# مدل یادگیری ماشین انباشته برای دسته‌بندی و پیش‌بینی بیماری‌های کبدی

**بابک آذرنوید[1*]، محسن عبدالحسین‌زاده[2]، حجت امامی[3]**

استادیار گروه ریاضی و علوم کامپیوتر، دانشکدهٔ علوم پایه، دانشگاه بناب، بناب، ایران[1*]

استادیار گروه ریاضی و علوم کامپیوتر، دانشکدهٔ علوم پایه، دانشگاه بناب، بناب، ایران[2]

دانشیار گروه مهندسی کامپیوتر، دانشکدهٔ فنی و مهندسی ، دانشگاه بناب، بناب، ایران[3]

## چکیده

**بیماری‌های کبدی یکی از علل اصلی مرگ‌ومیر هستند که تأثیر عمیقی بر زندگی افراد دارند و تشخیص آن‌ها در مراحل اولیه بسیار حیاتی است. هدف این پژوهش، توسعه و ارزیابی مدل یادگیری ماشین انباشته (SML) برای تشخیص و پیش‌بینی دقیق بیماری‌های کبدی است. مدل SML با استفاده از ساختار دولایه، الگوریتم‌های مختلف را ترکیب کرده تا مشکل بیش‌برازش را برطرف کند و دقت پیش‌بینی را افزایش دهد. در لایهٔ نخست، چهار الگوریتم شامل درخت تصادفی نامحدود (ET)، درخت تصمیم (DT)، جنگل تصادفی (RF) و تقویت گرادیان شدید (XGB) برای پیش‌بینی اولیه استفاده می‌شوند. در لایهٔ دوم، الگوریتم رگرسیون ترابری (LR) بر اساس خروجی لایهٔ نخست آموزش داده می‌شود تا پیش‌بینی نهایی انجام شود. تنظیم پارامترها با الگوریتم جست‌وجوی شبکهٔ توری (GS) انجام شده‌است. داده‌های مورد استفاده شامل ۶۱۵ نمونه‌داده با دوازده ویژگی از پایگاه دانشگاه کالیفرنیا در ایروین است که ۷۰٪ برای آموزش و ۳۰٪ برای آزمایشی اختصاص یافته است. نتایج اعتبارسنجی متقابل $k=5$ نشان می‌دهد که مدل پیشنهادی با صحت ۰.۹۹۴۰ و معیار F1 برابر ۰.۹۸۸۰، عملکرد برتری نسبت به سایر روش‌ها دارد. این پژوهش می‌تواند به کاهش مرگ‌ومیر ناشی از بیماری‌های کبدی کمک شایانی کند.**

**واژگان کلیدی: بیماری‌های کبد، تشخیص زودهنگام، یادگیری ماشین، مدل یادگیری ماشین انباشته، اعتبارسنجی متقابل.**

# Stacking machine learning model for classification and prediction of liver diseases

**Babak Azarnavid[1*], Mohsen Abdolhosseinzadeh[2], Hojjat Emami[3]**

Assistant Professor, Department of Mathematics and Computer Science, Basic Science Faculty, University of Bonab, Bonab, Iran[1*]

Assistant Professor, Department of Mathematics and Computer Science, Basic Science Faculty, University of Bonab, Bonab, Iran[2]

Associate Professor, Department of Computer Engineering, Faculty of Engineering, University of Bonab, Bonab, Iran[3]

## Abstract

**Liver diseases are among the leading causes of mortality worldwide, deeply influencing individuals' lives, often at younger ages when they are in the prime of their personal and professional lives. The insidious nature of these diseases lies in their early initial symptoms, which frequently goes unnoticed until the condition has progressed to an advanced stage. This delay in diagnosis not only diminishes the chances of successful treatment but also places an immense emotional and financial burden on patients as well as families. Early detection is therefore critical, as it can significantly alter the course of the disease, improving survival rates and quality of life. However, traditional diagnostic methods often fall**

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۳/۱۱/۰۷ • تاریخ پذیرش: ۱۴۰۴/۰۴/۲۹ • تاریخ انتشار: ۱۴۰۴/۶/۲۲ • نوع مطالعه: پژوهشی





[ DOI: 10.61882/jsdp.22.2.79 ]

**short in terms of speed, accuracy, and accessibility, particularly in resource-limited settings. This underscores the urgent need for innovative approaches to liver disease detection and its management.**
**Machine learning (ML) has been emerged as a powerful tool in this regard, offering the potential to revolutionize how we diagnose and predict liver diseases. By leveraging vast datasets—ranging from clinical records and laboratory results to imaging data—ML algorithms can uncover complex patterns and correlations that may elude human experts. These insights can lead to earlier and more accurate diagnoses, enabling timely interventions that can save lives. Among the various ML approaches, stacked machine learning (SML) models stand out for their ability to combine the strengths of multiple algorithms, mitigating the limitations of individual models and enhancing overall performance. This research focuses on developing and evaluating an SML model specifically designed for the accurate diagnosis, classification, and prediction of liver diseases, with the goal of addressing some of the most pressing challenges in this field.**
**The proposed SML model employs a sophisticated two-layer architecture to tackle common issues such as overfitting and improving prediction accuracy. In the first layer, the model integrates four robust base learner algorithms: Extremely Randomized Trees (ET), Decision Tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGB). Each of these algorithms contributes unique strengths, such as handling high-dimensional data, capturing non-linear relationships, and reducing variance. The predictions generated by these base learners are then fed into the second layer, where a Logistic Regression (LR) algorithm synthesizes the outputs to produce the final prediction. This layered approach ensures that the model benefits from the collective intelligence of multiple algorithms, resulting in more reliable and precise outcomes. To further optimize performance, the Grid Search (GS) algorithm was employed to fine-tune the parameters of the learning algorithms, ensuring that the model operates at its full potential. This study employs dataset from the University of California, Irvine (UCI) Machine Learning Repository. A sample size of 615 instances has been utilized to implement the proposed methodologies, with a stratified division of 70% for training and 30% allocated for testing purposes. The results of this research seems to be highly promising. Evaluation based on 5-fold cross-validation demonstrates that the proposed SML model outperforms existing methods, achieving an impressive 0.9940 accuracy and a 0.9880 F1-score on the test data. These metrics not only highlight the model's exceptional predictive capabilities but also underscore its potential to serve as a valuable tool for clinicians in real-world settings. By providing accurate and timely diagnoses, the SML model can help reduce the mortality and morbidity associated with liver diseases, offering hope to patients and their families.**
**Beyond the technical achievements, the human impact of this research cannot be overstated. For patients, the SML model represents a lifeline—a chance to detect liver diseases early, when treatment seems most effective, and to avoid the devastating consequences of late-stage diagnoses. For healthcare providers, it offers a reliable and efficient diagnostic tool that can enhance decision-making and improve patient outcomes. Also, for society as a whole, it signifies a step forward in the fight against a disease that disproportionately affects vulnerable populations, including those in underserved regions where access to advanced medical care is limited. In essence, this research is not just about developing a sophisticated algorithm; it is also about harnessing the power of machine learning to make a tangible difference in people's lives. By bridging the gap between cutting-edge technology and human care, the proposed SML model embodies the potential of computer science to address some of the most critical health challenges of our time. It is a testament to the transformative power of innovation, compassion, and collaboration in the pursuit of better health for all.**

**Keywords: Liver diseases, Early Diagnosis, Machine Learning, Cumulative Machine Learning Model, Cross-Validation.**

## ۱-مقدمه

کبد نقش حیاتی در عملکردهای مختلف بدن، از جمله سم‌زدایی، متابولیسم و سنتز پروتئین‌ها دارد و سلامت آن برای رفاه کلی ضروری است. بیماری‌های کبدی به‌عنوان یکی از چالش‌های مهم بهداشتی جهانی شناخته می‌شوند که به‌دلیل شیوع بالا، امکان بروز عوارض شدید و تأثیرات عمیق بر نرخ مرگ‌ومیر، توجه ویژه‌ای را به خود جلب کرده‌اند. هر ساله حدود دو میلیون نفر به‌دلیل مشکلات مربوط به کبد از جمله هپاتیت ویروسی و سیروز جان خود را از دست می‌دهند [۱،۲]. این بیماری‌ها نه‌تنها بر نرخ مرگ‌ومیر تأثیر می‌گذارند، بلکه کیفیت زندگی میلیون‌ها نفر در سراسر جهان را نیز به‌شدت تحت‌تأثیر قرار می‌دهند.

بیماری‌های کبدی ناشی از عوامل مختلفی هستند. ویروس‌ها و انگل‌ها می‌توانند کبد را آلوده کرده و باعث التهاب و اختلال در عملکرد آن شوند. هپاتیت‌های A، B، C، D و E رایج‌ترین ویروس‌هایی هستند که عفونت‌های کبدی ایجاد می‌کنند؛ همچنین مصرف بیش از حد الکل و تجمع چربی در کبد می‌تواند منجر به آسیب‌های کبدی مانند هپاتیت الکلی، بیماری کبد چرب و سیروز شود[۳]. بیماری مزمن کبدی می‌تواند به نتایج شدیدی مانند عدم جبران کبدی و هپاتوسلولار کارسینوما منجر شود؛ به‌ویژه در افرادی که دچار



[ DOI: 10.61882/jsdp.22.2.79 ]

فیبروز متوسط تا پیشرفته هستند[۴]. پیشرفت بیماری‌های کبدی به‌طورمعمول تا زمانی که به مراحل پیشرفته مانند سیروز برسد، نادیده گرفته می‌شود؛ جایی که آسیب غیرقابل‌برگشت رخ داده‌است. این موضوع اهمیت تشخیص زودهنگام و مداخله را برجسته می‌کند و ضرورت استفاده از روش‌های مؤثر برای پیش‌بینی و تشخیص بیماری‌های کبدی را در مراحل اولیه قوت می‌بخشد.

پیشرفت‌های اخیر در پردازش داده‌ها و یادگیری ماشین فرصت‌های امیدبخشی برای بهبود پیش‌بینی بیماری‌های کبدی و عوارض آن‌ها ارائه می‌دهند. با تجزیه‌وتحلیل داده‌های بالینی، شخصی و مولکولی، مدل‌های یادگیری ماشین می‌توانند الگوها و عوامل خطر را شناسایی کنند که ممکن است از طریق روش‌های تشخیص سنتی به‌راحتی قابل مشاهده نباشند. این رویکرد می‌تواند مداخلات به‌موقع را تسهیل و به‌طور بالقوه از پیشرفت بیماری کبدی جلوگیری کند و بار سلامت مرتبط را کاهش دهد. با توجه به افزایش شیوع بیماری‌های کبدی به‌دلیل عواملی مانند چاقی، سوءمصرف الکل و عفونت‌های ویروسی استفاده از فناوری برای بهبود پیش‌بینی به‌منظور به‌کارگیری در ابتکارات بهداشت عمومی که هدف آن‌ها مبارزه با بیماری کبد است، ضروری است.

ظهور هوش مصنوعی، به‌ویژه روش‌های یادگیری ماشین و داده‌کاوی فرصت‌های قابل توجهی را برای ارتقای خدمات بهداشتی از طریق تجزیه‌وتحلیل مجموعه‌داده‌های عظیم و شناسایی الگوهای معنی‌دار ارائه می‌دهد[۵، ۶، ۷ و ۸]. با وجود این پیشرفت‌ها، جامعهٔ پزشکی همچنان در پذیرش کامل این فناوری‌ها تردید دارد که منجر به عدم استفاده از پتانسیل عظیم آن می‌شود[۵]. یادگیری ماشین به‌عنوان یک ابزار پیشرفته در تشخیص بیماری‌ها به‌ویژه بیماری‌های کبدی و هپاتیت C اهمیت ویژه‌ای دارد. با تحلیل داده‌های گستردهٔ پزشکی و زیستی، الگوریتم‌های یادگیری ماشین می‌توانند الگوهای پیچیده‌ای را شناسایی کنند که ممکن است از دید انسان پنهان مانده باشند. این روش‌ها قابلیت افزایش دقت در تشخیص زودهنگام بیماری‌های کبدی را دارند، که به درمان سریع‌تر و مؤثرتر بیماران کمک می‌کنند[۹، ۱۰]؛ همچنین یادگیری ماشین می‌تواند در پیش‌بینی روند پیشرفت بیماری و پاسخ به درمان نیز مؤثر باشد و به پزشکان اطلاعات بهتری برای تصمیم‌گیری‌های بالینی ارائه دهد. با توجه به هزینهٔ بالای مادی و انسانی این بیماری‌ها در سطح جهانی، بهره‌گیری از یادگیری ماشین می‌تواند نقشی کلیدی در بهبود نتایج درمانی و کاهش عوارض ناشی از بیماری‌های کبدی ایفا کند.

در سال‌های اخیر چندین روش مبتنی بر یادگیری ماشین برای تشخیص بیماری‌های کبدی ارائه شده‌است؛ از جمله می‌توان به مدل تقویت سازگار (آدابوست)[1]، جنگل تصادفی [۱۱] ، شبکه‌های عصبی مصنوعی، ماشین بردار پشتیبان [۱۲]، درخت تصمیم، رگرسیون ترابری [۱۳] و درخت طبقه‌بندی افزایش گرادیان مبتنی بر هیستوگرام[2] [۵، ۶] اشاره کرد. این روش‌ها به کارایی قابل قبولی دست یافته‌اند؛ اما کارایی کسب‌شده هنوز با حالت ایده‌آل فاصله داشته و تلاش زیادی برای بهبود کارایی روش‌های موجود نیاز است؛ بنابراین در این پژوهش، چندین مدل یادگیری ماشین مورد ارزیابی قرار گرفته و درنهایت مدل یادگیری ماشین انباشته (SML) معرفی شد. هدف مدل پیشنهادی بهبود دقت پیش‌بینی از طریق ترکیب توانایی الگوریتم‌های متنوع یادگیری ماشین است؛ به همین منظور مدل پیشنهادی SML، با ترکیب پنج مدل داده‌محور شامل دسته‌بند بی‌نهایت تصادفی (ET)، درخت تصمیم (DT)، جنگل تصادفی (RF)، تقویت گرادیان شدید (XGB) و رگرسیون ترابری (LR) - در یک ساختار لایه‌ای یک مدل پیش‌بینی‌کنندهٔ توانمند ایجاد کرده که بر مشکل بیش‌برازش و کم‌برازش چیره و موجب بهبود دقت پیش‌بینی شده‌است.

به‌اختصار، نوآوری‌های پژوهش حاضر را می‌توان در موارد زیر خلاصه کرد:

- **مدل‌سازی مسئلهٔ تشخیص و پیش‌بینی بیماری‌های کبدی به صورت یک مسئلهٔ دسته‌بندی نظارتی[3]:** مسئلهٔ پیش‌بینی بیماری‌های کبدی به صورت یک مسئلهٔ یادگیری ماشین نظارتی مدل‌سازی شده‌است که موجب می‌شود به سادگی بتوان از طیف وسیعی از الگوریتم‌های یادگیری ماشین و تحلیل داده برای حل مسئله استفاده کرد.

- **ارائهٔ مدل یادگیری ماشین انباشته:** مدل SML با ترکیب هوشمندانهٔ الگوریتم‌های یادگیری متنوع در یک ساختار دو لایه‌ای موجب بهبود تشخیص و دسته‌بندی بیماری‌های کبدی می‌شود. مدل پیشنهادی با بهره‌گیری از چهار مدل یادگیر پایه شامل ET، DT، RF و XGB در لایهٔ نخست و الگوریتم LR در لایهٔ دوم مشکل بیش‌برازش را به نحو مطلوبی برطرف کرده و موجب بهبود پایداری و کارایی مدل‌های منفرد می‌شود؛ همچنین برای تنظیم فراپارامترهای الگوریتم‌های یادگیر پایه و فرامدل از الگوریتم جست‌وجوی شبکهٔ توری (GS) استفاده شده‌است.

[1] Adaptive boosting
[2] Histogram-based gradient boosting classification tree (HGBoost)
[3] Supervised




[ DOI: 10.61882/jsdp.22.2.79 ]

- **ارزیابی روش پیشنهادی:** نتایج آزمایش‌ها بر روی مدل استاندارد بیماری‌های کبدی و بر مبنای استراتژی اعتبارسنجی متقابل با $k$=5 نشان می‌دهد که مدل پیشنهادی SML بهبود قابل توجهی نسبت به سایر مدل‌های ماشین کسب کرده‌است. مدل SML بر روی مجموعه‌دادهٔ آزمایشی به‌ترتیب مقادیر ۰.۹۹۴۰، ۰.۹۸۸۰، ۰.۰۳ و ۰.۱۷۸۹ را بر معیارهای صحت، معیار F1، میانگین قدرمطلق خطا (MAE) و جذر میانگین مربعات خطا (RMSE) کسب کرده که بسیار بهتر از کارایی کسب‌شده به‌وسیلهٔ سایر مدل‌ها در ادبیات موضوع است. این موضوع نشان می‌دهد که مدل پیشنهادی می‌تواند به‌عنوان ابزاری مؤثر در تشخیص زودهنگام بیماری‌های کبدی مورداستفاده قرار گیرد.

ساختار ادامهٔ مقاله به‌صورت زیر سازمان‌دهی شده‌است. در بخش دوم کارهای مرتبط در زمینهٔ تشخیص و پیش‌بینی بیماری‌های کبدی بررسی، در بخش سوم به جزئیات روش پیشنهادی پرداخته، ارزیابی روش پیشنهادی به همراه نتایج آزمایش‌ها و بحث بر روی نتایج در بخش چهار و نتیجه‌گیری مقاله و نیز ارائه پیشنهادهایی برای انجام پژوهش‌های بیشتر در بخش پنجم آورده شده‌است.

## ۲- کارهای مرتبط

با توجه به این‌که بسیاری از آزمایش‌های پزشکی سنتی به لحاظ هزینه و زمان به‌صرفه نیستند و برای ارائهٔ تشخیص سریع و کم‌هزینه با چالش‌هایی مواجه‌اند، نیاز فوری و مبرمی به روش‌های تشخیصی جایگزین احساس می‌شود. متخصصان بهداشت از روش‌های آسیب‌شناسی مختلفی برای تفسیر گزارش‌های پزشکی و درنتیجه ارزیابی مؤثر وضعیت بیماران استفاده می‌کنند. پیشرفت‌های اخیر در هوش مصنوعی و توسعهٔ روش‌های یادگیری ماشین، تشخیص بیماری‌ها را بهبود داده و به پزشکان اجازه می‌دهد یافته‌های بالینی را با روش‌های پیشرفتهٔ تحلیل داده ادغام کنند [۱۲,۱۴]. در ادامه به بررسی مهم‌ترین روش‌های یادگیری ماشین در حوزهٔ پیش‌بینی بیماری‌های کبدی می پردازیم.

عارف و همکاران [۱۵] کاربرد الگوریتم‌های یادگیری ماشین را برای شناسایی الگوها و ارائهٔ پیش‌بینی‌های ارزشمند جهت تسهیل فرایندهای بالینی تصمیم‌گیری،بررسی کردند که مشتمل بر روش غیرتهاجمی و کم‌هزینه به ویژه برای بیمارانی است که نیاز به پایش دوره‌ای دارند؛ بر اساس نتایج آن‌ها مدل رگرسیون ترابری، مؤثرترین طبقه‌بندی برای پیش‌بینی پیشرفت هپاتیت C معرفی شده‌است. ترلسن و همکاران [۱۶] با به‌کارگیری روش‌های یادگیری ماشین، تحلیلی از ساختار داده‌ها، به‌منظور پاسخ اولیه به پرسش‌های بالینی و تحلیل داده‌های آزمایشگاهی ارائه کردند. سینگ و مانتری [۱۷] یک سازوکار پشتیبانی تصمیم‌گیری بالینی را مورد مطالعه قرار داده‌است که از روش‌های داده‌کاوی برای پایش و ارائهٔ خدمات بهداشتی استفاده می‌کند. آن‌ها با بررسی اطلاعات بیماری‌هایی مانند هپاتیت، مشکلات پوستی، بیماری‌های کبدی و اوتیسم سعی در افزایش دقت اثربخشی تصمیم‌گیری‌های بالینی داشتند.

چالش عدم تعادل داده‌ها، موانع قابل توجهی را در یادگیری ماشین، به‌ویژه در حوزه‌های مرتبط با سلامت مانند پژوهش‌ها بر روی بیماری‌های کبدی و ویروس هپاتیت C ایجاد می‌کند. یک مجموعه‌دادهٔ نامتعادل که با توزیع نابرابر طبقه‌ها مشخص می‌شود، تمایل به تولید مدل‌های پیش‌بینی‌کنندهٔ جانب‌دار به نفع طبقهٔ بیشینه را دارد که می‌تواند بر عملکرد مدل تأثیر منفی بگذارد. برای مقابله با این مشکل می‌توان از روش بیش‌نمونه‌گیری اقلیت مصنوعی[1] (SMOTE) استفاده کرد، که نمونه‌های مصنوعی از طبقهٔ کمینه را برای افزایش نمایش آن به‌منظور تسهیل تولید و یادگیری الگوهای کلی به‌کار می‌برد[۱۸]. لی و همکاران [۱۹]، یک روش طبقه‌بندی چندطبقه را ارائه کردند که با تحلیل ویژگی‌های خون بیماران احتمال بروز عفونت ویروس هپاتیت C[2] را به‌دست می‌آورد؛ همچنین با استفاده از روش بیش‌نمونه‌گیری، اقلیت مصنوعی و روش آبشاری دومرحله‌ای[3]، مدل الگوریتم‌های جنگل تصادفی[4] و رگرسیون ترابری[5] را با هم ادغام کردند و به‌وسیلهٔ الگوریتم کلونی مصنوعی زنبورعسل[6] آستانه ترکیب مؤثر مدل را بهینه‌سازی کردند. ارشادی و سیفی [۲۰] با بررسی روش‌های انتخاب، کاهش و خوشه‌بندی ویژگی‌ها و با معرفی یک روش انتخاب ویژگی به دنبال افزایش کارایی مدل‌های طبقه‌بندی در تشخیص پزشکی بودند؛ درضمن روش پیشنهادی خود را با الگوریتم ژنتیک[7] و بهینه‌سازی ازدحام ذرات[8] در جهت افزایش دقت طبقه‌بندی ترکیب کردند؛ همچنین ترکیب مدل‌های یادگیری ماشین با الگوریتم‌های بهینه‌سازی گلوگاهی دیپر[9] و بهینه‌سازی ازدحام ذرات [۲۱]، تبرید شبیه‌سازی‌شده[10] [۲۲]، برای افزایش دقت و کارایی الگوریتم‌ها در تشخیص و پیش‌بینی کارسینوم سلول کبدی [۲۱، ۹] مورد توجه پژوهش‌گران قرار گرفته است.

یاغان اوغلو [۲۳] از فرایندهای داده‌کاوی برای کشف الگوها و تخمین شیوع ویروس هپاتیت C استفاده کرده‌است. در این پژوهش با استفاده از تجزیه‌وتحلیل جامع و تجسم داده‌ها، ویژگی‌های کلیدی مرتبط با این بیماری شناسایی شده‌است. در مرجع [۲۰] الگوریتم‌های آدابوست و جنگل

[1] Synthetic minority oversampling technique
[2] Hepatitis C virus
[3] Cascade two-stage method
[4] Random forest
[5] Logistic regression
[6] Artificial bee colony
[7] Genetic algorithm
[8] Particle swarm optimization
[9] Dipper throated optimization
[10] Simulated annealing




[ DOI: 10.61882/jsdp.22.2.79 ]

تصادفی را برای پیش‌بینی بیماری‌های کبدی و عفونت ویروس هپاتیت C پیاده‌سازی و از الگوریتم جنگل تصادفی به‌عنوان یک یادگیرندهٔ ضعیف برای افزایش پایداری و دقت و درعین‌حال به کمترین حد رساندن بیش‌برازش استفاده شد. در مراجع [۱۲و۲۴] با هدف شناسایی عوامل مهم تشخیص بیماری کبد، الگوریتم‌های یادگیری ماشین را بر روی مجموعه‌ای از داده‌های جمع‌آوری‌شده از ۶۱۵ فرد مورد مطالعه قرار دادند. آن‌ها از تجسم داده‌ها و روش‌هایی مانند انتساب چندگانه به‌وسیلهٔ معادلات زنجیره‌ای برای مواجهه با داده‌های مفقود استفاده کردند و تحلیل مؤلفه‌های اصلی را برای کاهش ابعاد به‌کار بردند؛ همچنین رتبه‌بندی متغیرهای مهم حاصل از اندیس جینی را به‌وسیلهٔ تحلیل مؤلفه‌های اصلی مورد آزمایش قرار داده و الگوریتم‌های طبقه‌بندی دودویی از جمله شبکه‌های عصبی مصنوعی، جنگل‌های تصادفی و ماشین‌های بردار پشتیبان را برای تمایز بین نمونه‌های خون مبتلا به هپاتیت، فیبروز و سیروز به‌کار بردند.

هافمن و همکاران [۲۵] دو الگوریتم یادگیری ماشین مبتنی بر درخت تصمیم به همراه معیار درجهٔ فیبروز کبدی پیشرفته[1]، که تولید خودکار درختان تصمیم‌گیری را با استفاده از داده‌های آزمایشگاهی تسهیل می‌کنند، مورد بررسی قرار دادند. چیکو و ژورمن [۲۶] سوابق الکترونیکی سلامت[2] متعلق به ۵۴۰ فرد سالم و ۷۵ بیمار مبتلا به هپاتیت C را برای بررسی قابلیت طبقه‌بندهای یادگیری ماشین، به‌ویژه جنگل تصادفی، در پیش‌بینی تشخیص را در نظر گرفتند. آن‌ها با شناسایی متغیرهای حیاتی نظیر آمینوترانسفراز[3] (AST) و آلانین آمینوترانسفراز[4] (ALT) با به‌کاربستن یک گروه اعتبارسنجی شامل ۱۲۳ بیمار مبتلا به هپاتیت C و سیروز، اهمیت این آنزیم‌ها را مورد تأیید قرار داده و با ارزیابی‌های به‌عمل‌آمده، نشان دادند که مدل یادگیری گروهی، از نسبت سنتی AST/ALT عملکرد بهتری دارد؛ و درنتیجه اطلاعات بالینی را بهبود بخشیده و مراقبت از بیمار را ارتقا می‌دهد.

در مرجع [۳]، یک روش پیش‌پردازش داده‌های تطبیقی ارائه شده‌است؛ به‌طوری که عملکرد مدل‌های بنیادی یادگیری ماشین را از طریق روش‌هایی مانند پرکردن میانگین اختصاصی طبقه[5]، حذف داده‌های پرت[6]، نرمال‌سازی لگاریتمی[7]، انتخاب ویژگی[8]، مقیاس‌بندی ویژگی[9] و تعادل داده‌ها[10] بهبود می‌بخشد؛ همچنین آن‌ها برای تسهیل کاربرد عملی، یک رابط کاربری به‌منظور دریافت و پردازش داده‌های بیماران کبدی طراحی کردند. اروجی و کرمانی [۲۷] از مجموعه‌داده‌های هپاتیت C موجود در مخزن UCI[11] استفاده کرده و از روش چهارمرحله‌ای شامل پیش‌پردازش داده‌ها، داده‌کاوی، طراحی و تحلیل سامانه با ارائهٔ یک محیط کاربری در نرم‌افزار MATLAB استفاده کردند. آن‌ها با به‌کاربستن سه الگوریتم داده‌کاوی پرسپترون چندلایه، شبکهٔ بیزی و درخت تصمیم از روش اعتبارسنجی متقاطع دَه‌تایی به ارزیابی عملکرد پرداختند. فرغالی و همکاران [۲۸] یادگیری ماشین را برای پیش‌بینی ویروس هپاتیت C در بین کارکنان مراقبت‌های بهداشتی در مصر شامل ۸۵۹ بیمار و دوازده ویژگی استفاده کرده و دو رویکرد بدون انتخاب ویژگی و انتخاب روبه‌جلوی متوالی[12] را برای اصلاح مجموعه‌ویژگی‌ها مورد استفاده قرار دادند. علیزرگر و همکاران [۲۹] الگوریتم‌های یادگیری ماشین را روی مجموعه‌داده‌های موجود در NHANES [13] و UCI بررسی کرده و الگوریتم‌های بردار ماشین پشتیبان و تقویت گرادیان شدید[14] را به‌عنوان روش‌های پیش‌بینی بیماری‌های کبدی و هپاتیت C از روی داده‌های آزمایش خون پیشنهاد داده‌اند. بهاراتی و همکاران [۱۳] از مدل‌های طبقه‌بندی ماشین بردار پشتیبان، درخت تصمیم، رگرسیون ترابری و جنگل تصادفی با استفاده از انتخاب ویژگی و هم بدون آن برای پیش‌بینی ویروس هپاتیت C بهره گرفته‌اند.

در جدول (۱) کارایی مدل‌های یادگیری ماشین که برای پیش‌بینی بیماری‌های کبدی مورداستفاده قرار گرفته‌اند، ارائه شده‌است. در این جدول بهترین کارایی مدل‌ها گزارش شده‌است. توجه به این نکته ضروری است که برخی از این روش‌ها، همانند مدل مطرح‌شده در مرجع [۳]، نتایج را تنها بر روی مجموعه‌دادهٔ آموزشی بیان کرده و از ارزیابی مدل‌ها بر روی مجموعه‌دادهٔ آزمایشی خودداری کرده‌اند. بدیهی است که نتایج بر روی مجموعه‌داده‌های آزمایشی کمتر از مقادیر بیان‌شده در جدول (۱) خواهد بود. همان‌گونه که در این جدول مشاهده می‌شود، هنوز کارایی مدل‌ها از حالت ایده‌آل فاصله دارد و تلاش زیادی برای بهبود نتایج مورد نیاز است.

## ۳- مواد و روش‌ها

### ۳-۱-مجموعه‌داده

جدول (۲) مشخصات آماری مجموعه‌دادهٔ مورد استفاده در این پژوهش را نشان می‌دهد. این مجموعه‌داده شامل ۶۱۵ رکورد است که در آن هر رکورد شامل دوازده ویژگی است که عبارت‌اند از: سن، جنسیت، آلبومین[15]،

[1] Enhanced liver fibrosis (ELF) score
[2] Electronic health records
[3] Aspartate aminotransferase
[4] Alanine aminotransferase
[5] Class specific mean imputation
[6] Outlier rejection
[7] Log normalization
[8] Feature selection
[9] Feature scaling
[10] Data balancing

[11] UCI machine learning repository
[12] Sequential forward selection
[13] National health and nutrition examination Survey
[14] Extreme gradient boosting (XGB)
[15] Albumin (ALB)




[ DOI: 10.61882/jsdp.22.2.79 ]

**(جدول-۱): کارایی گزارش‌شده به‌وسیلهٔ روش‌های یادگیری ماشین برای پیش‌بینی بیماری‌های کبدی**

**(Table-1): Reported performance of machine learning methods for predicting liver diseases**

| سال | مرجع | مدل‌های مورداستفاده | مجموعه‌داده | صحت |
|---|---|---|---|---|
| ۲۰۲۱ | [۲۴] | ANN, RF, SVM | ۶۱۵ بیمار با ۱۴ ویژگی | RF: 0.9814 |
| ۲۰۲۱ | [۱۲] | ANN, RF, SVM | ۶۱۵ بیمار با ۱۴ ویژگی | SVM: 0.9823 |
| ۲۰۲۱ | [۲۷] | MLP, RF, J48 | ۶۱۵ بیمار با ۱۱ ویژگی | RF: 0.9990 |
| ۲۰۲۱ | [۲۶] | RF, DT, LR | ۶۱۵ بیمار با ۱۱ ویژگی | RF: 0.971 |
| ۲۰۲۱ | [۳۰] | SVM, NB, DT, RF, LR, KNN | ۶۱۵ بیمار با ۱۲ ویژگی | RF: 0.9729 |
| ۲۰۲۲ | [۲۳] | LR, KNN, DT, RF, SVM, ETC, ADA, NB | ۶۱۵ بیمار با ۱۴ ویژگی | DT: 0.9931 |
| ۲۰۲۲ | [۱۵] | LR, RF, RPART, CTREE, XGB | ۶۱۵ بیمار با ۱۱ ویژگی | LR: 0.948 |
| ۲۰۲۳ | [۲۹] | KNN, DT, ANN, SVM, LR, XGB | ۶۱۵ بیمار با ۱۲ ویژگی | SVM, XGB: 0.95 |
| ۲۰۲۳ | [۲۸] | NB, RF, KNN, LR | ۸۵۹ بیمار با ۱۲ ویژگی | RF: 0.9488 |
| ۲۰۲۴ | [۱۳] | DT, SVM, LR, RF | ۹۹۷ بیمار با ۱۰ ویژگی | RF: 0.89 |
| ۲۰۲۴ | [۱۷] | RST, KNN, LSVM, RBF SVM, DT, RF, NB | ۱۵۵ بیمار با ۲۰ ویژگی | RF: 0.8866 |
| ۲۰۲۴ | [۳] | LR, SVM, RF, DT, XB, Ensemble LR, Ensemble XB, Ensemble RF | UCI<br>۶۱۵ بیمار با ۱۲ ویژگی | Ensemble LR: 0.9984 |

**(جدول-۲): مشخصات مجموعه‌دادهٔ مورد ارزیابی**

**(Table-2): Characteristics of the data set being evaluated**

| معیار | رده | سن | جنسیت | ALB | ALP | ALT | AST | BIL | CHE | CHOL | CREA | GGT | PROT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| تعداد | ۶۱۵ | ۶۱۵ | ۶۱۵ | ۶۱۴ | ۵۹۷ | ۶۱۴ | ۶۱۵ | ۶۱۵ | ۶۱۵ | ۶۰۵ | ۶۱۵ | ۶۱۵ | ۶۱۴ |
| میانگین | - | ۴۷.۴۱ | ۰.۳۹ | ۴۱.۶۲ | ۶۸.۲۸ | ۲۸.۴۵ | ۳۴.۷۹ | ۱۱.۴۰ | ۸.۲۰ | ۵.۳۷ | ۸۱.۲۹ | ۳۹.۵۳ | ۷۲.۰۴ |
| انحراف معیار | - | ۱۰.۰۶ | ۰.۴۹ | ۵.۷۸ | ۲۶.۰۳ | ۲۵.۴۷ | ۳۳.۰۹ | ۱۹.۶۷ | ۲.۲۱ | ۱.۱۳ | ۴۹.۷۶ | ۵۴.۶۶ | ۵.۴۰ |
| کمینه | ۱ | ۱۹ | ۰ | ۱۴.۹ | ۱۱.۳ | ۰.۹ | ۱۰.۶ | ۰.۸ | ۱.۴۲ | ۱.۴۳ | ۸ | ۴.۵ | ۴۴.۸ |
| بیشینه | ۵ | ۷۷ | ۱ | ۸۲.۲ | ۴۱۶.۶ | ۳۲۵.۳ | ۳۲۴ | ۲۵۴ | ۱۶.۴۱ | ۹.۶۷ | ۱۰۷۹.۱ | ۶۵۰.۹ | ۹۰ |

آلکالین آمینوترانسفراز[1]، آسپارتات آمینوترانسفراز[2]، بیلی روبین[3]، کولین استراز[4]، کلسترول[5]، کراتینین[6]، گاماگلوتامین ترانسفراز[7] و پروتئین[8].

داده‌ها به پنج طبقهٔ اصلی تقسیم شده‌اند: اهداکنندهٔ خون[9]، اهداکنندهٔ خون مشکوک[10]، هپاتیت[11]، فیبروز[12] و سیروز[13]. مجموعه‌دادهٔ مورد استفاده در نشانی زیر قابل دسترسی است:

https://archive.ics.uci.edu/dataset/571/hcv+data

جدول (۳) توزیع داده‌ها در طبقه‌های مجموعه‌داده را نشان می‌دهد؛ همان‌گونه که در جدول مشاهده می‌شود، در مجموعه‌دادهٔ اصلی تعداد نمونه‌ها در هر طبقه برابر نیست و این عدم توازن می‌تواند بر تفسیر دقت مدل تأثیر بگذارد.

[1] Alanine aminotransferase (ALT)
[2] Aspartate aminotransferase (AST)
[3] Bilirubin (BIL)
[4] Cholinesterase (CHE)
[5] Cholesterol (CHOL)
[6] Creatinine (CREA)
[7] Gamma-glutamyl transferase (GGT)
[8] Protein (PROT)
[9] Blood donor
[10] Suspect blood donor
[11] Hepatitis
[12] Fibrosis
[13] Cirrhosis

شکل (۱) نمودار پراکندگی ماتریسی را برای مجموعه‌دادهٔ مورد بحث نشان می‌دهد که با استفاده از کتابخانهٔ Seaborn در پایتون ایجاد شده‌است. این نمودار هم‌بستگی بین متغیرها در مجموعه‌داده را نشان می‌دهد. درایه‌های قطر اصلی، توزیع مقادیر هر متغیر را به صورت هیستوگرام نشان می‌دهد. درایه‌های زیر قطر اصلی نمودار پراکندگی جفت متغیرها را نشان می‌دهد. خط قرمز رنگ، خط رگرسیون خطی است که جهت‌گیری ارتباط بین متغیرها را نشان می‌دهد. درایه‌های بالای قطر اصلی، ضرایب هم‌بستگی پیرسون بین جفت متغیرها را نشان می‌دهد. رنگ‌ها شدت و جهت هم‌بستگی را نشان می‌دهند. ماتریس نشان می‌دهد که بیشترین هم‌بستگی مثبت در بین دو متغیر Category و AST با مقدار ۰.۶۵ است؛ در مقابل، بیشترین همبستگی منفی بین دو متغیر ALB و Category با مقدار ۰.۲۹- است.

## ۳-۲-پیش‌پردازش

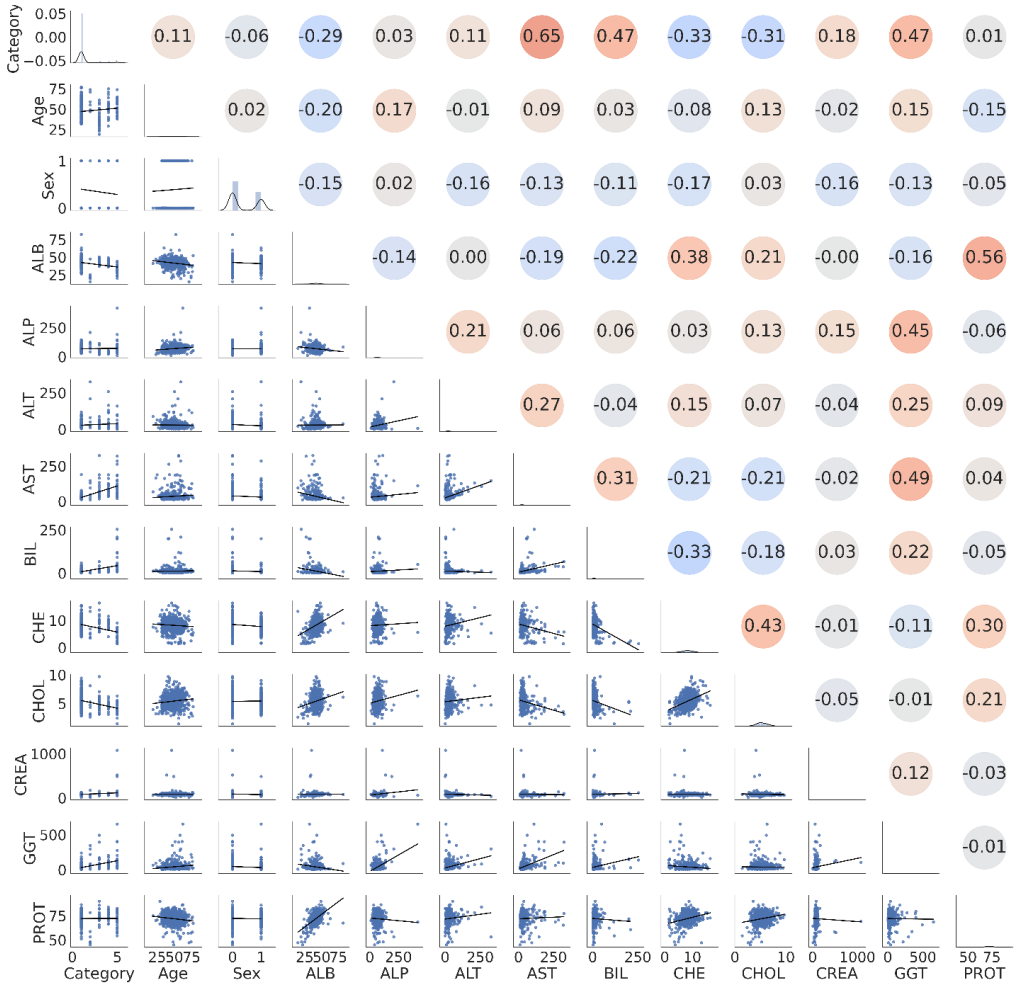
در مرحلهٔ پیش‌پردازش ابتدا به بررسی مقادیر میانگین و انحراف معیار ویژگی‌ها پرداخته شده‌است؛ همان گونه که در جدول (۲) مشاهده می‌شود، این مقادیر با یکدیگر اختلاف زیادی دارند. برای حل این مسئله از ماژول مقیاس‌بندی




[ DOI: 10.61882/jsdp.22.2.79 ]

استاندارد در پایتون استفاده شده‌است تا مقادیر رکوردهای مجموعه‌داده به طور مناسب پردازش شوند و تمام ویژگی‌های عددی در مقیاس قابل مقایسه قرار گیرند. این کار باعث می‌شود که هر ویژگی به طور مساوی در فرایند یادگیری مدل‌های یادگیری ماشین نقش داشته باشد.

مطابق جدول (۳)، در مجموعه‌دادهٔ مورد استفاده تعداد نمونه‌ها در هر طبقه برابر نیست و این عدم تعادل می‌تواند بر تفسیر دقت مدل تأثیر بگذارد؛ به همین دلیل، از روش SMOTE استفاده شده‌است. این روش یکی از رهیافت‌های کارآمد برای مقابله با مشکل عدم توازن در طبقه‌ها است. در بسیاری از مسائل یادگیری ماشین، به‌ویژه در حوزه‌های پزشکی، تعداد نمونه‌های طبقهٔ کمینه به‌طور قابل توجهی کمتر از طبقهٔ بیشینه است. عدم توازن می‌تواند منجر به عملکرد ضعیف مدل‌ها شود؛ زیرا مدل‌های یادگیری ماشین بیشتر تمایل دارند بر روی داده‌های طبقهٔ بیشینه تمرکز کنند و درنتیجه دقت پیش‌بینی برای طبقهٔ کمینه کاهش می‌یابد.




**(شکل-۱): ماتریس پراکندگی متغیرها در مجموعه‌داده به همراه ضرائب هم‌بستگی**

**(Figure-1): The dispersion matrix of variables in the data set along with correlation coefficients.of the pre-processing phase on a sample web page**

روش SMOTE به‌طور خاص با هدف افزایش تعداد نمونه‌های طبقهٔ کمینه و بهبود توازن بین طبقه‌ها طراحی شده‌است؛ این روش به جای رونوشت برداری ساده از نمونه‌های موجود در طبقهٔ کمینه، نمونه‌های جدیدی را به‌صورت مصنوعی تولید می‌کند. در الگوریتم SMOTE، ابتدا برای هر نمونهٔ موجود در طبقهٔ کمینه، $k$ همسایهٔ نزدیک، با استفاده از الگوریتم KNN شناسایی می‌شود؛ سپس برای تولید یک نمونهٔ جدید، یک همسایهٔ تصادفی از بین $k$ همسایه انتخاب می‌شود و یک نقطهٔ جدید در فضای ویژگی با استفاده از ترکیب خطی بین نمونهٔ اصلی و همسایهٔ انتخاب‌شده ایجاد می‌شود. این فرایند به تولید نمونه‌های جدیدی منجر می‌شود که در فضای ویژگی در نزدیکی نمونه‌های موجود قرار دارند، اما به اندازهٔ کافی متفاوت‌اند تا تنوع داده‌ها را افزایش دهند؛ همچنین داده‌های گم‌شده می‌توانند به‌دلایل مختلفی مانند خطا در جمع‌آوری داده‌ها یا عدم پاسخ‌گویی در نظرسنجی‌ها ایجاد شوند و وجود آن‌ها می‌تواند تأثیر قابل توجهی بر عملکرد مدل‌های یادگیری ماشین بگذارد. برای حل مشکل داده‌های گم‌شده از روش جایگزینی میانه[1] استفاده شده‌است که در آن مقادیر گم‌شدهٔ هر حوزه با میانهٔ مقادیر موجود

[1] Median



[ DOI: 10.61882/jsdp.22.2.79 ]

همان حوزه جایگزین می‌شود. روش میانه در مقابل داده‌های پرت[1] مقاوم است؛ بنابراین انتخاب مناسبی برای جایگزینی مقادیر گم‌شده محسوب می‌شود.

**(جدول-۳): توزیع داده‌ها در طبقه‌های مجموعه‌داده**
**(Table-3): Distribution of data points in the classes of benchmark dataset**

| نام طبقه | تعداد |
|---|---|
| اهداکنندهٔ خون | ۵۳۳ |
| اهداکنندهٔ خون مشکوک | ۷ |
| هپاتیت | ۲۴ |
| فیبروز | ۲۱ |
| سیروز | ۳۰ |

## ۳-۳-روش‌های یادگیری ماشین

### ۳-۳-۱-رگرسیون ترابری

رگرسیون ترابری[2] (LR) یک روش یادگیری نظارت‌شده است؛ این روش برخلاف نامش برای مسائل طبقه‌بندی دودویی استفاده می‌شود که در آن هدف پیش‌بینی احتمال تعلق یک ورودی داده‌شده به یک طبقهٔ خاص است. روش رگرسیون ترابری به‌دلیل سادگی و اثربخشی در زمینه‌های مختلفی مانند مالی، بهداشت و درمان، بازاریابی و علوم اجتماعی کاربرد فراوانی دارد. در این روش متغیر وابسته دوقطبی است و متغیر مستقل می‌تواند دو جمله‌ای، ترتیبی، بازه‌ای یا سطح نسبت باشد. رگرسیون ترابری ابتدا برای مسائلی با دو نتیجهٔ ممکن (برای مثال، بله/خیر، موفقیت/شکست) طراحی شده‌است؛ بااین‌حال می‌توان آن را به مسائل طبقه‌بندی چندطبقه نیز تعمیم داد. این مدل از تابع ترابری که به‌عنوان تابع سیگموئید نیز شناخته می‌شود، برای تبدیل خطی ترکیب ورودی‌ها به مقادیر احتمال استفاده می‌کند.

### ۳-۳-۲-تقویت گرادیان شدید

روش تقویت گرادیان شدید یک الگوریتم مقیاس‌پذیر و بهینه در یادگیری ماشین است که سرعت و عملکرد ماشین‌های تقویت گرادیان را بهبود بخشیده و در مسائل طبقه‌بندی و رتبه‌بندی کاربرد دارد. این روش از یک الگوریتم یادگیری مبتنی بر درخت تصمیم استفاده می‌کند و عملکرد مدل را با استفاده از محاسبات موازی و توزیع‌شده تسریع می‌کند. الگوریتم تقویت گرادیان شدید به‌طور مکرر درخت‌های تصمیم را اضافه می‌کند و هر درخت برای تصحیح خطاهای درختان قبلی آموزش داده می‌شود. این فرایند از گرادیان‌های تابع هزینه برای به‌کمینه‌رساندن خطاهای پیش‌بینی استفاده می‌کند و از روش‌های منظم‌سازی مانند لاسو و ریج برای جلوگیری از برازش بیش از حد استفاده می‌کند. پس از آموزش، این روش از مجموعه‌ای از درختان برای پیش‌بینی هدف در مجموعه‌دادهٔ آزمایشی استفاده می‌کند و پیش‌بینی نهایی به‌عنوان مجموع وزنی خروجی‌های همهٔ درختان محاسبه می‌شود.

تفاوت اصلی بین روش تقویت گرادیان و تقویت گرادیان شدید در پیاده‌سازی و روش‌های بهینه‌سازی آن‌ها نهفته است؛ درحالی‌که هر دو روش از مجموعه‌ای از درختان تصمیم برای بهبود دقت پیش‌بینی استفاده می‌کنند. روش تقویت گرادیان شدید به‌گونه‌ای طراحی شده‌است که کارآمدتر و مقیاس‌پذیرتر باشد. این روش ویژگی‌های پیشرفته‌ای مانند منظم‌سازی برای جلوگیری از برازش بیش از حد، پردازش موازی برای محاسبات سریع‌تر و مدیریت پیچیده‌تر مقادیر ازدست‌رفته را شامل می‌شود.

### ۳-۳-۳-طبقه‌بندی جنگل تصادفی

روش طبقه‌بندی جنگل تصادفی یک روش یادگیری مجموعه‌ای است که درخت‌های تصمیم‌گیری متعددی را در طول آموزش می‌سازد و از نتایج آن‌ها برای بهبود دقت پیش‌بینی استفاده می‌کند. در این روش از روش‌های بوت‌استرپینگ برای تولید درخت استفاده می‌شود که داده‌ها را دوباره نمونه‌برداری می‌کند تا تنوعی در بین درختان ایجاد کند و خطر بیش از حد برازش را کاهش دهد. روش جنگل تصادفی تفسیرهایی را در مورد اهمیت ویژگی‌هایی ارائه می‌دهد که تفسیرپذیری مدل‌های پیچیده را افزایش می‌دهد. ساختار آن نیز موازی‌پذیر است که کاربرد آن را برای مجموعه‌های داده بزرگ تسهیل می‌کند[۳۱]. درخت تصمیم یک درخت منفرد است که بر اساس تقسیم ویژگی‌ها تصمیم می‌گیرد؛ درحالی‌که جنگل تصادفی مجموعه‌ای از درختان تصمیم‌گیری است که هم‌زمان کار می‌کنند. درختان تصمیم‌گیری به‌ویژه با مجموعه‌داده‌های پیچیده مستعد بیش‌ازحد برازش هستند، اما جنگل‌های تصادفی این مشکل را با میانگین‌گیری نتایج چند درخت کاهش می‌دهند که منجر به تعمیم بهتر می‌شود؛ همچنین، روش جنگل‌های تصادفی با انتخاب تصادفی ویژگی‌ها برای تقسیم در هر گره به کاهش واریانس و بهبود تنوع مدل کمک می‌کند که این امر بهبود عملکرد کلی مدل را به همراه دارد.

### ۳-۳-۴-درختان بی‌نهایت تصادفی

درختان بی‌نهایت تصادفی[3] یک روش یادگیری ماشین برای دسته‌بندی است که از درخت‌های تصمیم‌گیری استفاده می‌کند و به‌عنوان یک جایگزین سریع‌تر برای روش جنگل تصادفی شناخته می‌شود. این الگوریتم مشابه جنگل تصادفی، چندین درخت تصمیم‌گیری ایجاد و پیش‌بینی‌های آن‌ها را ترکیب می‌کند. یکی از ویژگی‌های کلیدی این الگوریتم،

[1] Outlier
[2] Logistic Regression
[3] Extra trees classifier




[ DOI: 10.61882/jsdp.22.2.79 ]

انتخاب تصادفی ویژگی‌ها و نقاط تقسیم است. در مقایسه با جنگل تصادفی، درختان بی‌نهایت تصادفی به‌جای انتخاب بهترین تقسیم از یک زیرمجموعهٔ تصادفی از ویژگی‌ها، ویژگی‌ها و نقاط تقسیم را به‌طور کامل تصادفی انتخاب می‌کنند. این رویکرد به افزایش سرعت الگوریتم کمک می‌کند و زمان آموزش را کاهش می‌دهد؛ علاوه‌براین تصادفی‌بودن این روش به کاهش بیش‌برازش کمک می‌کند و عملکرد مدل را بر روی داده‌های نادیده بهبود می‌بخشد، به‌ویژه در مقایسه با مدل‌هایی که از ویژگی‌های تصادفی استفاده نمی‌کنند؛ این روش با استفاده از نمونه‌برداری تصادفی با جایگزینی، چندین زیرمجموعه از داده‌ها را ایجاد می‌کند و هر یک از این زیرمجموعه‌ها به‌طور مستقل برای آموزش تعدادی از درختان استفاده می‌شود. درختان بی‌نهایت تصادفی دارای هایپرپارامترهایی هستند که می‌توانند بر عملکرد مدل تأثیر بگذارند، مانند تعداد درختان و تعداد ویژگی‌های تصادفی که در هر تقسیم استفاده می‌شود؛ درنهایت در این روش پیش‌بینی نهایی با میانگین‌گیری پیش‌بینی‌ها از تمامی درختان به‌دست می‌آید؛ درکل درختان بی‌نهایت تصادفی به‌دلیل سرعت و کارایی بالای خود ابزاری قدرتمند در یادگیری دسته‌ای شناخته می‌شوند و می‌توانند در مسائل مختلف دسته‌بندی به‌کار گرفته شوند؛ همچنین درختان بی‌نهایت تصادفی به‌دلیل ساختار درختی خود می‌توانند به‌راحتی تفسیر شوند و به کاربران اجازه می‌دهند تا درک بهتری از فرایند تصمیم‌گیری داشته باشند.

### ۳-۳-۵-دسته‌بندی درخت تصمیم

دسته‌بندی درخت تصمیم[1] یک مدل دسته‌بندی غیرخطی است که به‌طور گسترده‌ای در بسیاری از کاربردها مورد استفاده قرار می‌گیرد. این روش با تقسیم داده‌ها به زیرمجموعه‌های کوچک‌تر بر اساس ویژگی‌های خاص، ساختاری درختی ایجاد می‌کند. هر گره درخت نمایان‌گر یک ویژگی و هر شاخه نشان‌دهندهٔ یک تصمیم یا نتیجه است که درنهایت به گره‌های برگ ختم می‌شود. در این مدل، مجموعه‌دادهٔ ورودی به‌طور مداوم به زیرمجموعه‌های کوچک‌تر بر اساس مقادیر ویژگی‌ها با استفاده از تقسیمات دوتایی بازگشتی تقسیم می‌شود. گره‌های داخلی آزمایشی‌های ویژگی را توصیف می‌کنند و گره‌های برگ نتایج پیش‌بینی‌شده را بر اساس بیشترین طبقه‌ها یا مقادیر انباشته‌شده از نمونه‌های آموزشی موجود در آن زیرمجموعه‌ها نشان می‌دهند.

معیارهای مختلفی مانند آنتروپی و شاخص جینی در انتخاب ویژگی‌ها و تقسیم داده‌ها مورد استفاده قرار می‌گیرند.

[1] Decision tree classifier

آنتروپی به عدم قطعیت موجود در یک مجموعه‌داده اشاره دارد و هدف آن کاهش آنتروپی از طریق تقسیم داده‌ها است؛ از سوی دیگر، شاخص جینی به‌احتمال اشتباه در پیش‌بینی طبقه‌ها اشاره داشته و هدف آن کمینه‌کردن این احتمال است. به‌دلیل ساختار درختی مدل درخت تصمیم، تفسیر این مدل‌ها آسان است و کاربران قادر خواهند بود مسیرهای تصمیم‌گیری را مشاهده کنند.

یکی از مزایای قابل‌توجه دسته‌بندی درخت تصمیم، توانایی آن در شناسایی روابط پیچیدهٔ درون داده‌ها بدون نیاز به مراحل پیش‌پردازش یا تبدیل گسترده است. درختان تصمیم می‌توانند برای داده‌های دسته‌ای و عددی به‌کار گرفته شوند و همچنین می‌توانند روابط بین ویژگی‌ها را به‌طور کارآمد مدل‌سازی کنند.

### ۳-۳-۶-روش یادگیری انباشته

شکل (۲) ساختار روش یادگیری انباشته (SML) را نشان می‌دهد. هدف از ارائهٔ این مدل بهبود دقت پیش‌بینی با ترکیب پیش‌بینی‌های چندین مدل یادگیری متنوع است. مدل SML بر اساس مفهوم استفاده از یک فرامدل[2] برای ترکیب پیش‌بینی‌های مدل‌های یادگیر پایه[3] استوار است. فرامدل با یادگیری از پیش‌بینی‌های مدل‌های پایه پیش‌بینی نهایی را که از دقت بالایی برخوردار است انجام می‌دهد.

مراحل مدل پیشنهادی SML به‌صورت زیر است:

- **ایجاد مدل‌های پایه:** در این مرحله چهار مدل پایه DTR، XGB، RFC، ETC بر روی داده‌های آموزشی ساخته می‌شوند. این الگوریتم‌ها به‌دلیل کارایی بالای خود در مسائل دسته‌بندی و پیش‌بینی انتخاب شدند. این مدل‌ها به‌خوبی می‌توانند با داده‌های غیرخطی و پیچیده که در بیماری‌های کبدی وجود دارد، سازگار شوند؛ برای مثال درخت تصمیم و جنگل تصادفی به‌خوبی قادر به شناسایی الگوهای پیچیده در داده‌ها هستند و در عین حال قابلیت تفسیر نتایج را نیز فراهم می‌کنند؛ همچنین الگوریتم XGB به‌عنوان یک روش تقویت‌کننده، به طور خاص برای بهبود دقت پیش‌بینی طراحی شده‌است. این الگوریتم با استفاده از روش‌های بهینه‌سازی پیشرفته می‌تواند به‌سرعت یاد بگیرد و دقت بالایی را در پیش‌بینی ارائه دهد.
- **پیش‌بینی‌های مدل‌های پایه:** هر کدام از مدل‌های پایه روی داده‌های آزمایشی اعمال شده و مقدار متغیر هدف را پیش‌بینی می‌کنند.

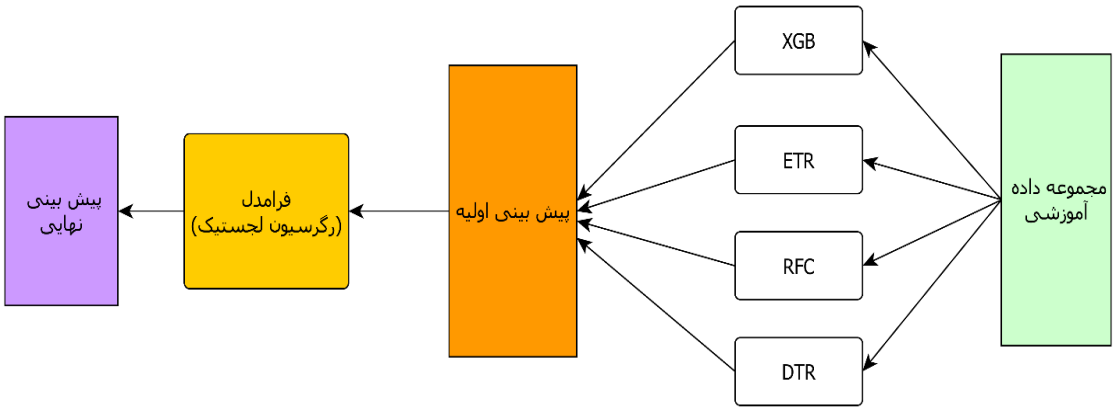
[2] Meta model
[3] Base models




[ DOI: 10.61882/jsdp.22.2.79 ]

– **ایجاد فرامدل:** یک فرامدل روی پیش‌بینی‌های تولیدشده به‌وسیلهٔ مدل‌های پایه ساخته می‌شود. وظیفهٔ فرامدل یادگیری رابطهٔ بین پیش‌بینی‌های مدل‌های پایه و خروجی واقعی است.

**پیش‌بینی نهایی:** فرامدل LR با استفاده از پیش‌بینی‌های مدل‌های پایه پیش‌بینی نهایی را ایجاد می‌کند.




**(شکل-۲): ساختار مدل پیشنهادی SML**

**(Figure-2): Structure of the proposed SML model**

# ۴-نتایج آزمایش‌ها و بحث

## ۴-۱-معیارهای ارزیابی

برای ارزیابی مدل‌ها و تحلیل نتایج از شش معیار کارایی استفاده شده‌است که عبارت‌اند از صحت[۱]، دقت[۲]، بازخوانی[۳]معیار F1، میانگین قدرمطلق خطا[۴] (MAE) و جذر میانگین مربعات خطا[۵] (RMSE) [۳۲].

## ۴-۲-نتایج آماری

جدول‌های (۴) و (۵) نتایج الگوریتم‌های یادگیری ماشین را بر روی مجموعه‌دادهٔ آموزشی و آزمایشی نشان می‌دهد. در تولید نتایج از روش اعتبارسنجی متقابل با k=5 استفاده شده‌است تا از بروز بیش‌برازش در داده‌ها جلوگیری شود. با توجه به نتایج ارائه‌شده در جدول‌های (۴) و (۵) می‌توان موارد زیر را استنباط کرد.

بر حسب معیار صحت بهترین الگوریتم مدل SML است که به‌ترتیب بر روی مجموعه‌داده‌های آزمایشی و آموزش نرخ صحت ۰.۹۹۹۱ و ۰.۹۹۴۰ را کسب کرده‌است. این موضوع نشان می‌دهد که این الگوریتم عملکرد مطلوبی در قیاس با سایر مدل‌ها دارد. الگوریتم ET جایگاه دوم را از نظر عملکرد ثبت کرده و پس از آن الگوریتم‌های DT و XGB در رتبه‌های بعدی قرار دارند. الگوریتم LR اگر چه مقدار صحت قابل قبولی کسب کرده‌است، اما در قیاس با سایر مدل‌ها در رتبهٔ آخر قرار دارد.

بر حسب معیار دقت، مدل پیشنهادی SML، به‌ترتیب با کسب مقادیر ۰.۹۹۸۲ و ۰.۹۸۴۹ بر روی مجموعه‌دادهٔ آموزشی و آزمایش بهترین نتایج را در قیاس با سایر مدل‌ها کسب کرده‌است. مقادیر دقت گزارش‌شده به‌وسیلهٔ بیشتر مدل‌ها قابل قبول و حاکی از عملکرد مطلوب آن‌ها در تفکیک طبقه‌ها است. مقادیر بالای دقت گزارش‌شده نشان می‌دهد که مدل پیشنهادی در پیش‌بینی بیماری‌های کبدی بسیار دقیق عمل می‌کند و احتمال تشخیص اشتباه افراد سالم به‌عنوان بیمار بسیار کم است.

بر حسب معیار بازخوانی، مدل‌های SML، ET و DT به‌ترتیب با نرخ بازخوانی ۰.۹۹۴، ۰.۹۹۸ و ۰.۹۷۹۴ بهترین نتایج را بر روی مجموعه‌دادهٔ آزمایشی و همچنین مدل SML بر روی مجموعه‌دادهٔ آموزشی با امتیاز ۰.۹۹۹۵ بهترین نرخ بازخوانی را کسب کرده‌اند. کسب مقادیر بالای بازخوانی به‌وسیلهٔ مدل پیشنهادی نشان از توانایی خوب آن در شناسایی نمونه‌های مثبت دارد.

بر حسب معیار F1 نیز بهترین نتایج برای مدل SML گزارش شده‌است. بر روی مجموعه‌دادهٔ آزمایشی، رتبهٔ دوم و سوم متعلق به مدل‌های ET و DT است. نتایج نشان می‌دهد که مدل SML توازن مطلوبی بین دقت و بازخوانی برقرار کرده‌است. این موضوع نشان می‌دهد که مدل هم در تشخیص بیماری خوب عمل می‌کند و هم از تشخیص اشتباه افراد سالم به‌عنوان بیمار تا حد امکان جلوگیری می‌کند.

بر حسب معیار MAE، بهترین نتایج به‌وسیلهٔ مدل ET کسب شده‌است که این مقدار برابر ۰.۰۰۳۲ برای مجموعه‌دادهٔ آموزشی و ۰.۰۲۴ برای مجموعه‌دادهٔ‌آزمون است. مدل SML با کمی اختلاف در ردهٔ دوم قرار گرفته است.

[1] Accuracy
[2] Precision
[3] Recall
[4] Mean Absolute Error
[5] Root Mean Squared Error




[ DOI: 10.61882/jsdp.22.2.79 ]

**(جدول-۴): نتایج مدل‌های یادگیری ماشین بر روی مجموعه‌دادهٔ آموزشی**

**(Table-4): Results of machine learning models on the training dataset**

| مدل | صحت | دقت | بازخوانی | F1 | MAE | RMSE |
|---|---|---|---|---|---|---|
| XGB | ۰.۹۹۳۱ | ۰.۹۹۳۲ | ۰.۹۹۳۱ | ۰.۹۹۳۱ | ۰.۰۱۲۵ | ۰.۱۵۶۵ |
| RF | ۰.۹۹۵۴ | ۰.۹۹۴۱ | ۰.۹۹۵۸ | ۰.۹۹۴۰ | ۰.۰۱۵۲ | ۰.۱۰۵۳ |
| LR | ۰.۹۵۵۲ | ۰.۹۵۵۷ | ۰.۹۵۵۱ | ۰.۹۵۴۹ | ۰.۰۵۵۰ | ۰.۲۷۴۴ |
| DTC | ۰.۹۷۶۰ | ۰.۹۷۹۴ | ۰.۹۸۱۵ | ۰.۹۸۲۰ | ۰.۰۳۵۶ | ۰.۲۵۵۲ |
| ETC | ۰.۹۹۸۶ | ۰.۹۹۸۲ | ۰.۹۹۷۷ | ۰.۹۹۸۶ | ۰.۰۰۳۲ | ۰.۰۸۸۶ |
| SML | ۰.۹۹۹۱ | ۰.۹۹۹۱ | ۰.۹۹۹۵ | ۰.۹۹۹۱ | ۰.۰۰۳۷ | ۰.۰۷۷۵ |

**(جدول-۵): نتایج مدل‌های یادگیری ماشین بر روی مجموعه‌دادهٔ آزمایشی**

**(Table-5): Results of machine learning models on the testing dataset**

| مدل | صحت | دقت | بازخوانی | F1 | MAE | RMSE |
|---|---|---|---|---|---|---|
| XGB | ۰.۹۷ | ۰.۹۷۱۶ | ۰.۹۷ | ۰.۹۶۹۸ | ۰.۰۵۴ | ۰.۳۲۵۶ |
| RF | ۰.۹۷۲ | ۰.۹۷۷۱ | ۰.۹۷۴ | ۰.۹۷۳۸ | ۰.۰۳۸ | ۰.۲۴۴۹ |
| LR | ۰.۹۵۴ | ۰.۹۵۶۸ | ۰.۹۵۴ | ۰.۹۵۲۹ | ۰.۰۶۶ | ۰.۳۳۱۷ |
| DT | ۰.۹۷۶ | ۰.۹۷۹۴ | ۰.۹۸۱۵ | ۰.۹۸۲ | ۱.۱۴۴ | ۰.۲۵۵۲ |
| ET | ۰.۹۸۴ | ۰.۹۸۴۹ | ۰.۹۸۸ | ۰.۹۸۵۹ | ۰.۰۲۴ | ۰.۱۸۴۴ |
| SML | ۰.۹۹۴ | ۰.۹۹۴۲ | ۰.۹۹۴ | ۰.۹۸۸ | ۰.۰۳ | ۰.۱۷۸۹ |

جایگاه سوم از منظر معیار MAE به مدل RF تعلق دارد. تحلیل مقادیر MAE نشان می‌دهد که تفاوت بین مشاهدات و پیش‌بینی مدل کمتر بوده و مدل در پیش‌بینی عملکرد خوبی دارد.

بر اساس معیار RMSE، بهترین نتایج متعلق به مدل پیشنهادی SML با مقادیر ۰.۷۵۵ و ۰.۱۷۸۹ به‌ترتیب برای مجموعه‌دادهٔ آموزشی و آزمایشی است. بعد از آن، مدل‌های ET و RF به‌ترتیب در رده‌های بعدی از منظر کارایی قرار دارند. مقادیر پایین RMSE نشان‌دهندهٔ این است که مدل SML به‌طور متوسط دارای خطاهای کمتری در پیش‌بینی‌هاست و پیش‌بینی‌های مدل به مشاهدات واقعی نزدیک‌ترند.

شکل (۳) خطای پیش‌بینی طبقه را که به‌وسیلهٔ الگوریتم‌های پیش‌طبقه‌بند تولید شده‌است، نشان می‌دهد؛ در این نمودار، محور افقی، طبقهٔ واقعی، محور عمودی تعداد پیش‌بینی‌های یک طبقهٔ خاص را برای یک طبقهٔ واقعی و رنگ‌های مختلف، طبقه‌های پیش‌بینی‌شدهٔ مختلف را نشان می‌دهند. نتایج بیان‌کنندهٔ آن است که الگوریتم SML در پیش‌بینی طبقه‌ها به نحو مطلوبی عمل کرده‌است.

به‌اختصار می‌توان گفت که مدل SML بهترین عملکرد را در مقایسه با تمام مدل‌ها دارد و رتبهٔ نخست را از منظر کارایی کلی کسب کرده‌است؛ همچنین رتبهٔ دوم از نظر کارایی کلی به مدل ET تعلق دارد که می‌تواند به‌عنوان گزینهٔ دوم برای پیش‌بینی در نظر گرفته شود.

شکل (۴) ماتریس درهم‌ریختگی را برای مدل‌های پیش‌بینی‌کننده نشان می‌دهد. محور عمودی نشان‌دهندهٔ «طبقهٔ واقعی» و محور افقی نشان‌دهندهٔ «طبقهٔ پیش‌بینی‌شده» است. هر عدد در ماتریس درهم‌ریختگی، تعداد نمونه‌هایی را نشان می‌دهد که در طبقهٔ واقعی هدف بوده‌اند و به‌وسیلهٔ مدل در طبقهٔ پیش‌بینی‌شده یک ستون هدف دسته‌بندی شده‌اند. سلول‌های قطر اصلی ماتریس (از بالا سمت چپ به پایین سمت راست)، نشان‌دهندهٔ دسته‌بندی‌های درست (مثبت واقعی)[1] هستند. سلول‌های خارج از قطر اصلی نشان‌دهندهٔ دسته‌بندی‌های نادرست (مثبت کاذب[2] و منفی کاذب)[3] هستند. تحلیل نتایج نشان می‌دهد که مدل SML در دسته‌بندی داده‌ها به‌خوبی عمل کرده‌است و نسبت به سایر مدل‌ها برتری دارد.

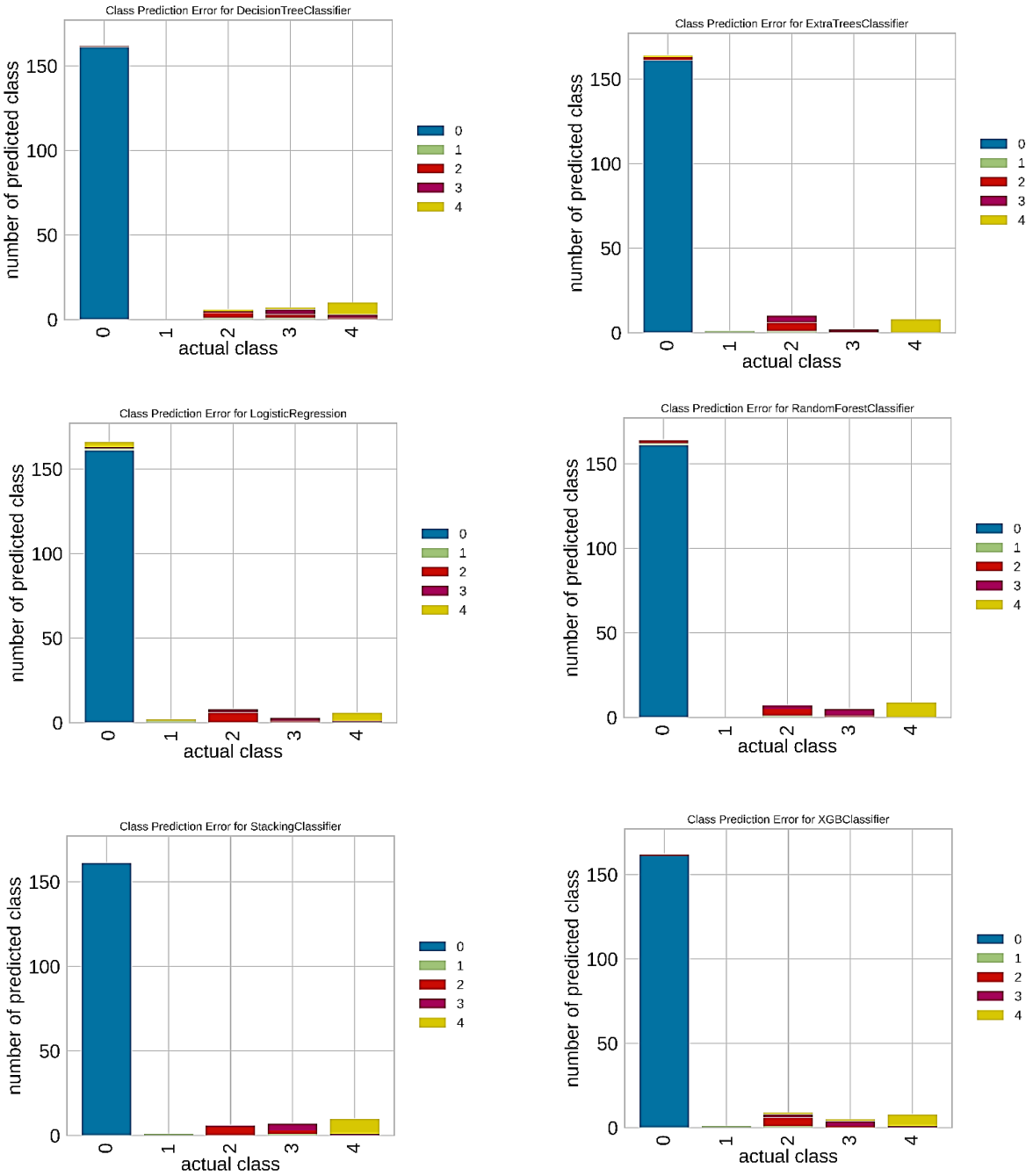
شکل (۵) منحنی‌های مشخصهٔ عملکرد[4] (ROC) را برای مدل‌های دسته‌بندی نشان می‌دهد. برای رسم و نمایش نمودار از کتابخانهٔ matplotlib و برای محاسبات داده‌های مورد نیاز نمودار ROC از کتابخانهٔ scikit-learn استفاده شده‌است. نمودار ROC به‌صورت گرافیکی نمایش می‌دهد که چقدر یک مدل دسته‌بندی عملکرد خوبی دارد و AUC (همان مساحت زیر نمودار) یک معیار عددی است که از نمودار ROC مشتق می‌شود و عملکرد کلی یک مدل را خلاصه می‌کند. در منحنی‌ها محور افقی نرخ مثبت کاذب و محور عمودی نرخ مثبت واقعی یا حساسیت را نشان می‌دهد. منحنی‌های ROC برای پنج طبقه در رنگ‌های مختلف نشان داده شده‌اند؛ همچنین دو منحنی

[1] True Positives
[2] False Positives
[3] False Negatives
[4] Receiver Operating Characteristic



[ DOI: 10.61882/jsdp.22.2.79 ]

میانگین با عنوان micro-average ROC curve و macro-average ROC curve در نمودارها آمده‌است که منحنی نخست عملکرد کلی مدل را در سطح هر نمونه‌داده نشان می‌دهد و منحنی دوم میانگین عملکرد مدل را در سطح طبقه‌ها نشان می‌دهد؛ همچنین خط مورب نقطه‌چین در نمودارها نشان‌دهندهٔ عملکرد یک دسته‌بند تصادفی (شانسی) است.




**(شکل-۳): خطای پیش‌بینی طبقه‌ها به‌وسیلهٔ الگوریتم‌های پیش‌بینی‌کننده بر روی مجموعه‌دادهٔ آموزشی**

**(Figure-3): Class prediction error by predictive algorithms on the training dataset**

رابطهٔ ریاضی برای محاسبهٔ macro-average ROC_AUC ($\alpha_1$) به صورت زیر است:

$$\alpha_1 = \frac{1}{C}\sum_{i=1}^{C}\beta_i \qquad (۱)$$

متغیر $\beta_i$ بیان‌کنندهٔ مساحت زیر منحنی ROC است که به صورت زیر محاسبه شده‌است:

$$\beta_i = \int_0^1 TPR\,(FPR)\,d\,(FPR) \qquad (۲)$$

در این رابطهٔ بالا TPR و FPR به ترتیب بیان‌کنندهٔ نرخ مثبت واقعی و نرخ مثبت کاذب است.

$$TPR = \frac{TP}{TP + FN} \qquad (۳)$$






[ DOI: 10.61882/jsdp.22.2.79 ]

$$TPR = \frac{FP}{FP + TN} \quad (۴)$$

در روابط بالا

TP: تعداد نمونه‌های مثبت که درست پیش‌بینی شده‌اند.

FN: تعداد نمونه‌های مثبت که اشتباه پیش‌بینی شده‌اند.

FP: تعداد نمونه‌های منفی که اشتباه پیش‌بینی شده‌اند.

*TN*: تعداد نمونه‌های منفی که درست پیش‌بینی شده‌اند.

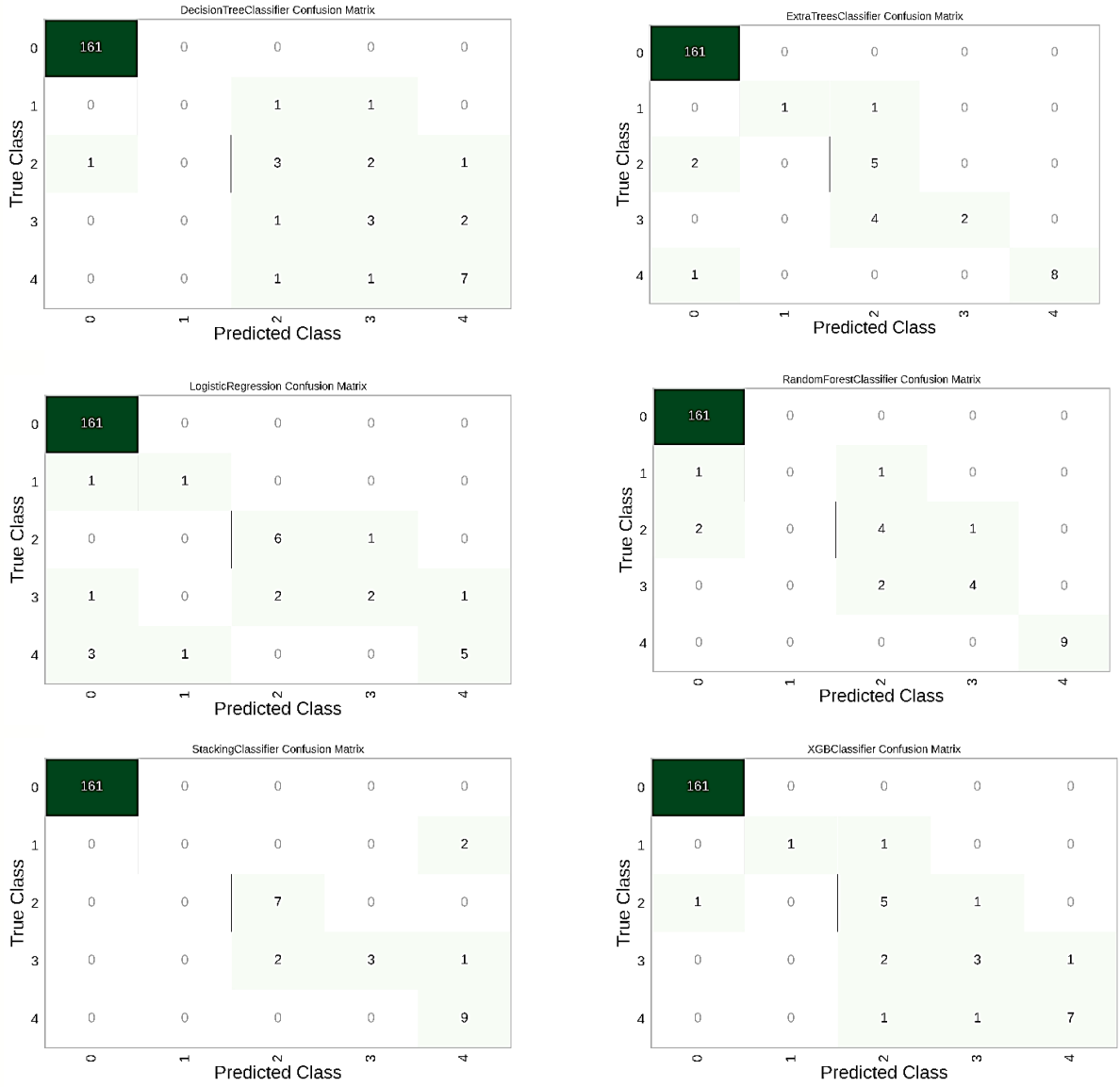
رابطهٔ ریاضی برای محاسبهٔ ( $\alpha_2$ ) micro-average ROC_AUC به صورت زیر است:

$$\alpha_2 = \int_0^1 TPR\,(FPR)\,d\,(FPR) \quad (۵)$$

$$TPR = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C} TP_i + \sum_{i=1}^{C} FN_i} \quad (۶)$$

$$FPR = \frac{\sum_{i=1}^{C} FP_i}{\sum_{i=1}^{C} FP_i + \sum_{i=1}^{C} TN_i} \quad (۷)$$

هر چه منحنی‌های ROC به گوشهٔ بالا سمت چپ نمودار نزدیک‌تر باشند، حاکی از آن است که عملکرد مدل بهتر است. با تحلیل منحنی‌های ROC متوجه می‌شویم که مدل‌های دسته‌بند در بیشتر موارد مقادیر بالای AUC (نزدیک به یک) را کسب کرده و به‌خوبی بین طبقه‌های مختلف تمایز قائل شده‌اند. عملکرد مدل SLM از سایر مدل‌ها بهتر است.



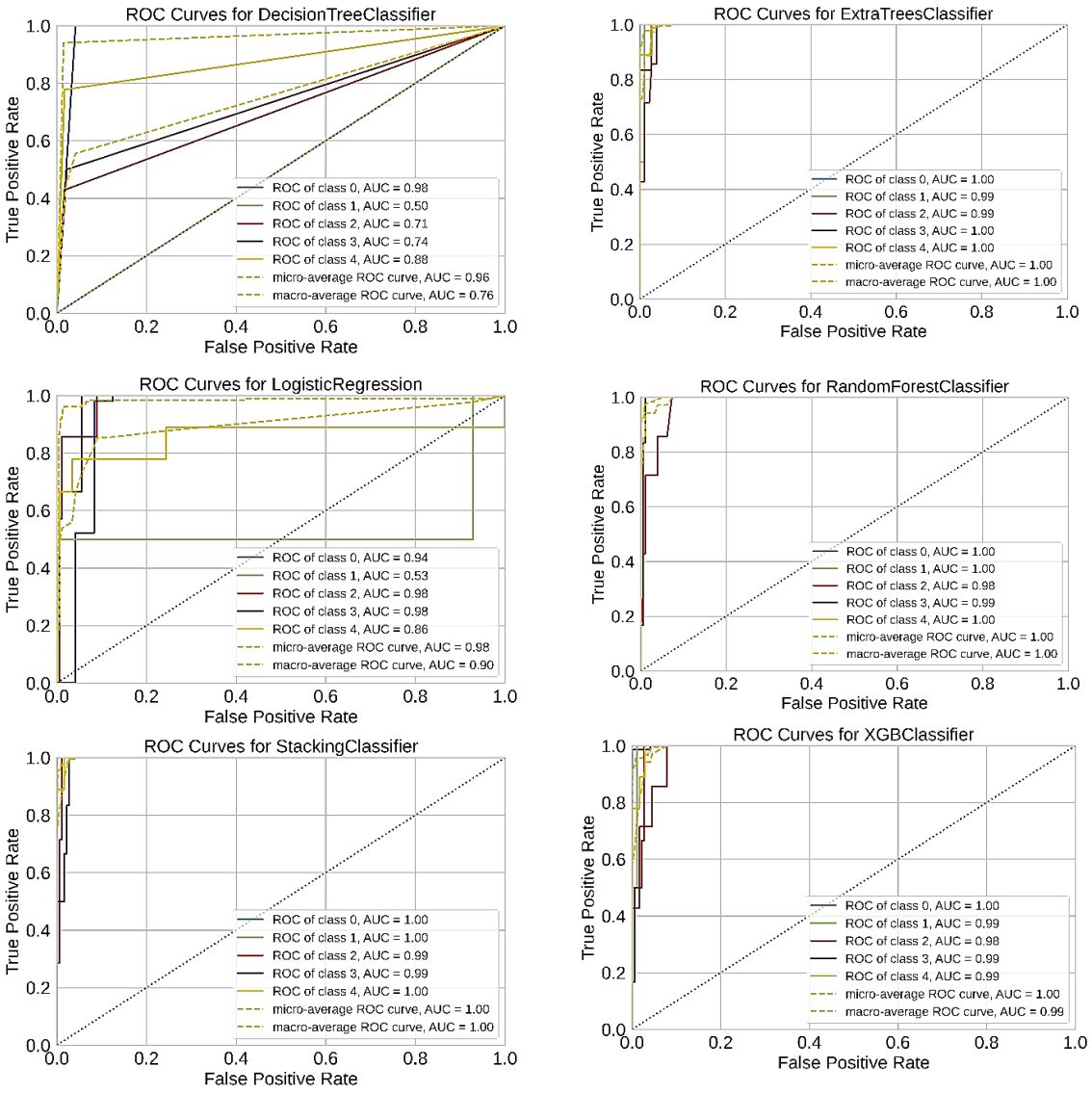

**(شکل-۴): ماتریس درهم ریختگی تولیدشده به‌وسیلهٔ الگوریتم‌های پیش‌بینی‌کننده**

**(Figure-4): Confusion matrix generated by predictive algorithms**





[ DOI: 10.61882/jsdp.22.2.79 ]




**(شکل-۵): منحنی‌های مشخصهٔ عملکرد برای مدل‌های یادگیری ماشین بر روی مجموعه‌دادهٔ آزمایش/آزمایشی**

**(Figure-5): Receiver Operating Characteristic curves for machine learning models on the test dataset**

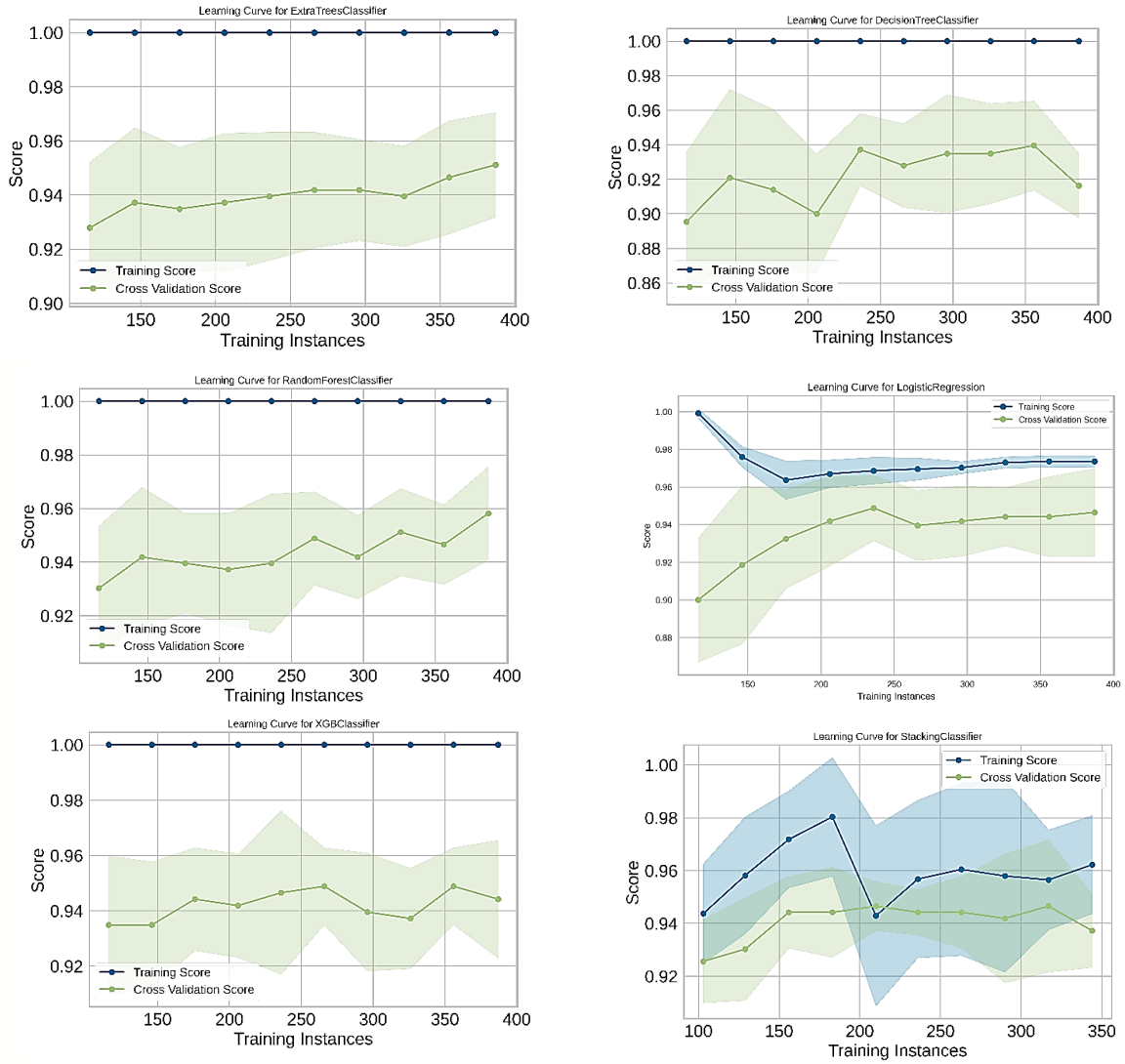
شکل (۶) نمودار مرز تصمیم را برای روش‌های یادگیری ماشین نشان می‌دهد. برای رسم نمودارها از دو ویژگی سن و جنسیت استفاده شده‌است. نتایج نشان می‌دهد که مدل‌های یادگیری ماشین در تفکیک داده‌های طبقه‌بندی‌شده کارایی قابل قبولی کسب کرده و نقاط داده در طبقه‌های مختلف به خوبی در نواحی مختلف تفکیک شده‌اند.

شکل (۷) منحنی یادگیری مدل‌ها را در پیش‌بینی طبقهٔ داده‌ها نشان می‌دهد. دو خط در نمودارها وجود دارد که امتیاز آموزشی و امتیاز اعتبارسنجی را برای مدل‌های مختلف نشان می‌دهد. امتیاز آموزشی نشان‌دهندهٔ دقت مدل بر روی داده‌های آموزشی است. در نمودارهای DT، ET، RF، و XGB امتیاز آموزش نزدیک به یک است که نشان می‌دهد این مدل‌ها می‌توانند به‌خوبی داده‌های آموزشی را یاد گرفته و برازش کنند، اما با خطر بیش‌برازش بر روی داده‌های آموزشی مواجه‌اند. امتیاز اعتبارسنجی متقابل نشان‌دهندهٔ دقت مدل در داده‌های اعتبارسنجی است، که بیان‌کنندهٔ توانایی مدل‌ها برای تعمیم به داده‌های دیده‌نشده است. در نمودارهای مدل‌های DT، ET، RF، و XGB امتیاز آموزش اعتبارسنجی متقابل با امتیاز آموزشی فاصله دارد که نشان می‌دهد این مدل‌ها با خطر بیش‌برازش مواجه‌اند. این موضوع در جدول‌های (۴) و (۵) قابل مشاهده است که بر روی داده‌های آزمایش/آزمایشی، کارایی این مدل‌ها کمتر از مدل SML است. نمودار یادگیری برای مدل‌های SML و LR نشان می‌دهد که تعادل مناسبی بین نمرهٔ آموزشی و نمرهٔ اعتبارسنجی فراهم می‌کند. در نمودار مدل SML امتیاز آموزشی و اعتبارسنجی بسیار نزدیک به یکدیگرند که نشان می‌دهد مشکل بیش‌برازش در داده‌های آموزشی وجود ندارد. این موضوع به ما اطمینان می‌دهد که مدل می‌تواند به‌خوبی به داده‌های دیده‌نشده پاسخ دهد. نمودارها نشان می‌دهند که امتیاز اعتبارسنجی با افزایش تعداد نمونه‌ها بهبود می‌یابد. این نشان می‌دهد که افزودن داده‌های بیشتر به مدل‌ها به بهبود عملکرد و عملکرد پیش‌بینی کمک می‌کند.





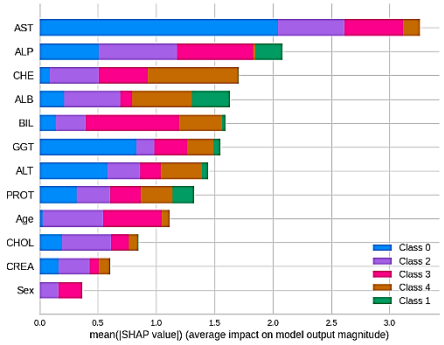
[ DOI: 10.61882/jsdp.22.2.79 ]




**(شکل-۷): منحنی یادگیری مدل‌ها بر روی مجموعه‌دادهٔ آموزشی**

**(Figure-7): Learning curve of models on the training dataset**

شکل (۸) اهمیت نسبی ویژگی‌ها را در پیش‌بینی مدل بر اساس نمودار توضیحات افزدونی شپلی[1] (SHAP) نشان می‌دهد. این نمودار اهمیت ویژگی‌ها را در یک مدل یادگیری ماشین نشان می‌دهد؛ به عبارت دیگر این نمودار نشان می‌دهد که کدام ویژگی‌ها در پیش‌بینی‌های مدل، بیشترین تأثیر را داشته‌اند. در شکل (۷)، نتایج بر اساس خروجی مدل XGB رسم شده‌است. محور افقی در نمودار، میانگین قدرمطلق مقادیر SHAP را برای هر ویژگی نشان می‌دهد؛ هر چه این مقدار بیشتر باشد، نشان‌دهندهٔ تأثیر بیشتر (میزان اهمیت) آن ویژگی در پیش‌بینی‌های مدل است. این محور نشان می‌دهد تأثیر ویژگی به‌طور میانگین چه اندازه در خروجی مدل تغییر ایجاد می‌کند. محور عمودی فهرستی از ویژگی‌های ورودی مدل را به‌ترتیب اهمیت نشان می‌دهد. ویژگی‌هایی که بالاتر قرار دارند، تأثیر بیشتری بر خروجی مدل دارند. هر کدام از رنگ‌ها در نمودار بیان‌کنندهٔ یکی از طبقه‌های خروجی مدل است. طول هر مستطیلِ رنگ نیز بیان‌کنندهٔ میانگین تأثیر آن ویژگی بر روی پیش‌بینی‌های مربوط به هر طبقه است. تحلیل نمودار نشان می‌دهد که AST به‌عنوان تأثیرگذارترین ویژگی در بین سایر ویژگی‌ها مطرح است؛ به‌ویژه در طبقهٔ صفر و چهار این تأثیر بیشتر است. ویژگی ALP بیشترین تأثیر را در طبقهٔ صفر و دو دارد. ویژگی‌های CHE، ALB و BIL تأثیر زیادی در پیش‌بینی‌ها دارند؛ اگر چه تأثیر آن‌ها از AST و ALP کمتر است.




**(شکل-۸): نمودار SHAP بر اساس خروجی مدل XGB بر روی مجموعه‌دادهٔ آموزشی**

**(Figure-8): SHAP chart based on the XGB model on the training dataset**

[1] SHapley Additive exPlanations



[ DOI: 10.61882/jsdp.22.2.79 ]

(جدول-۶): مقایسهٔ مدل پیشنهادی SML با سایر مدل‌های مطرح بر روی مجموعه‌دادهٔ آزمایش/آزمایشی

(Table-6): Comparison of the proposed SML model with other proposed models on the testing dataset

| مدل | مرجع | صحت | دقت | بازخوانی | F1 | MAE | RMSE |
|---|---|---|---|---|---|---|---|
| SVM | [۲۹] | ۰.۹۵۰۰ | ۰.۹۶۰۰ | ۰.۹۴۰۰ | ۰.۹۵۰۰ | --- | --- |
| RFC | [۱۷] | ۰.۸۸۶۶ | ۰.۷۴۴۶ | ۰.۷۵۱۷ | ۰.۷۴۸۱ | --- | ۰.۲۱۸ |
| HGBoost | [۶] | ۰.۸۳۴۳ | ۰.۸۳۰۰ | ۰.۸۳۰۰ | ۰.۸۴۰۰ | --- | ۰.۴۳۰۸ |
| Ensemble learning | [۳] | ۰.۹۸۲۰ | ۰.۹۸۱۲ | ۰.۹۷۸۰ | ۰.۹۸۰۰ | ۰.۰۳۶۰ | ۰.۲۰۴۹ |
| SML | -- | ۰.۹۹۴۰ | ۰.۹۹۴۲ | ۰.۹۹۴۰ | ۰.۹۸۸۰ | ۰.۰۳۰۰ | ۰.۱۷۸۹ |

## ۵-مقایسه با روش‌های مطرح

مدل پیشنهادی SML با چندین مدل لبهٔ دانشی در ادبیات موضوع بر اساس معیارهای کارایی مقایسه شده‌است؛ این روش‌ها عبارت‌اند از ماشین بردار پشتیبان (SVM) [۲۹]، طبقه‌بند جنگل تصادفی (RFC) [۱۷]، درخت طبقه‌بندی افزایش گرادیان مبتنی بر هیستوگرام (HGBoost) [۶] و یادگیری جمعی[1] [۳].

جدول (۶) نتایج کسب‌شده به‌وسیلهٔ مدل پیشنهادی و سایر مدل‌های همتا را بر حسب معیارهای کارایی مقایسه می‌کند. نتایج برتری مدل SML را در قیاس با سایر روش‌های همتا نشان می‌دهد. با توجه به نتایج گزارش‌شده در جدول (۶)، مدل پیشنهادی SML در معیار صحت نسبت به دیگر مدل‌ها برتری دارد و نشان‌دهندهٔ این است که درصد پیش‌بینی‌های صحیح این مدل به‌طور قابل توجهی بیشتر از سایرین است. این مدل در بین تمامی مدل‌های مورد مقایسه بهترین عملکرد را از نظر دقت کسب کرده‌است؛ به عبارت دیگر SML در شناسایی موارد مثبت واقعی و کاهش خطاهای مثبت کاذب بسیار موفق عمل کرده‌است؛ همچنین از آنجا که مدل SML توانسته است نمونه‌های مثبت بیشتری را شناسایی کند و تعداد بسیار کمی از منفی‌های کاذب را گزارش دهد، نرخ بازخوانی مطلوبی نسبت به سایر مدل‌ها دارد؛ علاوه‌براین، مدل پیشنهادی تعادل مناسبی بین دقت و بازخوانی برقرار کرده که این امر در نرخ بالای F1 به‌وضوح مشهود است. پیش‌بینی‌های مدل SML به مقادیر واقعی بسیار نزدیک بوده و به همین دلیل این مدل کمترین مقادیر را بر اساس معیارهای MAE و RMSE نسبت به سایر مدل‌ها کسب کرده‌است. برتری مدل پیشنهادی SML آن را به‌عنوان یک روش قابل اتکا در زمینهٔ پیش‌بینی بیماری‌های کبدی مطرح می‌کند.

## ۶- نتیجه‌گیری

در این پژوهش مسئلهٔ پیش‌بینی بیماری‌های کبدی بررسی شده و روش یادگیری ماشین انباشته (SML) مبتنی بر ترکیب مدل‌های یادگیر پایه و فرامدل ارائه شد. تحلیل نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی در مقایسه با سایر مدل‌های یادگیری ماشین و داده‌کاوی از کارایی بالایی برخوردار است. تحلیل منحنی یادگیری برای مدل‌های مختلف نشان داد که مدل پیشنهادی SML مشکل بیش‌برازش را به‌خوبی حل کرده و امتیاز آموزش و اعتبارسنجی مطلوبی تولید کرده‌است. مطابق دانش نویسندگان مقاله، روش پیشنهادی بالاترین دقت را در بین تمام روش‌های موجود در ادبیات موضوع کسب کرده‌است.

یکی از کارهای آینده، ارزیابی روش پیشنهادی بر روی مجموعه‌داده‌های متنوع با مقیاس بزرگ است تا محدودیت‌ها و نقاط قوت روش پیشنهادی بهتر مشخص شود؛ همچنین استفاده از مدل‌های یادگیر مختلف در لایه‌های مدل SML می‌تواند به بهبود هر چه بیشتر کارایی آن کمک کند.

[1] Ensemble learning

## ۷-مراجع 7-References


[1] C. Gan, Y. Yuan, H. Shen, et al., "Liver diseases: epidemiology, causes, trends and predictions," *Signal Transduction and Targeted Therapy*, vol. 10, no. 33, 2025, doi: 10.1038/s41392-024-02072-z.

[2] S. K. Asrani, H. Devarbhavi, J. Eaton, P. S. K.-J. of hepatology, and undefined 2019, "Burden of liver diseases in the world," *Elsevier*, 2019, doi: 10.1016/j.jhep.2018.09.014.

[3] A. Al Ahad, B. Das, M. R. Khan, N. Saha, A. Zahid, and M. Ahmad, "Multiclass liver disease prediction with adaptive data preprocessing and ensemble modeling," *Results in Engineering*, vol. 22, p. 102059, 2024.

[4] R. K. Sterling *et al.*, "AASLD Practice Guideline on blood-based noninvasive liver disease assessment of hepatic fibrosis and steatosis," *Hepatology*, 2024, doi: 10.1097/HEP.0000000000000845.

[5] S. Aminizadeh *et al.*, "Opportunities and challenges of artificial intelligence and distributed systems to improve the quality of healthcare service," *Artif Intell Med*, vol. 149, Mar. 2024, doi: 10.1016/J.ARTMED.2024.102779.

[6] P. Theerthagiri, "Liver disease classification using histogram-based gradient boosting classification tree with feature selection algorithm," *Biomed Signal Process*

*Control*, vol. 100, Feb. 2025, doi: 10.1016/J.BSPC.2024.107102.

[7] مجرد، موسی، پروین، حمید، نجاتیان، صمد، باقری فرد، کرم الله، «ترکیب یک روش خوشه‌بندی تجمعی و یک معیار شباهت جدید برای مدل‌سازی رفتار وراثتی بیماری‌ها»، *فصلنامهٔ پردازش علائم و داده‌ها*، دورهٔ ۱۸، شمارهٔ ۲، صص۹۷-۱۱۴، ۱۴۰۰.

[7] M. Mojarad, H. Parvin, S. Nejatiyan, and K. A. Bagheri Fard, "Combining an Ensemble Clustering Method and a New Similarity Criterion for Modeling the Hereditary Behavior of Diseases," *Signal and Data Processing*, vol. 18, no. 2, pp. 97–114, Oct. 2021, doi: 10.52547/JSDP.18.2.97.

[8] امامی، نسیبه، حسنی، زینب، «پیش‌بینی و تعیین عوامل مؤثر بر بقای پنج‌سالهٔ کلیهٔ پیوندی در داده‌های نامتوازن با رویکرد فراابتکاری و یادگیری ماشین»، *فصلنامهٔ پردازش علائم و داده‌ها*، دورهٔ ۱۵، شمارهٔ ۴، صص ۸۵-۹۴، ۱۳۹۷.

[8] N. Emami and Z. Hassani, "Prediction and determining the effective factors on the survival transplanted kidney for five-year in imbalanced data by the meta-heuristic approach and machine learning," *Signal and Data Processing*, vol. 15, pp. 85–94, 2019, doi: 10.29252/JSDP.15.4.85.

[9] S. Hashem *et al.*, "Machine Learning Prediction Models for Diagnosing Hepatocellular Carcinoma with HCV-related Chronic Liver Disease," *Comput Methods Programs Biomed*, vol. 196, p. 105551, Nov. 2020, doi: 10.1016/J.CMPB.2020.105551.

[10] K. Moulaei, H. Sharifi, K. Bahaadinbeigy, A. A. Haghdoost, and N. Nasiri, "Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis," *Int J Med Inform*, vol. 179, p. 105243, Nov. 2023, doi: 10.1016/J.IJMEDINF.2023.105243.

[11] D. A. Jadhav, "An enhanced and secured predictive model of Ada-Boost and Random-Forest techniques in HCV detections," *Mater Today Proc*, vol. 51, pp. 186–195, Jan. 2022, doi: 10.1016/J.MATPR.2021.05.071.

[12] F. B. Mostafa and M. E. Hasan, "Machine Learning Approaches for Inferring Liver Diseases and Detecting Blood Donors from Medical Diagnosis," *medRxiv*, Apr. 2021, doi: 10.1101/2021.04.26.21256121.

[13] P. T. Bharathi, S. N. Bindu, S. G. Deepthi, H. U. Gunakeerthi, and K. U. Harshitha, "AI based solution for Predicting Hepatitis C Virus from Blood Samples," *International Conference on Smart Systems for Applications in Electrical Sciences, ICSSES 2024*, 2024, doi: 10.1109/ICSSES62373.2024.10561391.

[14] M. Cedolin, M. E. Genevois, and Z. Canbulat, "Hepatitis C Diagnosis Using Computational Intelligence Techniques," *Lecture Notes in Networks and Systems*, vol. 1090 LNNS, pp. 29–36, 2024, doi: 10.1007/978-3-031-67192-0_4.

[15] M. Arif, M. A. Aslam, H. U. Rehman, M. Abbas, and S. Bukhari, "Laboratory Diagnostic Pathways Using Machine Learning," *VFAST Transactions on Software Engineering*, vol. 10, no. 1, pp. 78–85, Mar. 2022, doi: 10.21015/VTSE.V10I1.826.

[16] I. Trulson, S. Holdenrieder, and G. Hoffmann, "Using machine learning techniques for exploration and classification of laboratory data," *Journal of Laboratory Medicine*, vol. 48, no. 5, pp. 203–214, 2024.

[17] K. N. Singh and J. K. Mantri, "A clinical decision support system using rough set theory and machine learning for disease prediction," *Intelligent Medicine*, vol. 4, no. 3, pp. 200–208, Aug. 2024, doi: 10.1016/J.IMED.2023.08.002.

[18] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, Aug. 2019, doi: 10.1145/3343440.

[19] T.-H. S. Li, H.-J. Chiu, and P.-H. Kuo, "Hepatitis C virus detection model by using random forest, logistic-regression and ABC algorithm," *IEEE Access*, vol. 10, pp. 91045–91058, 2022.

[20] M. M. Ershadi and A. Seifi, "Applications of dynamic feature selection and clustering methods to medical diagnosis," *Appl Soft Comput*, vol. 126, p. 109293, Sep. 2022, doi: 10.1016/J.ASOC.2022.109293.

[21] M. Y. Shams, E. S. M. El-kenawy, A. Ibrahim, and A. M. Elshewey, "A hybrid dipper throated optimization algorithm and particle swarm optimization (DTPSO) model for hepatocellular carcinoma (HCC) prediction," *Biomed Signal Process Control*, vol. 85, p. 104908, Aug. 2023, doi: 10.1016/J.BSPC.2023.104908.

[22] J. S. Sartakhti, M. H. Zangooei, and K. Mozafari, "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)," *Comput Methods Programs Biomed*, vol. 108, no. 2, pp. 570–579, Nov. 2012, doi: 10.1016/J.CMPB.2011.08.003.

[23] M. Yağanoğlu, "Hepatitis C virus data analysis and prediction using machine learning," *Data Knowl Eng*, vol. 142, p. 102087, Nov. 2022, doi: 10.1016/J.DATAK.2022.102087.

[24] F. Mostafa, E. Hasan, M. Williamson, and H. Khan, "Statistical machine learning approaches to liver disease prediction," *Livers*, vol. 1, no. 4, pp. 294–312, 2021.

[25] G. Hoffmann, A. Bietenbeck, R. Lichtinghagen, and F. Klawonn, "Using machine learning techniques to generate laboratory diagnostic pathways—a case study," *J Lab Precis Med*, vol. 3, no. 6, 2018.

[26] D. Chicco and G. Jurman, "An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis," *IEEE Access*, vol. 9, pp. 24485–24498, 2021.

[27] A. Orooji and F. Kermani, "Machine learning based methods for handling imbalanced data in hepatitis diagnosis," *Frontiers in Health Informatics*, vol. 10, no. 1, p. 57, 2021.

[28] H. Mamdouh Farghaly, M. Y. Shams, and T. Abd El-Hafeez, "Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt," *Knowl Inf Syst*, vol. 65, no. 6, pp. 2595–2617, Jun. 2023, doi: 10.1007/S10115-023-01851-4/TABLES/7.



[ DOI: 10.61882/jsdp.22.2.79 ]

[29] A. Alizargar, Y. L. Chang, and T. H. Tan, "Performance Comparison of Machine Learning Approaches on Hepatitis C Prediction Employing Data Mining Techniques," *Bioengineering*, vol. 10, no. 4, p. 481, Apr. 2023, doi: 10.3390/BIOENGINEERING10040481/S1.
[30] R. Safdari, A. Deghatipour, M. Gholamzadeh, and K. Maghooli, "Applying data mining techniques to classify patients with suspected hepatitis C virus infection," *Intelligent Medicine*, vol. 2, no. 4, pp. 193–198, Nov. 2022, doi: 10.1016/J.IMED.2021.12.003.
[31] P. A. A. Resende and A. C. Drummond, "A Survey of Random Forest Based Methods for Intrusion Detection Systems," *ACM Computing Surveys (CSUR)*, vol. 51, no. 3, May 2018, doi: 10.1145/3178582.
[32] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

**حجت امامی** دانش‌آموختهٔ رشتهٔ مهندسی کامپیوتر گرایش هوش مصنوعی است. او دانشیار گروه مهندسی کامپیوتر دانشگاه بناب است. زمینه‌های پژوهشی وی شامل داده‌کاوی، یادگیری ماشین، سیستم‌های چندعاملی، الگوریتم‌های اکتشافی و فرااکتشافی، هوش جمعی و کاربردهای آن است. هم‌اکنون او در زمینهٔ استفاده از یادگیری ماشین در حوزهٔ مدیریت ترافیک هوایی، تشخیص بیماری در حوزهٔ پزشکی و جایابی گره‌ها در شبکه‌های نوری کار پژوهشی خود را ادامه می‌دهد.
نشانی رایانامهٔ ایشان عبارت است از:
emami@ubonab.ac.ir

**بابک آذرنوید** دانش‌آموختهٔ رشتهٔ ریاضی کاربردی و هم اکنون استادیار گروه ریاضی و علوم کامپیوتر دانشگاه بناب است. زمینه‌های پژوهشی او شامل روش‌های عددی حل معادلات دیفرانسیل، آنالیز عددی، یادگیری ماشین و کاربردهای آن‌هاست. در حال حاضر، وی در زمینهٔ استفاده از یادگیری ماشین برای تشخیص بیماری در حوزهٔ پزشکی و همچنین روش‌های عددی و فراابتکاری در حل معادلات دیفرانسیل به فعالیت‌های پژوهشی خود ادامه می‌دهد.
نشانی رایانامهٔ ایشان عبارت است از:
babakazarnavid@ubonab.ac.ir

**محسن عبدالحسین‌زاده** دانش‌آموختهٔ رشتهٔ ریاضی کاربردی گرایش پژوهش در عملیات است. او استادیار گروه ریاضی و علوم کامپیوتر دانشگاه بناب است. زمینه‌های پژوهشی وی شامل بهینه‌سازی شبکه، بهینه‌سازی ترکیباتی، الگوریتم‌های ابتکاری و فراابتکاری، تصمیم‌گیری غیرقطعی، هوش مصنوعی و یادگیری ماشین است. هم‌اکنون او در زمینهٔ استفاده از یادگیری ماشین در حوزهٔ مسیریابی، تشخیص بیماری در حوزهٔ پزشکی و الگوریتم‌های ترکیبی کار پژوهشی خود را ادامه می‌دهد.
نشانی رایانامهٔ ایشان عبارت است از:
mohsen.ab@ubonab.ac.ir



[ DOI: 10.61882/jsdp.22.2.79 ]