Industrial Management Journal

Print ISSN: 2008-5885

Online ISSN: 2423-5369

Unleashing the Power of Ensemble Learning: Predicting National Ranks in Iran's University Entrance Examination

Ahmad Jafarnejad Chaghoshi*

*Corresponding Author, Prof., Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. E-mail: jafarnjd@ut.ac.ir

Arman Rezasoltani

PhD., Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. E-mail: armanrezasoltani@ut.ac.ir

Amir Mohammad Khani

PhD., Department of Industrial Management, Faculty of Industrial Management and Technology, College of Management, University of Tehran, Tehran, Iran. E-mail: amir.mo.khani@ut.ac.ir

Abstract

Objective

This study seeks to explore and compare ensemble learning models for more accurate predictions of students' national ranks in Iran's nationwide university entrance examination, commonly known as the Konkur. The primary aim is to identify optimal models capable of predicting students' ranks with the highest precision by analyzing data from preparatory and simulated exams conducted before the Konkur. These identified models can then empower academic counselors to offer data-driven recommendations, assisting students in making informed decisions about their educational paths and academic planning.

Methods

Initially, Octoparse software was utilized to collect data related to the preparatory tests conducted by the Kanoon Farhangi Amoozesh (Cultural Center of Education, also known as Ghalamchi) to facilitate a precise analysis of students' performance. The dataset contained key information such as the average test scores, students' national and regional ranks in the exams, as well as the academic disciplines and universities where the students were admitted. Additionally, four advanced ensemble learning models—XGBoost, LightGBM, CatBoost, and Random Forest—were selected for comparison and evaluation based on their performance in predicting students' national ranks. To assess the accuracy and efficiency of these models, various metrics were used, including Mean Squared Error (MSE), Root Mean Squared Error (RMSE), coefficient of determination (R^2), and training and prediction times. Afterward, The collected data were split into training and testing sets to ensure optimal model training and evaluation. To achieve the best possible performance, the grid search

method was applied, enabling precise tuning of the model parameters to obtain optimal results.

Results

The study findings underscored the exceptional performance of the XGBoost and LightGBM models in predicting students' national ranks. XGBoost emerged as the most accurate model, delivering predictions that are closely aligned with the actual values and exhibiting the lowest error rates, positioning it as the top performer in this research. Similarly, LightGBM, with results closely mirroring those of XGBoost, was also recognized as a strong contender for predicting national ranks. Its blend of speed and precision made it a highly effective tool in this context. The Random Forest model also demonstrated superior accuracy compared to CatBoost, albeit requiring more time for training and prediction. In contrast, CatBoost was identified as the weakest option, displaying lower accuracy and higher error rates than the other models, despite its faster prediction times. This suggests that CatBoost may require further refinement to compete with the more successful models.

Conclusion

The research findings highlight the effectiveness of XGBoost and LightGBM models in ensemble learning, showcasing their exceptional performance in predicting students' ranks in nationwide examinations. With their high accuracy and optimal performance, these models can serve as valuable tools within educational systems, contributing to the improvement of learning processes. They have the potential to identify students who may encounter challenges in their educational journeys and aid in the development of more effective learning programs. Furthermore, the outcomes of this study can assist school leaders and educational policymakers in making well-informed decisions to promote educational equity and ensure equal learning opportunities for all students. Looking ahead, leveraging more advanced deep learning models and integrating additional data, such as social and economic factors and school types, could lead to enhanced prediction accuracy. Additionally, the integration of diverse machine learning models to create hybrid systems may enhance precision and minimize prediction errors. Overall, this research marks a significant milestone in the advancement of predictive and decision-making systems within the education sector, laying the groundwork for substantial progress in this field.

Keywords: Ensemble learning, National rank prediction, National university entrance examination, Machine learning models.

Citation: Jafarnejad, Ahmad; Rezasoltani, Arman & Khani, Amir Mohammad (2024). Unleashing the Power of Ensemble Learning: Predicting National Ranks in Iran's University Entrance Examination. *Industrial Management Journal*, 16(3), 457-481. (*in Persian*)

Industrial Management Journal, 2024, Vol. 16, No 3, pp. 457-481
Published by University of Tehran, Faculty of Management
https://doi.org/10.22059/IMJ.2024.381521.1008178
Article Type: Research Paper
© Authors
Received: June 26, 2024
Received in revised form: August 27, 2024
Accepted: September 15, 2024
Published online: October 01, 2024

مدیریت صنعتی
شاپا چاپی: ۵۸۸۵-۲۰۰۸
شاپا الکترونیکی: ۵۳۶۹-۲۴۲۳

مقایسه مدل‌های یادگیری جمعی برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری

احمد جعفرنژاد چقوشی *

* نویسندهٔ مسئول، استاد، گروه مدیریت صنعتی، دانشکده مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: jafarnjd@ut.ac.ir

آرمان رضاسلطانی

دانشجوی دکتری، گروه مدیریت صنعتی، دانشکده مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: armanrezasoltani@ut.ac.ir

امیرمحمد خانی

دانشجوی دکتری، گروه مدیریت صنعتی، دانشکده مدیریت صنعتی و فناوری، دانشکدگان مدیریت، دانشگاه تهران، تهران، ایران. رایانامه: amir.mo.khani@ut.ac.ir

چکیده

هدف: این پژوهش به بررسی و مقایسه مدل‌های یادگیری جمعی می‌پردازد تا بتواند به پیش‌بینی دقیق‌تری از رتبهٔ کشوری دانش‌آموزان در کنکور سراسری دست یابد. هدف اصلی این پژوهش شناسایی مدل‌های بهینه‌ای است که با تحلیل داده‌های آزمون‌های آمادگی که پیش از کنکور برگزار می‌شود، بتوانند رتبهٔ دانش‌آموزان را با بیشترین دقت پیش‌بینی کنند. مدل‌های شناسایی‌شده می‌توانند به مشاوران تحصیلی کمک کنند تا با ارائهٔ توصیه‌های دقیق‌تر و مبتنی بر داده، به دانش‌آموزان در انتخاب مسیر تحصیلی و برنامه‌ریزی آموزشی کمک کنند.

روش: در این پژوهش، به‌منظور تحلیل دقیق عملکرد دانش‌آموزان، از نرم‌افزار Octoparse برای جمع‌آوری داده‌های مرتبط با آزمون قلمچی کانون فرهنگی آموزش استفاده شد. مجموعه داده‌ها اطلاعات مهمی بودند، نظیر میانگین نمره آزمون، رتبه‌های کشوری و منطقه‌ای، رشته‌های تحصیلی و دانشگاه‌های قبولی دانش‌آموزان. همچنین در این مطالعه، چهار مدل پیشرفته یادگیری جمعی شامل XGBoost، LightGBM، CatBoost و Random Forest به‌منظور مقایسه و ارزیابی عملکرد در پیش‌بینی رتبهٔ کشوری دانش‌آموزان انتخاب شد. به‌منظور سنجش دقت و کارایی این مدل‌ها، از معیارهای متنوعی مانند میانگین مربعات خطا، جذر میانگین مربعات خطا، ضریب تعیین و همچنین زمان‌های آموزش و پیش‌بینی استفاده شد. داده‌های جمع‌آوری‌شده، به دو بخش آموزشی و آزمایشی تقسیم شدند تا مدل‌ها بتوانند به بهترین شکل آموزش ببینند و ارزیابی شوند. به‌منظور دستیابی به عملکرد بهینه، از روش جست‌وجوی شبکه‌ای بهره گرفته شد که به ما اجازه می‌دهد پارامترهای مدل‌ها را به‌طور دقیق تنظیم کنیم و بهترین نتایج ممکن را به‌دست آوریم.

یافته‌ها: نتایج این مطالعه نشان‌دهندهٔ عملکرد برجسته مدل‌های XGBoost و LightGBM در پیش‌بینی رتبهٔ کشوری دانش‌آموزان بود. مدل XGBoost به‌عنوان دقیق‌ترین گزینه شناخته شد؛ زیرا پیش‌بینی‌هایی ارائه داد که به مقادیر واقعی بسیار

نزدیک بود و کمترین میزان خطا را داشت. این دقت بالا باعث شد که XGBoost به‌عنوان مدل برتر در این پژوهش شناخته شود. مدل LightGBM نیز با نتایج بسیار مشابه XGBoost، به‌عنوان یکی دیگر از گزینه‌های برجسته، برای پیش‌بینی رتبهٔ کشوری انتخاب شد. این مدل به خاطر سرعت و دقت زیاد، مورد توجه قرار گرفت و به‌عنوان ابزاری مؤثر در این زمینه شناخته شد. مدل Random Forest نیز با دقتی بهتر نسبت به CatBoost عمل کرد؛ هرچند زمان بیشتری برای آموزش و پیش‌بینی نیاز داشت. در مقابل، مدل CatBoost به‌عنوان ضعیف‌ترین گزینه شناخته شد؛ زیرا در مقایسه با سایر مدل‌ها دقت کمتری داشت و مقادیر خطای بیشتری ارائه می‌کرد؛ هرچند سرعت پیش‌بینی آن بیشتر بود. به نظر می‌رسد که این مدل به بهبود نیاز دارد تا بتواند با دیگر مدل‌های موفق رقابت کند.

نتیجه‌گیری: یافته‌ها نشان می‌دهد که مدل‌های XGBoost و LightGBM به‌عنوان دو ابزار مؤثر در یادگیری جمعی، عملکرد بسیار خوبی برای پیش‌بینی رتبهٔ دانش‌آموزان در آزمون‌های سراسری دارند. این مدل‌ها با دقت بالا و عملکرد بهینه، می‌توانند به‌عنوان راهنماهایی ارزشمند در سیستم‌های آموزشی عمل کنند و به بهبود فرایندهای یادگیری کمک کنند. این مدل‌ها قادرند دانش‌آموزانی را که ممکن است در مسیر تحصیلی خود با چالش مواجه شوند، شناسایی کنند و به طراحی برنامه‌های یادگیری مؤثرتر یاری دهند. علاوه‌براین، نتایج این پژوهش می‌تواند به رهبران مدارس و سیاست‌گذاران برنامه‌های آموزشی کمک کند تا تصمیم‌های هوشمندانه‌تری در جهتِ ارتقای عدالت آموزشی اتخاذ کنند؛ به‌طوری که فرصت‌های یادگیری برابر، برای همهٔ دانش‌آموزان فراهم شود. در آینده، استفاده از مدل‌های پیشرفته‌تر یادگیری عمیق و اضافه‌کردن داده‌های مرتبط، مانند عوامل اجتماعی، اقتصادی و نوع مدارس، می‌تواند به بهبود دقت پیش‌بینی‌ها منجر شود. همچنین، ترکیب مدل‌های مختلف یادگیری ماشین برای ایجاد مدل‌های ترکیبی می‌تواند به افزایش دقت و کاهش خطاهای پیش‌بینی کمک کند. به‌طور کلی، این پژوهش می‌تواند نقطهٔ عطفی برای توسعهٔ سیستم‌های پیش‌بینی و تصمیم‌گیری در حوزهٔ آموزش در کشور شمرده شود و زمینه‌ساز بهبودهای چشمگیر در این زمینه باشد.

کلیدواژه‌ها: یادگیری جمعی، پیش‌بینی رتبهٔ کشوری، کنکور سراسری، مدل‌های یادگیری ماشین.

استناد: جعفرنژاد چقوشی، احمد؛ رضاسلطانی، آرمان و خانی، امیرمحمد (۱۴۰۳). مقایسه مدل‌های یادگیری جمعی برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری. *مدیریت صنعتی*، ۱۶(۳)، ۴۵۷- ۴۸۱.

مدیریت صنعتی، ۱۴۰۳، دوره ۱۶، شماره ۳، صص. ۴۵۷- ۴۸۱
ناشر: دانشکدهٔ مدیریت دانشگاه تهران
نوع مقاله: علمی پژوهشی
© نویسندگان
تاریخ دریافت: ۱۴۰۳/۰۴/۰۶
تاریخ ویرایش: ۱۴۰۳/۰۶/۰۶
تاریخ پذیرش: ۱۴۰۳/۰۶/۲۵
تاریخ انتشار: ۱۴۰۳/۰۷/۱۰
doi: https://doi.org/10.22059/IMJ.2024.381521.1008178

مقدمه

برای پژوهشگران و سیاست‌گذاران آموزشی، پیش‌بینی رتبه‌بندی دانش‌آموزان کشوری در کنکور، همواره یکی از موضوعات کلیدی بوده است. در واقع، با توجه به اهمیت آن در تعیین پیشرفت تحصیلی و مسیرهای شغلی دانش‌آموزان، مدل‌های پیش‌بینی قابل‌اعتماد، می‌توانند به‌طور چشمگیری بر جنبه‌هایی مانند افزایش نتایج یادگیری برای برابری بهتر در آموزش تأثیر بگذارند. این موضوع نکتهٔ جالبی را دربارهٔ اینکه چگونه رویکردهای یادگیری گروهی که از تعدادی تکنیک‌های یادگیری ماشینی متنوع تشکیل شده‌اند، به‌عنوان ابزاری مؤثر برای پیش‌بینی عملکرد تحصیلی توصیه شده‌اند، مطرح می‌کند. آزمون کنکور سراسری در نظام آموزشی ایران، به‌عنوان یکی از مهم‌ترین و سرنوشت‌سازترین آزمون‌ها در مسیر تحصیلی دانش‌آموزان شناخته می‌شود. این آزمون که به‌طور سالانه برگزار می‌شود، به‌عنوان معیاری برای ورود به دانشگاه‌ها و مؤسسه‌های آموزش عالی کشور به‌شمار می‌رود و در انتخاب رشته و دانشگاه دانش‌آموزان نقشی تعیین‌کننده دارد. از این رو، پیش‌بینی رتبهٔ کشوری دانش‌آموزان در این آزمون، برای والدین، دانش‌آموزان، مشاوران آموزشی و حتی سیاست‌گذاران آموزشی از اهمیت بالایی برخوردار است.

نظام آموزشی ایران بسیار رقابتی است و آزمون عالی سراسری عموماً در مسیر تحصیلی دانش‌آموزان مرحله‌ای حساس و تعیین‌کننده است. در نتیجه آن‌ها را همراه با خانواده خود در معرض فشارهای روانی و استرس شدید قرار می‌دهند. برعکس، این آزمون تنها عامل تعیین‌کننده برای ورود به دانشگاه است که انتقادهای مردمی را به‌دنبال داشته است. در واقع، تمرکز صرف بر نمره‌های آزمون کتبی ممکن است مهارت‌ها یا توانایی‌های فردی را نادیده بگیرد که به استفاده نابهینه از پتانسیل انسانی کشور منجر می‌شود. با بررسی چنین مجموعه‌ای از پارامترها، ایجاد مدل‌های پیش‌بینی دقیق برای قضاوت و پیش‌بینی عملکرد دانش‌آموزان در کنکور سراسری ضروری است. یک روش اخیراً مؤثر در این حوزه، مدل‌های یادگیری جمعی است. این نوع مدل‌ها می‌توانند دقت پیش‌بینی را از طریق ترکیبی از الگوریتم‌های مختلف یادگیری ماشین بهبود بخشند. در واقعیت، دانشگاه‌های گروه‌بندی شده، به‌عنوان یک استراتژی جدید در تجزیه‌وتحلیل داده‌ها و درک ماشین، نادرستی‌ها را حذف می‌کند و عملکرد کلی ماشین پیش‌بینی را با پیوستن به خروجی‌های مختلف مدل‌های متمایز افزایش می‌دهد.

مطالعات متعددی در زمینهٔ پیش‌بینی عملکرد تحصیلی دانش‌آموزان، به‌ویژه از طریق آزمون‌های مهم و قطعی مانند آزمون‌های کنکور سراسری، نشان می‌دهد که نمره‌های دانش‌آموزان را می‌توان به‌طور دقیق با کمک مدل‌های یادگیری جمعی پیش‌بینی کرد. برای مثال، تحقیقات ادجو و کلونی[1] (۲۰۱۸) نشان داده است که مدل‌های یادگیری جمعی ناهمگن، به‌طور چشمگیری دقت پیش‌بینی‌ها را بهبود می‌بخشد. ترکیبی از الگوریتم‌های مختلف مانند SVM، Random Forest و Adaboost مؤثر بود؛ زیرا آن‌ها فهمیدند که در مقایسه با مدل‌های مبتنی بر یک الگوریتم انحصاری، مدل‌های جمعی پیش‌بینی‌های قابل اعتمادتری ارائه می‌دهند. این موضوع را مطالعات اخیر تأیید کرده‌اند که الگوهای جمعی حالت قابل اعتمادتر از سایر تکنیک‌های پیش‌بینی در پیش‌بینی موفقیت تحصیلی دانش‌آموزان هستند

1. Adejo & Connolly

(بات و همکاران[1]، ۲۰۲۳). آن‌ها همچنین برای استفاده هنگام برخورد با مجموعه داده‌های عظیم و پیچیده توصیه می‌شوند. چنین الگوهایی علاوه‌براینکه ابزار بسیار مهمی برای سیاست‌گذاران یا برنامه‌ریزان آموزشی هستند، دقت پیش‌بینی را از طریق تکنیک‌هایی مانند بسته‌بندی[2] یا تقویت[3] بهبود می‌بخشند. علاوه‌براین، مطالعات هان، تانگ، چن، لیو و لیو[4] (۲۰۱۷) نشان داد که الگوریتم رتبه‌بندی Adaboost نسبت به سایر الگوریتم‌ها مانند درخت تصمیم، SVM و شبکه عصبی برتری دارد. بنابراین، این یافته‌ها بر چگونگی تأثیرگذاری تکنیک‌های یادگیری گروهی در افزایش دقت پیش‌بینی یا کاهش خطاهای سیستماتیک مرتبط با ارزشیابی دانش‌آموزان تأکید دارند.

در این پژوهش با هدف تحلیل تطبیقی مدل‌های مختلف یادگیری جمعی (XGBoost، LightGBM، CatBoost، Random Forest) در پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری، سعی شد تا دقت، سرعت آموزش و کارایی این مدل‌ها در شرایط دنیای واقعی مورد بررسی قرار گیرد. با استفاده از داده‌های آزمون‌های قلمچی، پژوهش تلاش کرد تا مدل‌هایی را شناسایی کند که بتوانند با دقت بیشتری رتبهٔ دانش‌آموزان را پیش‌بینی کنند. همچنین، بررسی شد که کدام‌یک از این مدل‌ها علاوه‌بر دقت بالا، سرعت مناسب‌تری برای آموزش و پیش‌بینی دارند و برای پیاده‌سازی در مقیاس‌های بزرگ مناسب‌ترند. هدف این پژوهش، تحلیل تطبیقی مدل‌های یادگیری جمعی در پیش‌بینی رتبهٔ کنکور و شناسایی مدلی است که با در نظر گرفتن معیارهای مختلفی همچون دقت، سرعت آموزش و زمان پیش‌بینی، بهترین عملکرد را داشته باشد. این مقایسه نه‌تنها می‌تواند به بهبود فرایندهای آموزشی و ارائه راه‌کارهای عملی در سیستم‌های مشاوره تحصیلی کمک کند، بلکه می‌تواند به سیاست‌گذاران آموزشی در اتخاذ تصمیم‌های مؤثرتر در جهت بهبود عدالت آموزشی و کاهش نابرابری‌ها در فرصت‌های یادگیری یاری رساند.

برای دستیابی به این اهداف، پژوهش حاضر به‌دنبال پاسخ سؤال‌های زیر است:

۱. کدام‌یک از مدل‌های یادگیری جمعی (XGBoost، LightGBM، CatBoost، Random Forest) برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری دقت بیشتری دارند؟
۲. چگونه می‌توان با استفاده از داده‌های آزمون‌های آمادگی قبل از کنکور، رتبهٔ کشوری دانش‌آموزان را با دقت پیش‌بینی کرد؟
۳. کدام‌یک از مدل‌های یادگیری جمعی، سرعت آموزش و پیش‌بینی بیشتری دارد و برای پیاده‌سازی در مقیاس‌های بزرگ مناسب‌تر است؟
۴. چگونه می‌توان از نتایج مدل‌های پیش‌بینی برای ارائهٔ راه‌کارهای عملی در سیستم‌های مشاوره تحصیلی استفاده کرد؟

این پژوهش با تمرکز بر داده‌های بی‌درنگ دانش‌آموزان و استفاده از مدل‌های پیشرفته یادگیری جمعی، به‌دنبال پاسخ به این سؤال‌هاست و نتایج آن می‌تواند به‌عنوان زیرساختی برای توسعهٔ سیستم‌های پیش‌بینی و تصمیم‌گیری در بخش آموزش و پرورش کشور استفاده شود و به ارتقای کیفیت آموزشی و افزایش عدالت آموزشی در سطح ملی بینجامد.

1. Butt et al.
2. Bagging
3. Boosting
4. Han, Tong, Chen, Liu & Liu

پیشینهٔ پژوهش

برای مقایسهٔ مدل‌های یادگیری جمعی در پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری با استفاده از نتایج آزمون قلمچی، تعدادی از مقاله‌های مرتبط با موضوع بررسی شد که به مقایسه و تحلیل کارایی مدل‌های یادگیری جمعی در پیش‌بینی عملکرد تحصیلی دانش‌آموزان پرداخته‌اند. این مطالعات نشان می‌دهد که مدل‌های یادگیری جمعی در پیش‌بینی رتبهٔ دانش‌آموزان در آزمون‌ها، دقت و کارایی بیشتری دارند و می‌توانند به‌طور مؤثری در سیستم‌های آموزشی و ارزیابی عملکرد دانش‌آموزان به کار گرفته شوند.

جدول ۱. پیشینه پژوهش

نویسندگان	عنوان مقاله	نتایج	مدل مورد استفاده
بات و همکاران (۲۰۲۳)	مدل‌های جمعی برای پیش‌بینی عملکرد دانشجویان در آموزش عالی	در این مطالعه از مدل‌های یادگیری جمعی مانند bagging، boosting، stacking، و voting برای پیش‌بینی عملکرد دانشجویان استفاده شده است. نتایج نشان می‌دهد که روش‌های boosting و stacking در مجموعه‌های داده کوچک و بزرگ عملکرد بهتری دارند.	Bagging, Boosting, Stacking, Voting
ساکری و صالح[1] (۲۰۲۰)	مدل هیبریدی یادگیری جمعی برای ارزیابی عملکرد دانشجویان	این مطالعه یک مدل هیبریدی جمعی (RHEM) را برای پیش‌بینی عملکرد دانشجویان معرفی می‌کند. این مدل با ترکیب الگوریتم‌های Naïve Bayes، Multilayer Perceptron، k-Nearest Neighbours و Decision Table با روش‌های جمعی مانند Bagging و Random Subspace به دقت ۹۱/۷درصد دست یافته است.	Naïve Bayes, Multilayer Perceptron, k-Nearest Neighbours, Decision Table, Bagging, Random Subspace
یان و لیو[2] (۲۰۲۰)	مدل پیش‌بینی عملکرد دانشجویان با استفاده از یادگیری ماشین	این مطالعه از مدل‌های یادگیری جمعی مانند Adaboost، Random Forest و SVM برای پیش‌بینی عملکرد دانش‌آموزان استفاده شده است. نتایج نشان داده است که این مدل‌ها نسبت به مدل‌های تکی دقت بیشتری دارند.	Adaboost, Random Forest, SVM
انجدات، موبید، نصیف و شامی[3] (۲۰۲۰)	انتخاب مدل جمعی سیستماتیک برای داده‌کاوی آموزشی	این مطالعه به بررسی و انتخاب مدل‌های یادگیری جمعی برای پیش‌بینی عملکرد دانش‌آموزان پرداخته و نشان داده است که مدل‌های یادگیری جمعی دقیق‌تر و کارآمدتر از مدل‌های تک الگوریتمی هستند.	Random Forest, Gradient Boosting, Adaboost, Bagging
ادجو و کلونی (۲۰۱۸)	پیش‌بینی عملکرد تحصیلی دانشجویان با استفاده از رویکرد جمعی ناهمگن چند مدل	این مطالعه نشان می‌دهد که استفاده از منابع داده‌های چندگانه و مدل‌های یادگیری جمعی، دقت و کارایی پیش‌بینی‌ها را افزایش می‌دهد.	Random Forest, Gradient Boosting, Stacking
هان و همکاران (۲۰۱۷)	کاربرد الگوریتم‌های یادگیری جمعی در پیش‌بینی عملکرد دانش‌آموزان	این مطالعه مقایسه‌ای بین الگوریتم‌های مختلف یادگیری جمعی مانند Adaboost و سایر الگوریتم‌ها مانند تصمیم‌گیری درختی، شبکه عصبی و SVM انجام داده و نتایج نشان می‌دهد که Adaboost بهترین عملکرد را داشته است.	Adaboost, Decision Tree, Neural Network, SVM

1. Sakri and Saleh
2. Yan and Liu
3. Injadat, Moubayed, Nassif & Shami

نویسندگان	عنوان مقاله	نتایج	مدل مورد استفاده
وانگ و شی[1] (۲۰۱۶)	پیش‌بینی نمرات آزمون‌های ورودی دانشگاه با استفاده از یادگیری ماشین	در این مطالعه از الگوریتم Adaboost برای پیش‌بینی نمرات آزمون‌های ورودی دانشگاه استفاده شده است که نشان‌دهندهٔ برتری این روش نسبت به مدل‌های پیش‌بینی سنتی است.	Adaboost
سالاری، رادفر و فقیهی (۱۴۰۳)	پیش‌بینی عملکرد دانشجویان با استفاده از الگوریتم‌های یادگیری ماشین و داده‌کاوی آموزشی (مطالعه موردی دانشگاه شاهد)	الگوریتم درخت تصمیم از سایر الگوریتم‌ها در پیش‌بینی عملکرد دانش‌آموزان با دقت ۸۴/۷۱درصد عملکرد بهتری داشت و فارغ‌التحصیلی را برای ۷۷/۸۸درصد از دانش‌آموزان ممتاز پیش‌بینی کرد.	Decision Tree
زنگوئی و فاطمی (۱۴۰۰)	پیش‌بینی دانش‌آموزان در معرض خطر افت تحصیلی با استفاده از تجزیه‌وتحلیل یادگیری در سیستم مدیریت یادگیری	شبکه LSTM در پیش‌بینی اولیه دانش‌آموزان در معرض خطر، به‌ترتیب با دقت ۹۴درصد و دقت ۸۸درصد، از الگوریتم ماشین بردار پشتیبانی بهتر عمل می‌کند. الگوریتم پیشنهادی به‌ترتیب با ۰/۹۳۶ و ۰/۸۸۲ عملکرد معلم و دانش‌آموز را به‌طور قابل‌توجهی بهبود می‌بخشد.	LSTM, SVM
اسدی و جوادزاده (۱۴۰۰)	پیش‌بینی نمرات دانش‌آموزان با استفاده از روش یادگیری ماشین (مقایسه ۳ الگوریتم درخت تصمیم، جنگل تصادفی و رگرسی	با استفاده از یک دیتاست که شامل نمرات و مقادیر مربوط به‌سایر ویژگی‌های دانش‌آموزان بوده، به بررسی میزان دقت و خطای هریک از این الگوریتم‌ها برای به‌دست آوردن نمره نهایی و یا احتمال قبولی و مردود شدن دانش‌آموزان در امتحان پایانی با توجه به اطلاعات قبلی در درس مربوطه پرداخته شد.	Decision Tree, Random Forest, Regression
قدوسی، میرسعیدی و کوشا (۱۳۹۹)	پیش‌بینی و تحلیل عملکرد دانشجویان به کمک تکنیک‌های داده‌کاوی به‌منظور بهبود عملکرد تحصیلی	مدل پیشنهادی از الگوریتم‌های داده کاوی برای پیش‌بینی عملکرد دانش‌آموز استفاده می‌کند که ابزار پشتیبانی تصمیم‌گیری مناسب در سیستم‌های آموزشی را فراهم می‌کند. فرایند انتخاب واحد، با استفاده از داده‌های موجود و علم داده‌کاوی، برای افزایش کارایی و تصمیم‌گیری آگاهانه، دوباره طراحی شد.	Decision Tree, Naïve Bayes, Neural Network
بابایی و شویدی نوش آبادی (۱۳۹۹)	مروری بر روش‌های پیش‌بینی عملکرد دانشجویان با استفاده از الگوریتم‌های یادگیری ماشین	این مقاله ۳ الگوریتم یادگیری ماشینی محبوب را مورد بحث قرار می‌دهد، جنگل تصادفی، درخت تصمیم و رگرسیون خطی. دقت و خطای آن‌ها را در کسب نمره‌های امتحان نهایی براساس ویژگی‌های دانش‌آموز بررسی می‌کند. این مقاله همچنین تکنیک‌های مورد استفاده برای بهبود نتایج را مورد بحث قرار می‌دهد و جداول را برای مقایسه ۳ الگوریتم برتر ارائه می‌دهد.	Random Forest, Decision Tree, Linear Regression

1. Wang and Shi

مطالعات مرور شده نشان می‌دهند که مدل‌های یادگیری جمعی، به‌ویژه مدل‌های Boosting، Stacking و Random Forest، در پیش‌بینی عملکرد تحصیلی دانش‌آموزان و دانشجویان با دقت زیادی عمل کرده‌اند. به‌عنوان مثال، بات و همکاران (۲۰۲۳) به این نتیجه رسیدند که روش‌های Boosting و Stacking به‌ویژه در مجموعه‌های داده کوچک و بزرگ عملکرد بهتری دارند. این یافته‌ها با نتایج یان و لیو (۲۰۲۰) نیز هم‌خوانی دارد که نشان دادند مدل‌های جمعی مانند Adaboost و Random Forest در مقایسه با مدل‌های تکی دقت بیشتری در پیش‌بینی عملکرد تحصیلی دارند. در مقابل، برخی مطالعات مانند سالاری و همکاران (۱۴۰۳) و زنگوئی و فاطمی (۱۴۰۰)، به‌طور خاص به استفاده از مدل‌های تکی مانند Decision Tree و LSTM پرداختند. هرچند این مدل‌ها نیز دقت قابل قبولی را نشان دادند، اما بررسی‌ها نشان می‌دهد که عملکرد آن‌ها به‌طور چشمگیری از مدل‌های جمعی کمتر است. این تفاوت‌ها می‌تواند ناشی از عدم ترکیب داده‌ها و استفاده محدود از مدل‌های جمعی باشد.

با وجود اینکه پژوهش‌های متعددی به بررسی مدل‌های یادگیری جمعی پرداخته‌اند، چندین شکاف مهم در این تحقیقات مشاهده می‌شود:

تمرکز ناکافی بر پیش‌بینی دقیق رتبهٔ کشوری: بیشتر مطالعات پیشین، عملکرد تحصیلی را به‌صورت کلی مورد بررسی قرار داده‌اند و پیش‌بینی دقیق رتبهٔ کشوری دانش‌آموزان در کنکور سراسری به‌طور خاص مورد توجه قرار نگرفته است. پژوهش حاضر به‌طور خاص به این مسئله می‌پردازد و از مدل‌های یادگیری جمعی برای پیش‌بینی دقیق رتبهٔ کشوری استفاده می‌کند.

مقایسهٔ جامع مدل‌های یادگیری جمعی: در تحقیقات پیشین مدل‌های مختلف یادگیری جمعی بررسی شده‌اند؛ اما مقایسهٔ جامع و دقیق بین آن‌ها در زمینهٔ پیش‌بینی رتبهٔ کشوری کنکور سراسری کمتر انجام شده است. این پژوهش با مقایسهٔ مدل‌های مختلف مانند Boosting، Bagging و Stacking در داده‌های خاص کنکور، قصد دارد شکاف موجود را پر کند.

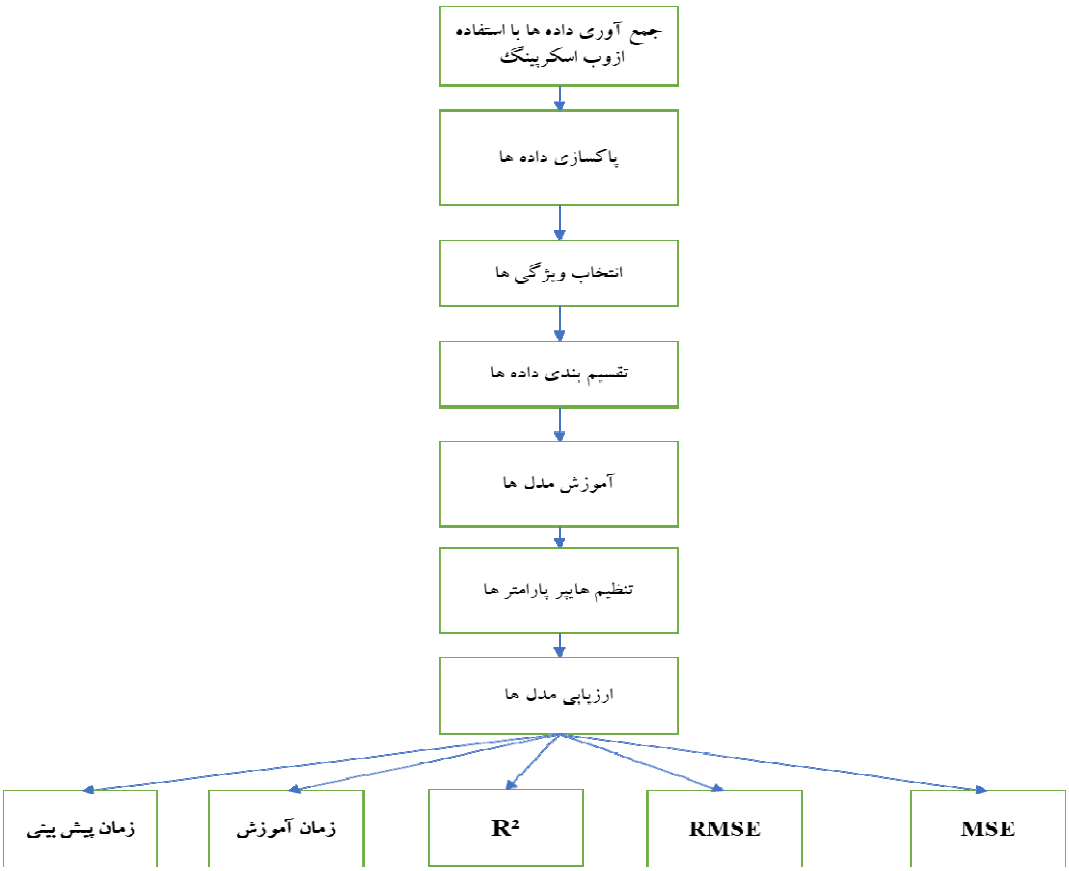
عدم تعمیم‌پذیری نتایج: مطالعاتی که بررسی شده‌اند، اغلب روی داده‌های خاص و در محیط‌های آموزشی خاصی انجام شده‌اند. این موضوع ممکن است قابلیت تعمیم نتایج به سایر محیط‌ها و گروه‌های آموزشی را کاهش دهد. بنابراین، نیاز به تحقیقاتی است که مدل‌های یادگیری جمعی را بر روی داده‌های متنوع‌تر و در محیط‌های مختلف آزمون کند.

تطبیق‌پذیری با سیستم آموزشی ایران: اغلب مطالعات گذشته در زمینه‌های آموزشی بین‌المللی انجام شده‌اند و تطبیق‌پذیری مدل‌های پیشنهادی با سیستم آموزشی ایران کمتر بررسی شده است. این پژوهش با استفاده از داده‌های آزمون های آزمایشی کنکور و رتبهٔ کشوری، نتایج عملی و کاربردی را برای سیستم آموزشی ایران ارائه می‌دهد.

با توجه به شکاف‌های شناسایی شده در تحقیقات گذشته، پژوهش حاضر با هدف مقایسه مدل‌های یادگیری جمعی در پیش‌بینی دقیق رتبهٔ کشوری دانش‌آموزان در کنکور سراسری انجام می‌شود. این پژوهش با استفاده از داده‌های آزمون‌های آزمایشی کنکور و تحلیل دقیق‌تر داده‌ها، می‌تواند بهبودهای مهمی در دقت پیش‌بینی ارائه دهد و به‌عنوان ابزاری مفید در سیستم‌های آموزشی به‌کار گرفته شود. نتایج این پژوهش می‌تواند راهنمایی‌های کاربردی برای بهبود عملکرد دانش‌آموزان و بهینه‌سازی فرایندهای آموزشی فراهم کند.

روش‌شناسی پژوهش

این پژوهش به‌منظور مقایسه مدل‌های یادگیری جمعی برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری بر اساس داده‌های آزمون‌های آمادگی پیش از کنکور انجام شده است. داده‌های مورد استفاده در این پژوهش با استفاده از وب اسکرپینگ از وبسایت کانون فرهنگی آموزش[1] استخراج شده‌اند. ابزار مورد استفاده برای استخراج داده‌ها نرم‌افزار Octoparse است که یکی از ابزارهای قدرتمند در زمینه وب اسکرپینگ است.



شکل ۱. فرایند اجرای مدل‌های یادگیری جمعی برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری

جمع‌آوری داده‌ها

منبع داده‌ها

داده‌های مورد استفاده در این پژوهش، از وبسایت رسمی کانون فرهنگی آموزش (قلمچی) استخراج شده است. این وبسایت اطلاعات مفصلی دربارهٔ عملکرد دانش‌آموزان در آزمون‌های قلمچی، شامل میانگین تراز آزمون، رتبه‌های کشوری و منطقه‌ای، رشتهٔ تحصیلی، دانشگاه محل قبولی و سایر جزئیات مربوط به قبولی‌ها ارائه می‌دهد. داده‌های استخراج‌شده مربوط به بازهٔ زمانی سال‌های ۱۳۹۰ تا ۱۴۰۰ است.

1. https://www.kanoon.ir

فرایند وب اسکریپینگ

برای جمع‌آوری داده‌ها از وبسایت کانون فرهنگی آموزش، از نرم‌افزار Octoparse استفاده شده است. Octoparse ابزار قدرتمند وب اسکریپینگ است که به کاربران امکان می‌دهد تا داده‌های وبسایت‌ها را به‌صورت خودکار و بدون نیاز به برنامه‌نویسی استخراج کنند. مراحل اصلی وب اسکریپینگ به شرح زیر است:

تعیین آدرس صفحات وب: آدرس صفحات وب شامل اطلاعات دانش‌آموزان شناسایی و به نرم‌افزار Octoparse داده شد.

استخراج داده‌ها: با استفاده از ابزارهای کشیدن و رها کردن Octoparse، بخش‌های مختلف صفحات وب شامل نام شهر، رتبهٔ کشوری، میانگین تراز آزمون، سابقهٔ کانونی، تعداد آزمون‌ها، رشتهٔ تحصیلی، دانشگاه محل قبولی و سایر اطلاعات مرتبط انتخاب و استخراج شد.

ذخیره‌سازی: داده‌های استخراج‌شده به‌صورت خودکار در قالب یک فایل اکسل ذخیره شدند و سپس برای تجزیه‌وتحلیل‌های بیشتر در Python مورد استفاده قرار گرفتند.

ویژگی‌های دیتاست

تعداد رکوردها و ویژگی‌ها

دیتاست نهایی شامل ۷۳/۸۳۸ رکورد از داده‌های دانش‌آموزان است. این رکوردها مربوط به دانش‌آموزانی است که در آزمون‌های قلمچی شرکت کرده‌اند؛ سپس در کنکور سراسری قبول شده‌اند. داده‌ها مربوط به بازه زمانی سال‌های ۱۳۹۰ تا ۱۴۰۰ است. دیتاست ۱۱ ویژگی دارد که به شرح جدول ۲ است.

جدول ۲. ویژگی‌های دیتاست برای پیش‌بینی رتبهٔ کشوری در کنکور

ویژگی	توضیحات
شهر	نام شهر محل سکونت یا تحصیل دانش‌آموز
رتبهٔ کشوری	رتبهٔ دانش‌آموز در کنکور سراسری در سطح کشور
رتبه در منطقه	رتبهٔ دانش‌آموز در منطقهٔ تحصیلی خود (منطقه‌های ۱، ۲ یا ۳)
منطقه	منطقهٔ تحصیلی دانش‌آموز (بر اساس شهر به منطقه‌های ۱، ۲ یا ۳)
سابقهٔ کانونی	مدت زمانی که دانش‌آموز در آزمون‌های قلمچی شرکت کرده است
میانگین تراز کانون	میانگین نمره تراز دانش‌آموز در آزمون‌های قلمچی
تعداد آزمون	تعداد آزمون‌هایی که دانش‌آموز در کانون فرهنگی آموزش شرکت کرده است
رشته قبولی	رشتهٔ تحصیلی که دانش‌آموز در کنکور سراسری در آن قبول شده است
دانشگاه قبولی	دانشگاهی که دانش‌آموز در آن قبول شده است
رشته مدرسه	رشته‌ای که دانش‌آموز در مدرسه خوانده است
سال	سالی که دانش‌آموز در کنکور سراسری شرکت کرده و قبول شده است

در این پژوهش، ۶ ویژگی اصلی برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری انتخاب شد که همگی با آزمون‌های آمادگی پیش از کنکور مرتبط است. این ویژگی‌ها به‌طور خاص شامل میانگین تراز کانونی، سابقهٔ کانونی، تعداد آزمون‌ها، رشتهٔ مدرسه، منطقه و سال می‌شود. انتخاب این ویژگی‌ها به‌دلیل تأثیر مستقیم آن‌ها بر عملکرد دانش‌آموزان در آزمون‌های کنکور و قابلیت پیش‌بینی دقیق‌تر رتبهٔ کشوری انجام شده است. ویژگی منطقه یکی از ویژگی‌های کلیدی در این پژوهش است. این ویژگی بر اساس شهر محل سکونت یا تحصیل دانش‌آموزان تعیین شده است و دانش‌آموزان را بر اساس امکانات و شرایط تحصیلی به سه منطقهٔ ۱، ۲، و ۳ طبقه‌بندی می‌کند. استفاده از ویژگی منطقه به‌جای شهر، به دلایل زیر بوده است:

تأثیر امکانات آموزشی: پژوهش‌ها نشان داده‌اند که منطقه‌بندی بر اساس امکانات آموزشی و اجتماعی، بر عملکرد تحصیلی دانش‌آموزان تأثیر مستقیم دارد (دانیل[1]، ۲۰۲۱). این تأثیر در رتبه‌بندی کنکور نیز مشاهده می‌شود؛ به‌طوری که دانش‌آموزان مناطق مختلف با توجه به دسترسی به منابع آموزشی متفاوت، عملکردهای متفاوتی دارند.

قابلیت تعمیم‌دهی و دسته‌بندی: منطقه‌بندی شهرها به سه منطقه مختلف بر اساس شرایط و امکانات آموزشی، به تعمیم‌پذیری و دسته‌بندی بهتر داده‌ها کمک می‌کند. استفاده از شهر به تنهایی ممکن است اطلاعات بیشتری را وارد مدل کند؛ اما قابلیت تعمیم‌دهی و تفسیر نتایج را کاهش دهد (گیبسون و وب[2]، ۲۰۱۵).

ویژگی‌هایی که به نتایج نهایی کنکور و انتخاب رشته مرتبط هستند، مانند رشتهٔ قبولی یا دانشگاه قبولی، در این پژوهش حذف شدند. دلایل عدم استفاده از این ویژگی‌ها عبارت‌اند از:

جلوگیری از تأثیرگذاری داده‌های خروجی: استفاده از ویژگی‌هایی که مستقیماً به نتیجهٔ نهایی کنکور یا انتخاب رشته مرتبط هستند، می‌تواند باعث ورود سوگیری و داده‌های خروجی به مدل شود. این ویژگی‌ها به‌جای پیش‌بینی، صرفاً نتیجه‌ای هستند که پس از وقوع آزمون و انتخاب رشته حاصل می‌شوند و بنابراین نمی‌توانند به بهبود دقت پیش‌بینی رتبهٔ کشوری کمک کنند (ناوارو و همکاران[3]، ۲۰۲۱).

تمرکز بر عملکرد تحصیلی قبل از کنکور: هدف این پژوهش پیش‌بینی رتبهٔ کشوری بر اساس عملکرد تحصیلی و آزمون‌های آمادگی قبل از کنکور است؛ بنابراین ویژگی‌هایی مرتبط به نتایج پس از کنکور، مانند رشته یا دانشگاه قبولی، اطلاعاتی پسینی هستند که در زمان پیش‌بینی رتبهٔ کشوری قابل استفاده نیستند و می‌توانند باعث پیچیدگی و کاهش دقت مدل شوند (یاغچی[4]، ۲۰۲۲).

پیش‌پردازش داده‌ها

پاکسازی داده‌ها

بررسی مقادیر گمشده و نامعتبر: در این مرحله، داده‌ها به‌طور دقیق از نظر وجود مقادیر گم‌شده[5] و موارد نامعتبر[6]

1. Daniele
2. Gibson and Webb
3. Navarro et al.
4. Yağcı
5. Missing Values
6. Invalid Entries

بررسی شدند. برای این منظور، از روش‌هایی مانند تحلیل توزیع داده‌ها، محاسبهٔ درصد مقادیر گم‌شده و مقایسه با استانداردهای موجود استفاده شد.

مدیریت مقادیر گم‌شده: برای رکوردهایی که دارای مقادیر گم‌شده بودند، دو رویکرد به کار گرفته شد:

- حذف رکوردهای ناقص: رکوردهایی که مقادیر گم‌شدهٔ آن‌ها بیش از حد مجاز (بیشتر از ۵۰ درصد ویژگی‌ها) بودند، حذف شدند.
- جایگزینی مقادیر گم‌شده: برای رکوردهایی که تعداد محدودی مقدار گم‌شده داشتند، از روش‌های جایگزینی مانند میانگین‌گیری[1] برای ویژگی‌های عددی و مد رایج[2] برای ویژگی‌های دسته‌بندی‌شده استفاده شد.

بررسی و حذف موارد نامعتبر: مواردی که خارج از محدوده منطقی بودند (مانند نمره‌های غیرمنطقی یا داده‌های نامعتبر)، شناسایی، حذف یا اصلاح شدند. این فرایند شامل تحلیل‌های آماری و شناسایی نقاط پرت[3] بود تا اطمینان حاصل شود که داده‌های نهایی معتبرند.

تقسیم‌بندی داده‌ها

اهمیت تقسیم‌بندی داده‌ها: برای ارزیابی عملکرد مدل‌ها و جلوگیری از overfitting، داده‌ها به دو بخش آموزش[4] و آزمون[5] تقسیم شدند. این کار اجازه می‌دهد تا مدل‌ها روی یک بخش از داده‌ها آموزش ببینند و روی داده‌های جدید و نادیده‌گرفته‌شده ارزیابی شوند.

روش تقسیم‌بندی: داده‌ها با نسبت ۷۰:۳۰ به دو بخش آموزش و آزمون تقسیم شدند. این نسبت به‌دلیل تعادل بین حجم داده‌های آموزشی و ارزیابی انتخاب شده است تا مدل‌ها بتوانند به‌طور مناسب آموزش ببینند و در عین حال، عملکرد آن‌ها با داده‌های جدید بررسی شود.

تقسیم‌بندی تصادفی: برای اطمینان از نمایندگی مناسب داده‌ها در هر دو مجموعه، تقسیم‌بندی به‌صورت تصادفی[6] انجام شد. این روش کمک می‌کند تا مدل‌ها روی نماینده‌های مناسبی از داده‌ها آموزش ببینند و ارزیابی شوند (هان، پی و کامبر[7]، ۲۰۱۱).

مدل‌های یادگیری ماشین مورد استفاده

انتخاب مدل‌ها

در این پژوهش، ۴ مدل یادگیری جمعی[8]، برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان انتخاب و مقایسه شدند:

1. Mean Imputation
2. Mode Imputation
3. Outliers
4. Training
5. Test
6. Random Split
7. Han, Pei & Kamber
8. Ensemble Learning

XGBoost[1]

یکی از الگوریتم‌های یادگیری ماشین برپایه درخت تصمیم است که از تکنیک گرادیان بوستینگ[2] برای ایجاد مدل‌های پیش‌بینی استفاده می‌کند. در این الگوریتم، مدل‌ها به صورت متوالی ساخته می‌شوند و هر مدل جدید تلاش می‌کند تا خطاهای مدل قبلی را کاهش دهد.

هدف اصلی XGBoost حداقل کردن تابع هزینه است که معمولاً شامل یک تابع از دست دادن L[3] و یک عبارت تنبیهی برای جلوگیری از بیش‌برازش است. تابع هزینهٔ کلی به صورت زیر است:

$$L(\theta) = \sum_{i=1}^{n} L\left(Y_i, \hat{Y}_i\right) + \sum_{k=1}^{K} \Omega(f_k)$$ رابطهٔ ۱)

که در آن، Y_i مقدار واقعی نمونه iام؛ $\hat{Y}_i$ مقدار پیش‌بینی‌شده برای نمونه iام؛ $L\left(Y_i, \hat{Y}_i\right)$ تابع از دست دادن، به‌عنوان مثال، خطای مربعات میانگین (MSE) برای رگرسیون یا آنتروپی کراس برای طبقه‌بندی؛ $\Omega(f_k)$ عبارت تنبیهی برای پیچیدگی مدل که به صورت زیر تعریف می‌شود:

$$\Omega(f) = \gamma T + \frac{1}{2}\,\lambda \sum_{j} \omega_j^2$$ رابطهٔ ۲)

که در آن، T تعداد برگ‌های درخت تصمیم؛ ω_j وزن‌های گره‌های برگ؛ γ و λ ابرپارامترهایی هستند که پیچیدگی مدل را کنترل می‌کنند.

به‌روزرسانی مدل در هر مرحله: در هر مرحله از الگوریتم XGBoost، به‌جای حداقل‌سازی مستقیم تابع هزینه، تقریب مرتبهٔ دوم برای تابع از دست‌دادن به‌کار می‌رود. این بهینه‌سازی با استفاده از گرادیان[4] و هسین[5] انجام می‌شود:

$$L^{(t)} \approx \sum_{j=1}^{n} \left[g_j f_t(x_j) + \frac{1}{2} h_j f_t^2(x_i) \right] + \Omega(f_t)$$ رابطهٔ ۳)

که در آن، g_i گرادیان تابع از دست دادن و h_i هسین تابع از دست دادن می‌باشند (چن و گسترین[6]، ۲۰۱۶)

LightGBM[7]

یک الگوریتم تقویت گرادیان است که از روش‌های بهینه‌سازی پیشرفته برای افزایش سرعت و کارایی استفاده می‌کند و به‌ویژه برای مجموعه داده‌های بزرگ و پیچیده بسیار مناسب است. این الگوریتم برای ساختن درخت‌های تصمیم از

1. Extreme Gradient Boosting
2. Gradient Boosting
3. Loss function
4. Gradient
5. Hessian
6. Chen and Guestrin
7. Light Gradient Boosting Machine

روش‌های نوآورانه‌ای مانند تقسیم مبتنی‌بر برگ[1] و روش هیستوگرام برای تقسیم‌بندی[2] استفاده می‌کند. هدف LightGBM، مشابه XGBoost، حداقل‌سازی یک تابع هزینه است که یک تابع از دست‌دادن و یک عبارت تنبیهی برای پیچیدگی مدل را شامل می‌شود.

یکی از تفاوت‌های اصلی LightGBM با سایر الگوریتم‌ها، استفاده از تقسیم‌بندی هیستوگرامی برای کاهش زمان محاسباتی است. در این روش، ویژگی‌های پیوسته به دسته‌هایی (bins) تقسیم و محاسبات روی این دسته‌ها انجام می‌شود. این روش باعث کاهش قابل‌توجه پیچیدگی محاسباتی می‌شود و الگوریتم را سریع‌تر می‌کند.

به‌روزرسانی مدل در هر مرحله: مشابه XGBoost، LightGBM نیز از تقریب مرتبهٔ دوم برای تابع از دست دادن استفاده می‌کند. این بهینه‌سازی‌ها به LightGBM کمک می‌کنند تا درخت‌های تصمیمی با دقت بیشتر و کارایی بهتر ایجاد کنند (کی و همکاران[3]، ۲۰۱۷).

CatBoost

یک الگوریتم قدرتمند برپایه تقویت گرادیان است که به‌طور خاص برای داده‌های دسته‌ای طراحی شده است. برخلاف بسیاری از الگوریتم‌های یادگیری ماشین که به پیش‌پردازش داده‌های دسته‌ای (مانند One-Hot Encoding) نیاز دارند، CatBoost می‌تواند مستقیماً با این نوع داده‌ها کار کند که این یکی از مزایای مهم آن است. CatBoost نیز مانند سایر مدل‌های تقویت گرادیان (مانند XGBoost و LightGBM) یک تابع هزینه را بهینه می‌کند که شامل دو بخش است: تابع از دست‌دادن و عبارت تنبیهی؛ اما مهم‌ترین ویژگی CatBoost استفاده از روش Ordered Boosting است. در روش‌های معمول تقویت گرادیان، مدل‌ها با استفاده از کل داده‌ها آموزش می‌بینند که ممکن است باعث overfitting شود. اما در CatBoost، داده‌ها به صورت مرتب‌شده استفاده می‌شوند تا مدل‌های جدید تنها از داده‌هایی استفاده کنند که قبلاً توسط مدل‌های قبلی دیده نشده‌اند. این روش به کاهش overfitting کمک می‌کند و مدل نهایی را بهبود می‌بخشد.

Ordered Boosting در CatBoost: یکی از تفاوت‌های کلیدی CatBoost با دیگر الگوریتم‌های تقویت گرادیان، استفاده از Ordered Boosting به جای Standard Boosting است. در این روش، داده‌ها به‌ترتیب خاصی مرتب می‌شوند و هر مدل جدید براساس اطلاعاتی که تا آن مرحله جمع‌آوری‌شده، آموزش می‌بینند. این باعث کاهش نشت داده و جلوگیری از بیش‌برازش می‌شود.

به‌روزرسانی مدل در هر مرحله: مشابه سایر الگوریتم‌های تقویت گرادیان، در هر مرحله از الگوریتم CatBoost، مدل‌ها با استفاده از اطلاعات گرادیان و هسین به‌روزرسانی می‌شوند که مشابه سایر مدل‌های تقویت گرادیان است؛ اما تفاوت اصلی CatBoost در نحوهٔ استفاده از داده‌های دسته‌ای و Ordered Boosting است (پروخورنکووا، گوسف، ووروبف، دوروگوش و گولین[4]، ۲۰۱۸).

1. Leaf-wise growth
2. Histogram-based splitting
3. Ke et al.
4. Prokhorenkova, Gusev, Vorobev, Dorogush & Gulin

Random Forest

یک مدل مبتنی بر درخت تصمیم‌گیری است که با ترکیب چندین درخت به صورت تصادفی، دقت پیش‌بینی را افزایش می‌دهد. در Random Forest، چندین درخت تصمیم‌گیری ساخته می‌شود که هرکدام، براساس یک نمونه تصادفی از داده‌های آموزشی[1] ایجاد می‌شوند. سپس، پیش‌بینی نهایی با ترکیب نتایج این درخت‌ها به دست می‌آید. پیش‌بینی نهایی به صورت زیر محاسبه می‌شود:

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$ رابطهٔ ۴)

$\hat{f}(x)$ پیش‌بینی نهایی است؛ $f_b(x)$ پیش‌بینی bامین درخت برای نمونه x است؛ B تعداد درخت‌ها در Random Forest است.

برای رگرسیون، میانگین پیش‌بینی‌های تمام درخت‌ها محاسبه می‌شود:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$ رابطهٔ ۵)

برای طبقه‌بندی، رأی‌گیری اکثریت از پیش‌بینی‌های درخت‌ها انجام می‌شود:

$$\hat{y} = mode\ (f_1(x), f_2(x), \dots, f_B(x))$$ رابطهٔ ۶)

فرمول برای محاسبهٔ خطا در Random Forest: یکی از جنبه‌های مهم Random Forest، توانایی اندازه‌گیری خطای خارج از کیسه[2] است. این روش از داده‌هایی استفاده می‌کند که در ساختن یک درخت خاص استفاده نشده‌اند و برای ارزیابی عملکرد مدل، بدون نیاز به مجموعه داده آزمایشی جداگانه به کار می‌رود. فرمول محاسبهٔ خطای OOB به صورت زیر است:

$$OOB\ Error = \frac{1}{N}\sum_{i=1}^{N} L\ (y_i, \hat{y}oo_{B,i})$$ رابطهٔ ۷)

که در آن، N تعداد نمونه‌ها؛ y_i مقدار واقعی؛ $\hat{y}oo_{B,i}$ پیش‌بینی با استفاده از درخت‌هایی که این نمونه را ندیده‌اند؛ L تابع خطا، مثلاً MSE برای رگرسیون یا Cross-Entropy برای طبقه‌بندی (بریمن[3]، ۲۰۰۱).

آموزش مدل‌ها

هر یک از این مدل‌ها روی داده‌های آموزشی، آموزش دیدند؛ سپس با استفاده از داده‌های آزمون ارزیابی شدند. برای هر مدل، هایپرپارامترها با استفاده از روش جست‌وجوی شبکه‌ای[4] بهینه‌سازی شدند تا بهترین عملکرد ممکن به دست آید.

1. Bootstrap Sample
2. Out-of-Bag Error, OOB Error
3. Breiman
4. Grid Search

ارزیابی مدل‌ها

مدل‌ها با استفاده از معیارهای زیر ارزیابی شدند.

MSE[1] معیاری است که برای سنجش میانگین مربع خطاهای پیش‌بینی‌شده استفاده می‌شود. این معیار میزان اختلاف بین مقادیر واقعی و پیش‌بینی‌شده توسط مدل را اندازه‌گیری می‌کند (کاظمیان، افشارکاظمی، فتحی هفشجانی و معتدل، ۱۴۰۲) . MSE با استفاده از فرمول زیر محاسبه می‌شود:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$ رابطهٔ ۸)

که در آن، n تعداد نمونه‌ها؛ y_i مقدار واقعی نمونه i؛ $\hat{y}_i$ مقدار پیش‌بینی‌شده برای نمونه i؛ MSE یک معیار حساس به انحرافات بزرگ است، بنابراین اگر مدل پیش‌بینی‌های بزرگی داشته باشد که از مقدار واقعی به دور باشند، MSE افزایش می‌یابد (جیمز، ویتن، هستی و تبشیرانی[2]، ۲۰۱۳).

RMSE[3] همان MSE است؛ اما با ریشه‌گیری که به واحد اصلی متغیر پیش‌بینی‌شده برمی‌گردد (هادسون[4]، ۲۰۲۲). این معیار نشان‌دهندهٔ میزان انحراف استاندارد خطاهاست و به‌طور مستقیم قابل‌تفسیر است. RMSE برای تفسیر دقیق‌تر خطاها مفید است؛ زیرا مستقیماً به همان واحد داده‌ها مرتبط است (هیندمن و کوهلر[5]، ۲۰۰۶). فرمول RMSE به شکل زیر است:

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$ رابطهٔ ۹)

R^2[6] معیاری رایج برای سنجش قدرت توضیح‌دهندگی مدل است که نشان می‌دهد چه درصدی از واریانس مقادیر وابسته توسط مدل توضیح داده می‌شود (خانی، کزازی و بیرامی، ۱۴۰۱). فرمول R^2 به شکل زیر است:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(\bar{y} - \hat{y}_i)^2}$$ رابطهٔ ۱۰)

که در آن $\bar{y}$ میانگین مقادیر واقعی است و مخرج کسر واریانس مقادیر واقعی است. مقدار R^2 بین صفر و ۱ است، جایی که ۱ نشان‌دهندهٔ مدل کامل است که تمام واریانس را توضیح می‌دهد و صفر نشان‌دهندهٔ مدلی است که هیچ واریانسی را توضیح نمی‌دهد (فراست[7]، ۲۰۱۹).

1. Mean Squared Error
2. James, Witten, Hastie & Tibshirani
3. Root Mean Squared Error
4. Hodson
5. Hyndman and Koehler
6. R-squared
7. Frost

Training Time مدت زمانی است که مدل برای یادگیری و تنظیم پارامترهای خود از مجموعه داده‌ها صرف می‌کند. این معیار نشان‌دهندهٔ کارایی زمانی مدل در فرایند آموزش است و هرچه مدل سریع‌تر آموزش ببیند، از نظر عملیاتی بهتر است، به‌ویژه برای داده‌های بزرگ (گودفلا، بنژیو و کورویل[1]، ۲۰۱۶).

Inference Time به مدت زمانی اشاره دارد که مدل پس از آموزش برای انجام پیش‌بینی صرف می‌کند. این معیار به‌ویژه برای کاربردهای بی‌درنگ[2] اهمیت دارد. مدل‌هایی که زمان استنتاج کمتری دارند، در کاربردهای آنلاین و بی‌درنگ مفیدترند (ژانگ، بنژیو، هاردت، ریچ و وینالز[3]، ۲۰۱۷).

نتایج هر مدل با استفاده از معیارهای ارزیابی مختلف مورد بررسی و مقایسه قرار گرفت. این مقایسه به‌منظور شناسایی مدل بهینه برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان انجام شد.

یافته‌های پژوهش

در این پژوهش، ۴ مدل یادگیری جمعی شامل XGBoost، LightGBM، CatBoost و Random Forest برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری ارزیابی شده است. برای ارزیابی و مقایسه این مدل‌ها از معیارهای متعددی که در بخش قبل بیان شد، استفاده شده است. تمامی مدل‌ها روی سیستمی با پردازنده Intel Core i5-7200U و ۸ گیگابایت رم و پایتون ۳/۱۲ اجرا شده است.

ارزیابی عملکرد مدل‌ها

جدول ۳ نتایج حاصل از اجرای مدل‌ها را بر اساس معیارهای مختلف نشان می‌دهد.

جدول ۳. مقایسه عملکرد مدل‌های یادگیری جمعی برای پیش‌بینی رتبهٔ کشوری

Model	MSE	RMSE	R^2	Training Time (s)	Inference Time (s)
XGBoost	۱۵۶۹۸۵۱/۷۴	۱۲۵۲/۹۴	۰/۷۵۱۹	۰/۴۵	۰/۰۳
LightGBM	۱۵۸۲۸۰۶/۴۲	۱۲۵۸/۱۰	۰/۷۴۹۸	۰/۲۴	۰/۰۶
CatBoost	۱۷۰۶۲۸۴/۳۸	۱۳۰۶/۲۵	۰/۷۳۰۳	۰/۷۴	۰/۰۰
Random Forest	۱۶۸۰۶۵۲/۳۳	۱۲۹۶/۴۰	۰/۷۳۴۴	۲۶/۷۱	۰/۹۸

تحلیل و تفسیر نتایج

دقت پیش‌بینی

XGBoost: این مدل با کمترین مقدار MSE (۱۵۶۹۸۵۱/۷۴) و RMSE (۱۲۵۲/۹۴) بهترین عملکرد را در بین مدل‌های دیگر داشته است. مقدار R^2 برابر با ۰/۷۵۱۹ نشان می‌دهد که این مدل توانسته است بخش بزرگی از واریانس داده‌ها را توضیح دهد. دقت بالای این مدل و نیز خطای کمتر آن، XGBoost را به گزینه‌ای ایدئال برای پیش‌بینی رتبهٔ کشوری تبدیل می‌کند.

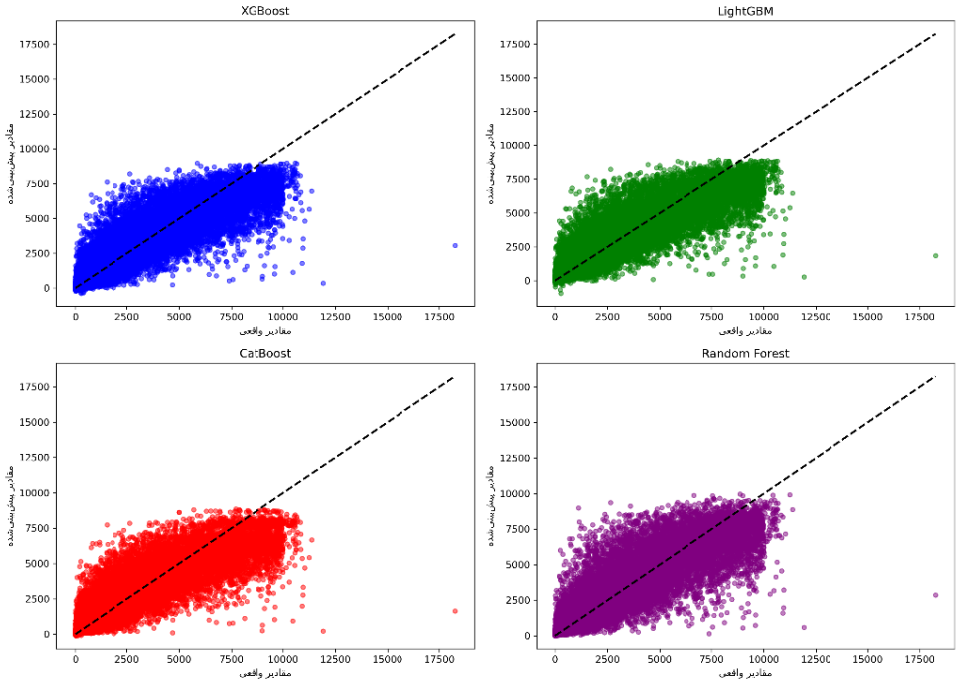
1. Goodfellow, Bengio & Courville
2. Real-time
3. Zhang, Bengio, Hardt, Recht & Vinyals

LightGBM: مدل LightGBM نیز عملکردی بسیار نزدیک به XGBoost داشته و با MSE (۱۵۸۲۸۰۶/۴۲) و RMSE (۱۲۵۸/۱۰) دومین بهترین مدل در این پژوهش بوده است. مقدار R^2 برابر با ۰/۷۴۹۸ نشان می‌دهد که این مدل نیز در توضیح واریانس داده‌ها توانمند است. با توجه به اینکه LightGBM زمان آموزش کمتری نسبت به XGBoost دارد، در کاربردهایی که زمان آموزش اهمیت دارد، می‌تواند ترجیح داده شود.

Random Forest: مدل Random Forest با MSE (۱۶۸۰۶۵۲/۳۳) و RMSE (۱۲۹۶/۴۰) عملکرد بهتری نسبت به CatBoost داشته است. مقدار R^2 برابر با ۰/۷۳۴۴ نشان می‌دهد که این مدل توانسته است بخش قابل‌توجهی از واریانس داده‌ها را توضیح دهد. با وجود اینکه زمان آموزش این مدل (۲۶/۷۱ ثانیه) بسیار طولانی‌تر از سایر مدل‌ها است، اما همچنان از نظر دقت پیش‌بینی بهتر از CatBoost عمل کرده است.

CatBoost: مدل CatBoost با MSE (۱۷۰۶۲۸۴/۳۸) و RMSE (۱۳۰۶/۲۵) ضعیف‌ترین عملکرد را در بین مدل‌های مورد بررسی داشته است. مقدار R^2 برابر با ۰/۷۳۰۳ نشان‌دهنده این است که این مدل قادر به توضیح کمتری از واریانس داده‌ها نسبت به سایر مدل‌ها بوده است. زمان پیش‌بینی این مدل بسیار سریع (۰/۰۰ ثانیه) است؛ اما به‌دلیل دقت کمتر، نسبت به Random Forest عملکرد ضعیف‌تری دارد.

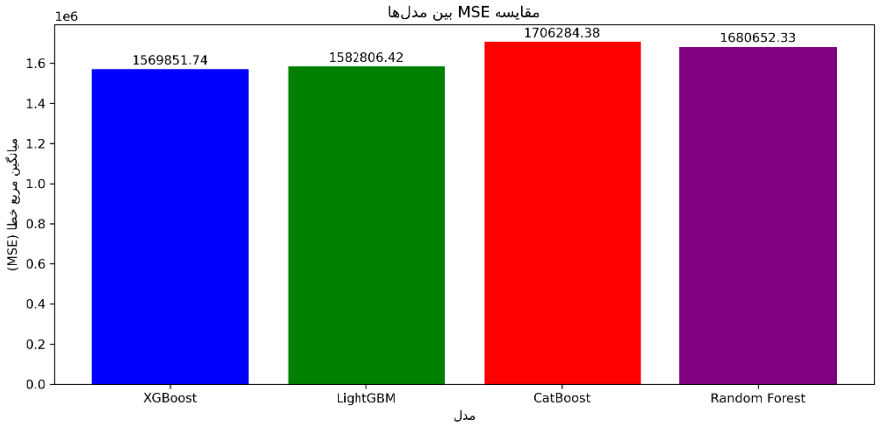
نمودار پراکندگی در شکل ۲، مقایسه‌ای بین مقادیر واقعی و پیش‌بینی‌شده برای هریک از مدل‌ها را نشان می‌دهد. نقاطی که نزدیک به خط ایدئال (خط‌سیاه نقطه‌چین) قرار دارند، نشان‌دهنده پیش‌بینی‌های دقیق‌ترند.



شکل ۲. مقایسه مقادیر پیش‌بینی‌شده و واقعی رتبه‌های کنکور توسط مدل‌های مختلف یادگیری جمعی

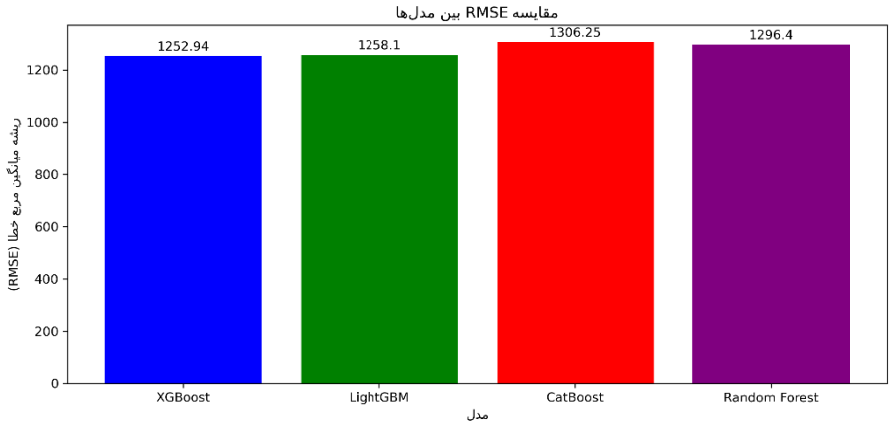
در این شکل مدل XGBoost به وضوح کمترین پراکندگی نسبت به خط ایدئال را نشان می‌دهد که این مسئله حاکی از دقت بالای آن است.

در شکل ۳، مقادیر MSE برای هر مدل به‌صورت میله‌ای نمایش داده شده است. این نمودار نشان می‌دهد که XGBoost به‌عنوان دقیق‌ترین مدل با کمترین مقدار MSE مشخص شده است که نشان می‌دهد این مدل توانسته است پیش‌بینی‌های نزدیکی به مقادیر واقعی ارائه دهد. LightGBM نیز عملکردی بسیار نزدیک به XGBoost داشته و در جایگاه دوم قرار گرفته است. Random Forest، در زمان آموزش نسبت به XGBoost و LightGBM کندتر بوده است؛ اما از نظر دقت پیش‌بینی بهتر از CatBoost عمل کرده است. CatBoost با بیشترین مقدار MSE، نسبت به دیگر مدل‌ها دقت کمتری در پیش‌بینی‌ها داشته است.



شکل ۳. مقایسه میانگین مربعات خطا (MSE) برای مدل‌های مختلف یادگیری جمعی در پیش‌بینی رتبهٔ کشوری

شکل ۴ مقادیر RMSE برای هر مدل را به‌صورت میله‌ای نمایش می‌دهد. این نمودار نشان می‌دهد که XGBoost کمترین خطای استاندارد را داشته و به‌عنوان دقیق‌ترین مدل از این نظر شناخته شده است. LightGBM نیز به‌دنبال XGBoost، خطای کمی داشته و عملکرد چشمگیری در پیش‌بینی‌ها از خود نشان داده است. Random Forest با خطای بیشتر نسبت به XGBoost و LightGBM، اما بهتر از CatBoost، در جایگاه بعدی قرار گرفته است. CatBoost بیشترین مقدار RMSE را دارد که نشان‌دهنده خطای بیشتری در پیش‌بینی‌های این مدل نسبت به سایر مدل‌ها است.



شکل ٤. مقایسه ریشهٔ میانگین مربعات خطا (RMSE) برای مدل‌های مختلف یادگیری جمعی در پیش‌بینی رتبهٔ کشوری

کارایی زمانی (Training Time و Inference Time)

XGBoost: این مدل علاوه‌بر دقت بالا، زمان آموزش کوتاهی (۰/۴۵ ثانیه) و زمان پیش‌بینی بسیار کمی (۰/۰۳ ثانیه) دارد. این ویژگی‌ها XGBoost را به گزینه‌ای مناسب برای کاربردهایی تبدیل می‌کند که به دقت و سرعت زیاد نیاز دارند.

LightGBM: مدل LightGBM با زمان آموزش کمتر (۰/۲۴ ثانیه) و زمان پیش‌بینی کمی بیشتر (۰/۰۶ ثانیه) نسبت به XGBoost، گزینه‌ای مناسب برای شرایطی است که زمان آموزش مهم است.

Random Forest: این مدل با زمان آموزش بسیار طولانی (۲۶/۷۱ ثانیه) و زمان پیش‌بینی نسبتاً زیاد (۰/۹۸ ثانیه)، از نظر زمانی بهینه نیست؛ اما از نظر دقت پیش‌بینی بهتر از CatBoost عمل کرده است.

CatBoost: این مدل با زمان آموزش (۰/۷۴ ثانیه) و زمان پیش‌بینی بسیار سریع (نزدیک به صفر)، در بخش زمان پیش‌بینی عالی عمل کرده است؛ اما از نظر دقت و عملکرد کلی نسبت به Random Forest ضعیف‌تر است.

با توجه به نتایج به‌دست‌آمده، مدل‌های XGBoost و LightGBM به‌عنوان بهترین گزینه‌ها برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان شناسایی شدند. این مدل‌ها نه‌تنها دقت زیادی در پیش‌بینی دارند، بلکه از کارایی زمانی مناسبی نیز برخوردارند. مدل Random Forest نیز عملکرد خوبی از خود نشان داده است؛ اما زمان طولانی آموزش آن ممکن است در برخی از کاربردها محدودیت ایجاد کند. مدل CatBoost به دلیل دقت کمتر و خطای بالاتر نسبت به سایر مدل‌ها، به‌عنوان ضعیف‌ترین مدل شناسایی شد، اگرچه زمان پیش‌بینی بسیار سریعی دارد.

بحث و نتیجه‌گیری

این پژوهش ۴ مدل یادگیری جمعی XGBoost، LightGBM، CatBoost و Random Forest را برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری، بر اساس داده‌های آزمون‌های قلمچی بررسی و مقایسه کرد. داده‌های استفاده‌شده از وبسایت کانون فرهنگی آموزش استخراج شد و پس از پیش‌پردازش و پاک‌سازی، مدل‌های یادگیری ماشین روی آن‌ها آموزش داده شدند. ارزیابی مدل‌ها با استفاده از معیارهایی نظیر Mean Squared Error (MSE)، Root Mean Squared Error (RMSE)، R-squared (R^2)، زمان آموزش و زمان پیش‌بینی انجام شد.

نتایج نشان داد که مدل‌های XGBoost و LightGBM از نظر دقت و کارایی بهترین عملکرد را داشتند. این مدل‌ها با مقادیر پایین‌تر MSE و RMSE و همچنین مقادیر بالاتر R^2 توانایی بیشتری را در پیش‌بینی دقیق رتبهٔ کشوری دانش‌آموزان نشان دادند. علاوه‌براین، زمان‌های آموزش و پیش‌بینی این مدل‌ها نیز نسبتاً کوتاه بود که نشان‌دهندهٔ کارایی بالای آن‌ها است. مدل Random Forest نیز عملکرد مناسبی داشت، هرچند زمان آموزش و پیش‌بینی آن طولانی‌تر بود، از نظر دقت پیش‌بینی بهتر از CatBoost عمل کرد. CatBoost با مقادیر بالاتر MSE و RMSE و مقادیر پایین‌تر R^2 عملکرد ضعیف‌تری نسبت به سایر مدل‌ها داشت، اما زمان پیش‌بینی بسیار سریعی را ارائه داد.

مطالعات مختلفی به بررسی کارایی مدل‌های یادگیری جمعی در پیش‌بینی عملکرد تحصیلی دانش‌آموزان پرداخته‌اند. در پژوهش حاضر، مدل‌های XGBoost و LightGBM به‌عنوان دقیق‌ترین مدل‌ها برای پیش‌بینی رتبهٔ کشوری دانش‌آموزان در کنکور سراسری شناخته شدند. این یافته‌ها با نتایج پژوهش بات و همکاران (۲۰۲۳) که نشان دادند روش‌های Boosting و Stacking در مجموعه‌های داده کوچک و بزرگ عملکرد بهتری دارند، هم‌خوانی دارد. همچنین، یافته‌های این پژوهش مشابه با نتایج یان و لیو (۲۰۲۰) است که نشان دادند مدل‌های Adaboost و Random Forest در پیش‌بینی عملکرد تحصیلی دقت زیادی دارند. در مقابل، مدل CatBoost که در این پژوهش عملکرد ضعیف‌تری داشت، برخلاف برخی مطالعات، مانند پروخورنکووا و همکاران (۲۰۱۸) بود آن‌ها این مدل را به‌دلیل توانایی آن در کار با داده‌های دسته‌ای و کاهش بیش‌برازش بسیار مؤثر دانسته‌اند. این تفاوت ممکن است به‌دلیل نوع داده‌های مورد استفاده در این پژوهش یا پارامترهای انتخاب شده برای آموزش مدل باشد. علاوه‌براین، نتایج پژوهش حاضر نشان می‌دهد که مدل‌های جمعی، به‌طور کلی نسبت به مدل‌های تکی دقت بیشتری دارند که این یافته با تحقیقات هان و همکاران (۲۰۱۷) هم‌راستاست. آن‌ها نیز نشان دادند که الگوریتم‌های جمعی، به‌ویژه Adaboost نسبت به الگوریتم‌های تکی مانند درخت تصمیم و شبکه عصبی عملکرد بهتری داشتند.

علاوه‌براین، در حالی که برخی مطالعات مانند ادج و کلونی (۲۰۱۸) به استفاده از مدل‌های جمعی ناهمگن مانند SVM و Random Forest پرداخته‌اند، این پژوهش نشان داد که مدل‌های مبتنی بر Boosting مانند XGBoost در شرایط داده‌های آزمون کنکور سراسری ایران عملکرد بهتری دارند. این تفاوت‌ها می‌تواند به‌دلیل تفاوت در نوع داده‌ها، ویژگی‌های مورد استفاده یا شرایط آموزشی متفاوت باشد. یکی از تفاوت‌های عمدهٔ این پژوهش با مطالعات گذشته، تمرکز بر پیش‌بینی دقیق رتبهٔ کشوری دانش‌آموزان در کنکور سراسری است. برخلاف بسیاری از مطالعات که به بررسی کلی عملکرد تحصیلی پرداخته‌اند، این پژوهش به‌طور خاص به پیش‌بینی رتبهٔ کشوری پرداخت و نشان داد که مدل‌های XGBoost و LightGBM در این زمینه بسیار مؤثرند.

این پژوهش می‌تواند کاربردهای متعددی در زمینهٔ آموزش و پیش‌بینی تحصیلی داشته باشد. از جمله پیشنهادهای کاربردی، می‌توان به استفاده از مدل‌های پیش‌بینی رتبهٔ کشوری دانش‌آموزان در سیستم‌های مشاوره تحصیلی و انتخاب رشته اشاره کرد. مشاوران تحصیلی می‌توانند با بهره‌گیری از این مدل‌ها، توصیه‌های دقیق‌تری به دانش‌آموزان ارائه دهند و در تعیین مسیر تحصیلی آنان نقش مؤثری ایفا کنند. همچنین مدیران و تصمیم‌گیران حوزهٔ آموزش می‌توانند با تحلیل نتایج پیش‌بینی، تصمیم‌گیری‌های بهتری در تخصیص منابع آموزشی و خدمات به دانش‌آموزان داشته باشند. این مدل‌ها همچنین می‌توانند به شناسایی دانش‌آموزانی که در معرض افت تحصیلی هستند، کمک کند تا برنامه‌های آموزشی ویژه‌ای برای آن‌ها تدوین شود. در نهایت، استفاده از این مدل‌ها در بهینه‌سازی برنامه‌های آموزشی، می‌تواند به معلمان و مدیران در تدوین برنامه‌های کارآمدتر و مناسب‌تر برای دانش‌آموزان کمک کند.

از دیدگاه پژوهشی، پیشنهاد می‌شود در پژوهش‌های آینده از مدل‌های پیشرفته‌تر یادگیری عمیق مانند شبکه‌های عصبی چندلایه و شبکه‌های عصبی کانولوشنی استفاده شود. این مدل‌ها، به‌ویژه در صورت افزایش حجم داده‌ها می‌توانند دقت بالاتری نسبت به مدل‌های فعلی داشته باشند. علاوه‌براین، افزودن ویژگی‌های جدید، مانند عوامل اجتماعی ـ

اقتصادی، نوع مدارس و سوابق تحصیلی، می‌تواند به بهبود دقت پیش‌بینی‌ها کمک کند. استفاده از مدل‌های هیبریدی که نقاط قوت چندین مدل یادگیری را ترکیب می‌کنند، نیز می‌تواند به نتایج دقیق‌تر منجر شود. در ادامه، پیشنهاد می‌شود که مدل‌ها در شرایط مختلف، از جمله حجم داده‌های متفاوت و محیط‌های آموزشی متنوع، ارزیابی شوند تا کارایی آن‌ها در کاربردهای گوناگون سنجیده شود. همچنین، انجام تحلیل حساسیت برای شناسایی تأثیر هر ویژگی بر پیش‌بینی رتبهٔ کشوری دانش‌آموزان، می‌تواند به بهینه‌سازی مدل‌ها و درک بهتر از عوامل مؤثر در عملکرد دانش‌آموزان کمک کند.

منابع

اسدی، علی اکبر و جوادزاده، محمدعلی (۱۴۰۰). پیش‌بینی نمرات دانش‌آموزان با استفاده از روش یادگیری ماشین (مقایسه سه الگوریتم درخت تصمیم، جنگل تصادفی و رگرسیون خطی). *ششمین کنفرانس بین‌المللی پژوهش‌های کاربردی در کامپیوتر، برق و فناوری اطلاعات*.

بابایی، مرضیه و شویدی نوش آبادی، مهدی (۱۳۹۹). مروری بر روش‌های پیش‌بینی عملکرد دانشجویان با استفاده از الگوریتم‌های یادگیری ماشین. ششمین کنفرانس بین‌المللی نوآوری و تحقیق در علوم مهندسی.

خانی، امیر محمد؛ کزازی، ابولفضل و بیرامی، ثریا (۱۴۰۱). مدل‌سازی ساختاری بر مبنای مدیریت زنجیره‌تأمین در رابطه با مدیریت کیفیت جامع، نگهداری‌و تعمیرات بهره‌ور فراگیر، سازمان یادگیرنده و عملکرد عملیاتی. *مطالعات مدیریت صنعتی*، ۲۰(۶۵)، ۳۹-۸۴

زنگوئی، حمید و فاطمی، سید امید (۱۴۰۰). پیش‌بینی دانش‌آموزان در معرض خطر افت تحصیلی با استفاده از تجزیه‌وتحلیل یادگیری در سیستم مدیریت یادگیری. *مجله ایرانی آموزش از دور*، ۳(۲)، ۳۲-۴۴.

سالاری، مژده؛ رادفر، رضا و فقیهی، مهدی (۱۴۰۳). پیش‌بینی عملکرد دانشجویان با استفاده از الگوریتم‌های یادگیری ماشین و داده کاوی آموزشی (مطالعه موردی دانشگاه شاهد). *مطالعات مدیریت کسب‌وکار هوشمند*، ۱۲(۴۷)، ۳۱۵-۳۶۶.

قدوسی، محمد، میرسعیدی، فاطمه و کوشا، حمیدرضا (۱۳۹۹). پیش‌بینی و تحلیل عملکرد دانشجویان به کمک تکنیک‌های داده‌کاوی به‌منظور بهبود عملکرد تحصیلی. *فناوری آموزش*، ۱۴(۴)، ۸۲۱-۸۳۴.

کاظمیان، مینا؛ افشارکاظمی، محمدعلی؛ فتحی هفشجانی، کیامرث و معتدل، محمدرضا (۱۴۰۲). ارائه مدل هوشمند تعیین قیمت فولاد با رویکرد ترکیبی نظریه بازی‌ها و الگوریتم‌های یادگیری ماشین. *مدیریت صنعتی*، ۱۵(۳)، ۴۷۸- ۵۰۷.

References

Adejo, O. W. & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. *Journal of Applied Research in Higher Education*, *10*(1), 61–75. https://doi.org/10.1108/jarhe-09-2017-0113

Asadi, A.A. & Javadzadeh, M.A. (2021). Prediction of students' grades using machine learning method (comparison of three algorithms of decision tree, random forest and linear regression). *The 6th International Conference on Applied Research in Computer, Electricity and Information Technology*. (*in Persian*)

Babaei, M. & Shovidi Noushabadi, M. (2019). A review of student performance prediction methods using machine learning algorithms. *The 6th International Conference on Innovation and Research in Engineering Sciences*, https://civilica.com/doc/1033467 (*in Persian*)

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Butt, N. A., Mahmood, Z., Shakeel, K., Alfarhood, S., Safran, M. & Ashraf, I. (2023). Performance Prediction of students in Higher Education using Multi-Model Ensemble Approach. *IEEE Access*, *11*, 136091–136108. https://doi.org/10.1109/access.2023.3336987

Chen, T. & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Daniele, V. (2021). Socioeconomic inequality and regional disparities in educational achievement: The role of relative poverty. *Intelligence*, *84*, 101515. https://doi.org/10.1016/j.intell.2020.101515

Frost, J. (2019). *Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models*. Statistics by Jim Publishing. https://doi.org/10.1007/978-3-030-36077-8

Ghodoosi, M., Mirsaeedi, F. & Koosha, H. (2020). Predicting and analyzing the performance of students through data mining techniques to improve academic performance. *Technology of Education Journal (TEJ)*, *14*(4), 821-834. doi: 10.22061/jte.2019.4902.2134 (*in Persian*)

Gibson, D. C. & Webb, M. E. (2015). Data science in educational assessment. *Education and Information Technologies*, *20*, 697-713. https://doi.org/10.1007/s10639-015-9411-7

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press. https://doi.org/10.5555/3086952

Han, J., Pei, J. & Kamber, M. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5

Han, M., Tong, M., Chen, M., Liu, J. & Liu, C. (2017). Application of Ensemble Algorithm in Students' Performance Prediction. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), 735-740.*

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, *15*(14), 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Injadat, M., Moubayed, A., Nassif, A. B. & Shami, A. (2020). Systematic ensemble model selection approach for educational data mining. *Knowledge-Based Systems*, *200*, 105992. https://doi.org/10.1016/j.knosys.2020.105992

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kazemian, M., Afsharkazemi, M. A., Fathi Hafashjani, K. & Motadel, M. (2023). Presenting Smart Steel Pricing Model: An Integration of Game Theory and Machine Learning Algorithms. *Industrial Management Journal*, *15*(3), 478-507. doi: 10.22059/imj.2023.356697.1008039 (*in Persian*)

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30. https://doi.org/10.5555/3294996.3295074

Khani, A. M., Kazzazi, A. & Birami, S. (2022). Structural modeling based on supply chain management in relation to total quality management, maintenance and comprehensive productivity, learning organization and operational performance. *Industrial Management Studies*, *20*(65), 39-84. doi: 10.22054/jims.2022.62763.2688 (*in Persian*)

Navarro, C. L. A., Damen, J. A., Takada, T., Nijman, S. W., Dhiman, P., Ma, J. , ... & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ*, *375*. https://doi.org/10.1136/bmj.n2281

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31. https://doi.org/10.48550/arXiv.1706.09516

Sakri, S. & Saleh, A. (2020). RHEM: a robust hybrid ensemble model for students' performance assessment on cloud computing course. *International Journal of Advanced Computer Science and Applications*, *11*(11). https://doi.org/10.14569/ijacsa.2020.0111150

Salari, M., Radfar, R. & Faghihi, M. (2024). Predicting students' performance using machine learning algorithms and educational data mining (a case study of Shahed University). *Business Intelligence Management Studies*, *12*(47), 315-366. doi: 10.22054/ims.2023.75523.2375 (*in Persian*)

Wang, N. Z. & Shi, N. Y. (2016). Prediction of the admission lines of college entrance examination based on machine learning. *2016 2nd IEEE International Conference on Computer and Communications (ICCC), 332-335.* https://doi.org/10.1109/compcomm.2016.7924718

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments,* 9(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yan, L. & Liu, Y. (2020). An ensemble prediction model for potential student recommendation using machine learning. *Symmetry*, *12*(5), 728. https://doi.org/10.3390/sym12050728

Zangooei, H. & Fatemi, O. (2021). Predicting students at risk of academic failure using learning analytics in the learning management system. *Quarterly of Iranian Distance Education Journal*, *3*(2), 32-44. doi: 10.30473/idej.2022.63913.1104 (*in Persian*)

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2017). *Understanding deep learning requires rethinking generalization*. https://doi.org/10.48550/arXiv.1611.03530