предимство на ансамбловите модели при прогнозиране и класификация.
- Python workflows за импорт, почистване, моделиране и визуализация.
- Ранно откриване на рискови студенти и подкрепа на вътрешната оценка на качеството.
- Приложимост в обучението по машинно обучение и обработка на естествен език.

2 Ключови резултати и интерпретация



97.8%

ТОЧНОСТ

Gradient Boosting



91.5%

ТОЧНОСТ

Random Forest



95.5%

ТОЧНОСТ

SVM RBF



>93%

средна прецизност

Random Forest

Анализ на резултатите

- 1 При прогнозиране на намеренията за висше образование (Q1) с Random Forest (фиг. 1) верните класификации са по диагонала, а грешките са главно между съседни ординални категории; клас 4 остава най-труден за разпознаване.
- 2 При ансамбловите модели за намеренията за продължаване на образованието (фиг. 2) Voting и CatBoost разпознават най-добре положителните случаи. Voting е по-балансиран, а Stacking и LightGBM са с по-нисък Recall.
- 3 В многокласовия sentiment анализ (фиг. 3) крайните класове се разпознават по-успешно, а нестемираните данни се представят по-добре. При професионалната реализация (фиг. 4–5) Random Forest отчита 174 TN, 18 FP, 16 FN, 193 TP и Average Precision = 0.93.
- 4 Сравнението на десетте алгоритъма (фиг. 6) поставя Gradient Boosting, Random Forest и Histogram-based Gradient Boosting на първите места по accuracy. При спам класификацията (фиг. 7–8) броят възли влияе върху точността, а SVM + Logistic Regression + MLP достига над 99.5%.

3 Експериментални резултати



Фиг. 1. Матрица на обръкване – Random Forest npu прогнозиране на намеренията за висше образование



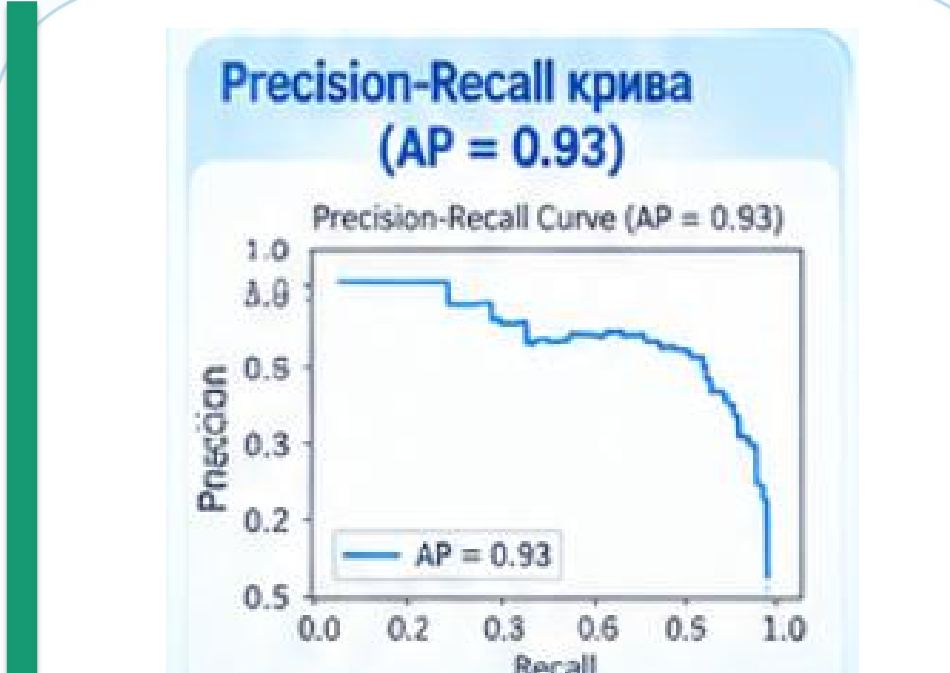
Фиг. 2. Матрици на обръкване – ансамблови модели



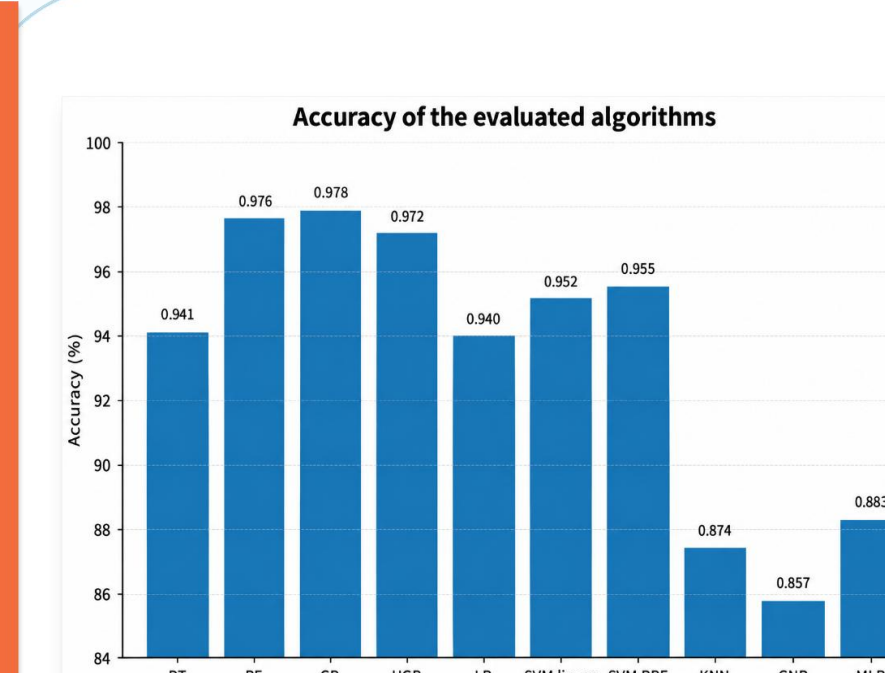
Фиг. 3. Нормализирана матрица на обръкване – балансиран набор без стеминг (5 класа)



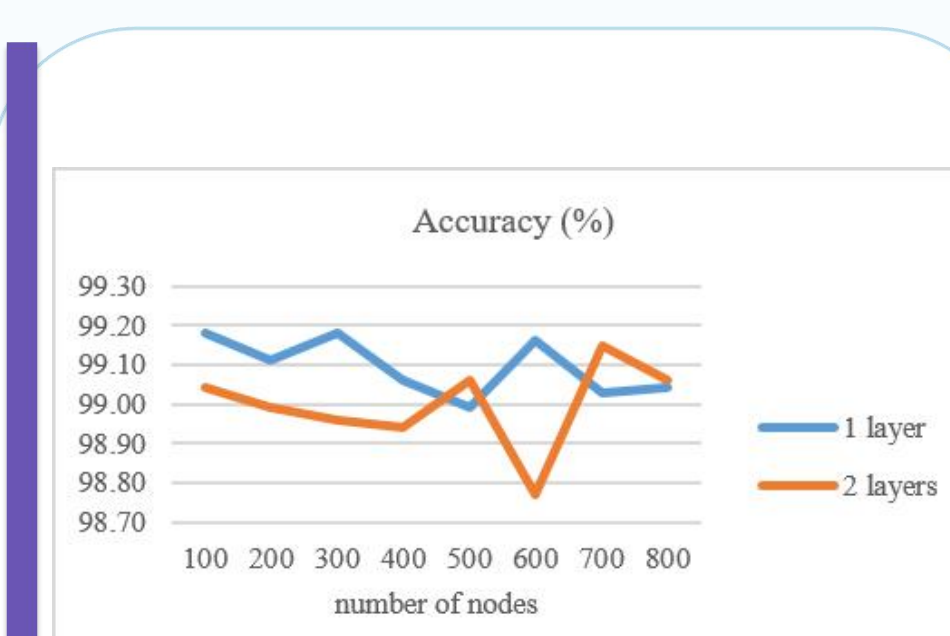
Фиг. 4. Матрица на обръкване – Random Forest



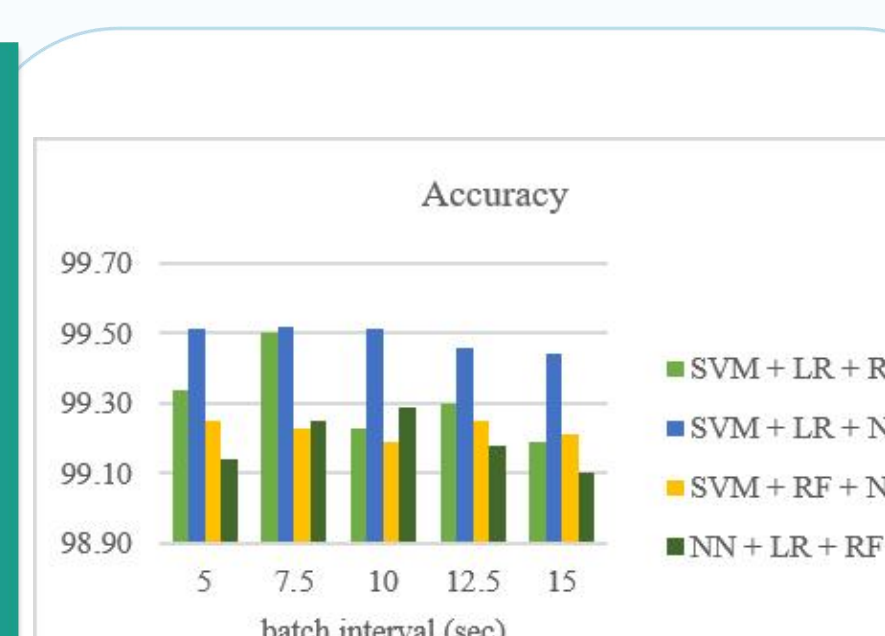
Фиг. 5. Крива прецизност–пълнота – Random Forest



Фиг. 6. Точност на оценените алгоритми



Фиг. 7. Влияние на броя възли в скрито ниво върху класификационна точност на спам класификатор



Фиг. 8. Класификационна точност на ансамблов модел

4 Основни изводи

- Ансамбловите модели осигуряват най-висока точност и по-стабилни резултати.
- Random Forest се отличава с надеждно прогнозиране на професионалната реализация.
- Анализът на анкетно поведение подобрява качеството и надеждността на данните.
- Получените резултати имат пряка приложимост в образователната аналитика.
- Създадените програмни решения могат да се надграждат в бъдещи проекти.

5 Научни публикации

1. Integrated analysis of student surveys with machine learning
2. Machine Learning Approaches for Predicting Students' Intentions to Pursue Higher Education
3. Automated Detection of Inattentive and Extreme Response Styles in Survey Data
4. Predicting student placement status using tree-based ensemble classifiers
5. Decision trees, random forests, and hybrid ensembles for ordinal survey data analysis
6. Application of the simplex method in tasks for determining an optimal production program
7. Teaching linear regression with Python and Excel: Methods, tools, and practical tasks
8. Random forest versus logistic regression for student placement prediction
9. Evaluating Neural Network Parameter Effects on Spam Classification Accuracy
10. Impact of Classifier Choice on Automatic Spam Detection: An Experimental Analysis
11. Разширяване на мрежи за достигане на пълна свързаност
12. Алгоритмично моделиране с дълбочинно търсене
13. Анализ на граф чрез последователно премахване на листни възли
14. Обзор на моделите за детекция на обекти в компютърното зрение
15. Моделиране на задачи чрез побитови представяния

6 Благодарности

Проект НП4/2025

Финансиран целево от държавния бюджет чрез Технически университет – Варна.

Благодарим на изследователския екип!

Заедно
към по-добро бъдеще!