(تحلیل و انتخاب مدل بهینه برای پیش‌بینی سکته مغزی با استفاده از الگوریتم‌های طبقه‌بندی)

علی ایزدی
دانشجوی کارشناسی ارشد مهندسی کامپیوتر، دانشگاه بجنورد
ali.izadi.ce@gmail.com

راشد شهابی
دانشجوی کارشناسی مهندسی کامپیوتر، دانشگاه بجنورد
r.shahabi2001@gmail.com

چکیده

در این تحقیق، یک مدل پیش‌بینی وقوع سکته مغزی بر اساس مجموعه داده‌های سلامت طراحی و توسعه داده شده است. هدف این مطالعه، شناسایی مدل با بهترین عملکرد در میان شش الگوریتم طبقه‌بندی شامل رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده، ماشین بردار پشتیبان و K-نزدیک‌ترین همسایه است. پس از انجام مراحل پیش‌پردازش داده‌ها و استفاده از روش تکنیک افزایش مصنوعی نمونه‌های اقلیت (SMOTE) برای متوازن‌سازی داده‌های نامتوازن، هر یک از الگوریتم‌ها آموزش داده شده و نتایج آن‌ها با استفاده از معیارهای ارزیابی شامل دقت کلی (Accuracy)، بازخوانی (Recall)، دقت (Precision)، امتیاز F1 (F1-Score)، میانگین کلان (Macro Avg) و میانگین وزنی (Weighted Avg) تحلیل شده‌اند.

برای تحلیل دقیق‌تر، نمودارهای مشخصه عملکرد گیرنده، دقت-بازخوانی و ماتریس درهم‌ریختگی برای هر مدل رسم و بررسی شدند. نتایج نشان داد که مدل جنگل تصادفی به دلیل دقت بالا، بازخوانی مناسب و توانایی در مدیریت داده‌های پیچیده و غیرخطی، بهترین عملکرد را در پیش‌بینی وقوع سکته مغزی داشته است. در مقابل، مدل‌های لجستیک و بیز ساده ضعف بیشتری در شناسایی موارد مثبت نشان دادند.

این مطالعه نشان می‌دهد که الگوریتم‌های طبقه‌بندی می‌توانند ابزارهای مؤثری در حوزه‌های پزشکی برای پیش‌بینی و تصمیم‌گیری باشند و انتخاب مدل مناسب تأثیر بسزایی در بهبود دقت پیش‌بینی‌ها دارد.

برای مشاهده کد پروژه، به این لینک مراجعه کنید.

کلمات کلیدی: الگوریتم‌های طبقه‌بندی، پیش‌بینی سکته مغزی، جنگل تصادفی، یادگیری ماشین، داده‌های پزشکی

۱. مقدمه

در دنیای امروز، سلامت انسان‌ها و پیشگیری از بیماری‌ها یکی از چالش‌های اصلی جوامع محسوب می‌شود. در این راستا، تشخیص و پیش‌بینی به موقع بیماری‌های مختلف، از جمله سکته مغزی[۱]، می‌تواند نقش مهمی در کاهش عوارض و مرگ و میر ایفا کند. با گسترش دسترسی به داده‌های پزشکی و پیچیده‌تر شدن روش‌های تحلیل، بهره‌گیری از مدل‌های پیش‌بینی برای تشخیص و تحلیل بیماری‌ها بیش از پیش مورد توجه قرار گرفته است. (مهدی‌پور، ۱۳۹۵) با ظهور الگوریتم‌های طبقه‌بندی[۲] و یادگیری ماشین[۳]، روش‌های پیش‌بینی در این حوزه دچار تحول عظیمی شده‌اند، به گونه‌ای که اکنون امکان پیش‌بینی با دقت بیشتر فراهم شده است. (Chen et al, 2021)

در گذشته، روش‌های سنتی آماری و تحلیل‌های اولیه برای پیش‌بینی و تشخیص بیماری‌ها به کار گرفته می‌شدند که غالباً به دلیل محدودیت در داده‌ها و پیچیدگی کمتر، توانایی کافی برای تحلیل شرایط چندوجهی و پیچیده را نداشتند (Zhou, 2014). با پیشرفت تکنولوژی و افزایش توان محاسباتی، به کارگیری الگوریتم‌های پیشرفته یادگیری ماشین در سال‌های اخیر منجر به ارتقای دقت و توانایی تحلیل داده‌های پزشکی شده است. این پیشرفت به خصوص در زمینه تشخیص بیماری‌ها و مدیریت داده‌های پیچیده به وضوح قابل مشاهده است. (Rajula et al, 2020)

الگوریتم‌های طبقه‌بندی مختلف، مانند رگرسیون لجستیک (LaValley, 2008)، درخت تصمیم (Song and Ying, 1986)، جنگل تصادفی (Rigatti, 2017)، بیز ساده (Webb et al, 2010)، ماشین بردار پشتیبان(Pisner and Schnyer, 2020) و K-نزدیک‌ترین همسایه (Peterson, 2009)، در حوزه‌های مختلف از جمله پزشکی، مهندسی و اقتصاد کاربرد گسترده‌ای دارند. این الگوریتم‌ها، هر یک با ویژگی‌ها و توانایی‌های منحصر به فرد خود، برای تحلیل و طبقه‌بندی داده‌های پیچیده و پیش‌بینی نتایج در بسیاری از حوزه‌ها به کار گرفته می‌شوند. (King et al, 1995)

در حوزه پزشکی، پیش‌بینی سکته مغزی یکی از مهم‌ترین موضوعات تحقیقاتی است. این پیش‌بینی می‌تواند به بیماران، پزشکان و سیاست‌گذاران بهداشتی در مدیریت بهتر عوامل خطر و بهبود فرآیند درمان کمک کند. مدل‌های مختلف طبقه‌بندی، از جمله جنگل تصادفی، توانسته‌اند با تحلیل داده‌های گسترده پزشکی، از جمله سن، سابقه بیماری، فشار خون و شاخص‌های دیگر، پیش‌بینی دقیقی از وقوع سکته مغزی ارائه دهند. (Bonkhoff and Grefkes, 2022)

در این مقاله، به ارزیابی مدل‌های مختلف طبقه‌بندی برای یافتن بهترین مدل پیش‌بینی وقوع سکته مغزی پرداخته‌ایم. مدل‌هایی مانند رگرسیون لجستیک[۴]، درخت تصمیم[۵]، جنگل تصادفی[۶]، بیز ساده[۷]، ماشین بردار پشتیبان[۸] و K-نزدیک‌ترین همسایه[۹] مورد ارزیابی قرار گرفته‌اند. هدف این پژوهش، تحلیل عملکرد این مدل‌ها و تعیین بهترین الگوریتم برای پیش‌بینی دقیق‌تر سکته مغزی است. بر اساس نتایج به‌دست‌آمده، انتظار می‌رود که مدل جنگل تصادفی به دلیل توانایی در تحلیل داده‌های پیچیده و روابط غیرخطی، عملکرد بهتری نسبت به دیگر مدل‌ها ارائه دهد.

۲. روش تحقیق

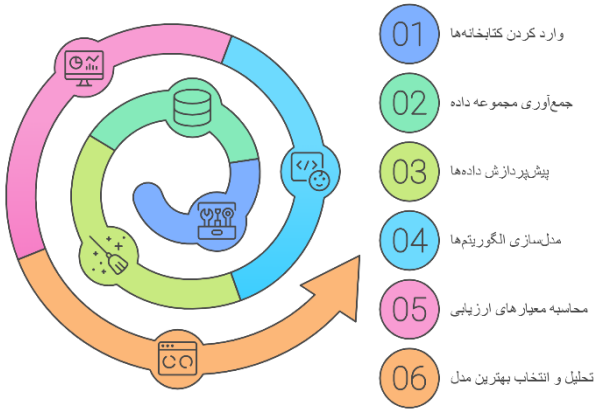
این پژوهش با استفاده از یک مجموعه داده بین‌المللی شامل ۵۱۱۰ سطر و ۱۲ ویژگی، به بررسی و پیش‌بینی سکته مغزی پرداخته است. هدف اصلی، ارزیابی عملکرد چندین مدل طبقه‌بندی با استفاده از زبان برنامه‌نویسی پایتون است. مراحل انجام این تحقیق در شکل ۱ نمایش داده شده است.

---

[1] Stroke
[2] Classification
[3] Machine Learning
[4] Logistic Regression
[5] Decision Tree
[6] Random Forest
[7] Naïve Bayes
[8] Support Vector Machine
[9] K-Nearest Neighbors

در گام نخست، کتابخانه‌های مورد نیاز برای تحلیل داده‌ها و توسعه مدل‌ها وارد شدند. سپس، مجموعه داده مربوط به سکته مغزی آماده‌سازی و در فرآیند مدل‌سازی استفاده شد. در مرحله بعد، داده‌ها تحت پیش‌پردازش قرار گرفتند تا برای الگوریتم‌های یادگیری ماشین آماده شوند. مرحله بعدی به آموزش و ارزیابی مدل‌های مختلف طبقه‌بندی اختصاص دارد. در پایان، معیارهای ارزیابی مدل‌ها محاسبه شده و با مقایسه عملکرد الگوریتم‌ها وتحلیل نمودارها، بهترین مدل برای پیش‌بینی سکته مغزی انتخاب شد.

در ادامه، جزئیات هر یک از این مراحل به طور کامل توضیح داده می‌شود.



شکل ۱. روند انجام کار

۲-۱. ورود کتابخانه‌های مورد نیاز

در فرآیند توسعه مدل‌های یادگیری ماشین و تحلیل داده‌ها، استفاده از کتابخانه‌های پیشرفته نخستین گام در فرآیند پژوهش بود. این کتابخانه‌ها ابزارهای تخصصی و قدرتمندی را در اختیار پژوهشگران قرار می‌دهند که امکان انجام تحلیل‌ها و مدل‌سازی‌ها را با دقت و کارایی بیشتری فراهم می‌کنند.

در این پژوهش، کتابخانه Pandas برای مدیریت و پردازش داده‌ها و NumPy برای محاسبات عددی استفاده شدند. همچنین، کتابخانه Matplotlib و Seaborn برای مصورسازی داده‌ها و ارائه نتایج گرافیکی به کار گرفته شدند. برای توسعه و ارزیابی مدل‌های یادگیری ماشین، ازScikit-learn بهره گرفته شد. علاوه براین، از کتابخانه imblearn برای مدیریت داده‌های نامتوازن استفاده گردید که نقشی حیاتی در بهبود عملکرد مدل‌ها داشت. این ترکیب از کتابخانه‌ها چارچوبی جامع و قدرتمند برای تحلیل داده‌ها و پیاده‌سازی مدل‌های طبقه‌بندی ارائه داد.

۲-۲. مجموعه داده

در این بخش از تحقیق، از مجموعه داده‌ای به نام Stroke Prediction Dataset استفاده شده است که از سایت Kaggle دریافت شده است. (Fedesoriano, 2020) این مجموعه داده شامل اطلاعات جمع‌آوری‌شده از افراد مختلف است که ویژگی‌هایی مانند جنسیت، سن، سابقه بیماری قلبی، فشار خون[10]، سطح گلوکز[11]، شاخص توده بدنی[12]، وضعیت تأهل، نوع شغل، محل سکونت و وضعیت استعمال دخانیات را در بر می‌گیرد. متغیر هدف این مجموعه داده، وقوع یا عدم وقوع سکته مغزی است.

[10] Hypertension
[11] Glucose
[12] BMI

این مجموعه داده شامل 5110 نمونه و 12 ویژگی است که برای تحلیل و مدل‌سازی استفاده شده‌اند. داده‌ها به گونه‌ای طراحی شده‌اند که اطلاعات جامع و متنوعی را برای پیش‌بینی و تحلیل سکته مغزی فراهم کنند. در ادامه به طور کلی، به توصیف آن‌ها می‌پردازیم.

- شناسه (ID): شناسه منحصربه‌فرد هر فرد در مجموعه داده.
- جنسیت (Gender): جنسیت فرد که می‌تواند "مذکر" یا "مونث" باشد.
- سن (Age): سن فرد به سال.
- فشار خون (Hypertension): مشخص می‌کند که آیا فرد سابقه فشار خون بالا دارد یا خیر (۱: بله، ۰: خیر).
- بیماری قلبی (Heart Disease): مشخص می‌کند که آیا فرد دچار بیماری قلبی است یا خیر (۱: بله، ۰: خیر).
- وضعیت تأهل (Ever Married): وضعیت تأهل فرد که می‌تواند "بله" یا "خیر" باشد.
- نوع شغل (Work Type): نوع شغل فرد شامل "شغل خصوصی"، "شغل آزاد"، "کارهای دولتی" و "بیکار".
- نوع محل سکونت (Residence Type): نوع محل سکونت فرد که می‌تواند "شهری" یا "روستایی" باشد.
- میانگین سطح گلوکز خون (Avg Glucose Level): میانگین سطح گلوکز خون فرد بر حسب میلی‌گرم بر دسی‌لیتر.
- شاخص توده بدنی (BMI): شاخص توده بدنی فرد که از نسبت وزن به قد محاسبه شده است.
- وضعیت استعمال دخانیات (Smoking Status): وضعیت استعمال دخانیات فرد شامل "هرگز سیگار نکشیده"، "سابقاً سیگار کشیده"، و "در حال سیگار کشیدن".
- وقوع سکته مغزی (Stroke): متغیر هدف که نشان می‌دهد آیا فرد دچار سکته مغزی شده است یا خیر (۱: بله، ۰: خیر).

۲-۳. پیش‌پردازش داده‌ها

جهت پیش‌پردازش داده‌ها در این پژوهش، مراحل زیر انجام شده است:

- حذف ستون‌های غیرضروری: ستون شناسه به دلیل نداشتن تأثیر مستقیم در پیش‌بینی حذف شد.
- مدیریت مقادیر گمشده: مقادیر گمشده در ستون شاخص توده بدنی با استفاده از میانگین مقدار ستون تکمیل شدند.
- کدگذاری متغیرهای دسته‌بندی‌شده: ستون‌های دسته‌بندی‌شده مانند جنسیت، وضعیت تأهل، نوع شغل، نوع سکونت، و وضعیت استعمال دخانیات با استفاده از روش برچسب‌گذاری[۱۳] به صورت عددی تبدیل شدند.
- تعریف ویژگی‌ها و متغیر هدف: متغیر هدف (stroke) برای پیش‌بینی سکته مغزی انتخاب شد و سایر ستون‌ها به‌عنوان ویژگی‌های مستقل در نظر گرفته شدند.
- متعادل‌سازی داده‌ها: به دلیل عدم توازن در داده‌های مربوط به وقوع سکته مغزی، از تکنیک افزایش مصنوعی نمونه‌های اقلیت[۱۴] برای متعادل‌سازی داده‌ها استفاده شد و داده‌ها به‌صورت متعادل درآمدند.
- تقسیم داده‌ها: داده‌ها به دو مجموعه آموزشی (۸۰٪) و آزمایشی (۲۰٪) تقسیم شدند تا عملکرد مدل‌ها به‌صورت عادلانه ارزیابی شود.
- استانداردسازی ویژگی‌ها: ویژگی‌های عددی با استفاده از StandardScaler استانداردسازی شدند تا مقیاس تمام متغیرها یکسان شود و تأثیر ویژگی‌های با مقیاس بزرگ کاهش یابد.

این مراحل پیش‌پردازش داده‌ها شرایط لازم را برای تحلیل دقیق‌تر و آموزش مدل‌های یادگیری ماشین فراهم کرد.

۲-۴. آموزش مدل‌های مختلف طبقه‌بندی

در این مرحله، داده‌ها با استفاده از الگوریتم‌های مختلف طبقه‌بندی آموزش داده می‌شوند. این الگوریتم‌ها شامل رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده، ماشین بردار پشتیبان و K-نزدیک‌ترین همسایه هستند. هر یک از این

[13] Label Encoding
[14] Synthetic Minority Oversampling Technique (SMOTE)

مدل‌ها با توجه به ویژگی‌های تأثیرگذار موجود در داده‌ها، مانند سن، فشار خون، سطح گلوکز و شاخص توده بدنی برای پیش‌بینی وقوع یا عدم وقوع سکته مغزی آموزش می‌بینند.
این فرآیند شامل ارزیابی عملکرد هر مدل و بررسی معیارهایی مانند دقت کلی، بازخوانی و امتیاز F1 است. نتایج به دست آمده از این مرحله به ما امکان می‌دهد تا عملکرد مدل‌ها را مقایسه کرده و بهترین الگوریتم را برای پیش‌بینی سکته مغزی انتخاب کنیم.

۲-۵. نمودارها

پس از آن که مدل‌های آموزش‌دیده‌شده با الگوریتم‌های ذکرشده ایجاد شدند، نمودارهای منحنی مشخصه عملکرد گیرنده[15] (ROC)، نمودار دقت-بازخوانی[16] و ماتریس درهم‌ریختگی[17] برای هر یک از مدل‌های آموزش ترسیم شدند. نمودار ROC نموداری است که از نمایش نرخ مثبت‌های صحیح[18] در مقابل نرخ مثبت‌های کاذب[19] به دست می‌آید. این نمودار به طور گسترده برای ارزیابی توانایی مدل در تفکیک کلاس‌ها استفاده می‌شود.
نمودار نمودار دقت-بازخوانی نیز نشان‌دهنده رابطه بین دقت و بازخوانی مدل‌ها است و برای تحلیل عملکرد مدل‌ها در داده‌های نامتوازن بسیار مفید است. این نمودار معمولاً برای بررسی نحوه تعادل مدل بین بازخوانی و دقت به کار می‌رود.
ماتریس درهم‌ریختگی نیز برای نمایش تعداد پیش‌بینی‌های درست و نادرست مدل در هر کلاس مورد استفاده قرار گرفت. این ماتریس شامل مقادیر پیش‌بینی‌های صحیح و خطاها برای هر دسته‌بندی است و تحلیل آن به ما امکان می‌دهد تا نقاط قوت و ضعف مدل‌ها را در پیش‌بینی هر کلاس شناسایی کنیم.
در بخش نتایج، این نمودارها برای مقایسه و بررسی عملکرد مدل‌های مختلف ارائه و تحلیل شده‌اند.

۲-۶. محاسبه معیارهای ارزیابی

در این بخش، به بررسی و محاسبه معیارهای مختلف ارزیابی برای الگوریتم‌های به‌کاررفته می‌پردازیم. این معیارها به‌منظور سنجش عملکرد مدل‌های طبقه‌بندی و تعیین دقت پیش‌بینی آن‌ها استفاده می‌شوند. معیارهای مورد استفاده شامل دقت[20]، بازخوانی[21]، امتیاز[22] F1، دقت کلی[23]، میانگین کلان[24]، میانگین وزنی[25] و مساحت زیر منحنی[26] هستند که در ادامه شرح داده می‌شوند:
- دقت: این معیار، نسبت پیش‌بینی‌های درست مثبت به کل پیش‌بینی‌های مثبت را نشان می‌دهد. به عبارتی، دقت مشخص می‌کند که از میان تمام پیش‌بینی‌های مثبت، چه تعداد آن‌ها واقعاً صحیح بوده‌اند. دقت بالا زمانی حاصل می‌شود که مدل بتواند مثبت‌های واقعی را با کمترین خطای مثبت کاذب پیش‌بینی کند. مقدار دقت باید به ۱ نزدیک باشد؛ مقادیر پایین نشان‌دهنده وجود تعداد زیادی پیش‌بینی‌های مثبت کاذب است که می‌تواند در کاربردهایی مانند پزشکی مشکل‌ساز باشد. این معیار با استفاده از رابطه (۱) محاسبه می‌شود.

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)+False Positive (FP)}} \quad (۱)$$

[15] Receiver Operating Characteristic (ROC)
[16] Precision-Recall Curve
[17] Confusion Matrix
[18] True Positive Rate (TPR)
[19] False Positive Rate (FPR)
[20] Precision
[21] Recall
[22] F1-Score
[23] Accuracy
[24] Macro Average
[25] Weight Average
[26] Area Under Curve (AUC)

- بازخوانی: این معیار، نسبت پیش‌بینی‌های درست مثبت به کل مقادیر واقعی مثبت را اندازه‌گیری می‌کند. بازخوانی نشان می‌دهد که از تمام موارد مثبت واقعی، چه تعداد آن‌ها به‌درستی توسط مدل شناسایی شده‌اند. بازخوانی بالا نشان‌دهنده توانایی مدل در شناسایی موارد مثبت است. مقدار پایین این معیار به معنای از دست دادن تعداد زیادی از موارد مثبت واقعی است که در موارد حساس مانند پیش‌بینی بیماری بسیار مشکل‌ساز است. این معیار طبق رابطه (۲) محاسبه می‌شود. از Recall به عنوان TPR نیز نام برده می‌شود.

$$(۲)\ \text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)+False Negative (FN)}}$$

- امتیاز F1: امتیاز F1 میانگین هارمونیک دقت و بازخوانی است و تعادلی بین این دو معیار ایجاد می‌کند. این معیار برای ارزیابی عملکرد مدل در داده‌های نامتوازن بسیار مناسب است. مقدار F1 نزدیک به ۱ نشان‌دهنده تعادل مناسب بین دقت و بازخوانی است. مقادیر پایین این معیار ممکن است به معنای ضعف مدل در یکی از این جنبه‌ها باشد. این معیار طبق رابطه (۳) محاسبه می‌شود.

$$(۳)\ \text{F1-Score} = \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} * 2$$

- دقت کلی: دقت کلی نسبت تعداد پیش‌بینی‌های درست (برای هر دو کلاس مثبت و منفی) به کل پیش‌بینی‌ها را نشان می‌دهد. این معیار برای داده‌های متوازن مفید است. در داده‌های نامتوازن، مقدار دقت ممکن است گمراه‌کننده باشد، زیرا مدل می‌تواند با پیش‌بینی کلاس غالب به دقت بالا دست یابد. مقدار دقت نزدیک به ۱ برای داده‌های متوازن مطلوب است. فرمول دقت کلی در رابطه (۴) آورده شده است.

$$(۴)\ \text{Accuracy} = \frac{\text{True Positive (TP)+True Negative (TN)}}{\text{Total Samples}}$$

- میانگین کلان: این معیار، میانگین ساده دقت و بازخوانی برای تمامی کلاس‌ها را بدون توجه به تعداد نمونه‌های هر کلاس محاسبه می‌کند. این معیار برای داده‌های نامتوازن که کلاس‌های کم‌تعداد نیز اهمیت دارند، مناسب است. مقدار بالای Macro Avg نشان‌دهنده عملکرد خوب مدل برای همه کلاس‌ها است. فرمول محاسبه این معیار در رابطه (۵) ارائه شده است.

$$(۵)\ \text{Macro Avg} = \text{Metric}_i \sum_{i=1}^{N} \frac{1}{N}$$

- میانگین وزنی: این معیار، میانگین دقت و بازخوانی برای تمامی کلاس‌ها را با توجه به تعداد نمونه‌های هر کلاس محاسبه می‌کند. این معیار تأثیر کلاس‌های پرتکرار را در نظر می‌گیرد و برای داده‌های نامتوازن کاربردی است. مقدار بالای Weighted Avg به معنای عملکرد قابل‌قبول مدل برای کلاس‌های پرتکرار است، اما ممکن است تأثیر کلاس‌های کم‌تعداد کمتر نمایان باشد. فرمول این معیار در رابطه (۶) آمده است.

$$(۶)\ \text{Weight Avg} = \text{Metric}_i \frac{n_i}{N_{total}} \sum_{i=1}^{N} \frac{1}{N}$$

- مساحت زیر منحنی: این معیار، مساحت زیر منحنی مشخصه عملکرد گیرنده (ROC) را اندازه‌گیری می‌کند. مقدار AUC بین ۰ تا ۱ قرار دارد و نشان‌دهنده توانایی مدل در تفکیک صحیح کلاس‌ها است. AUC نزدیک به ۱ نشان می‌دهد که مدل دارای عملکرد بسیار خوب در شناسایی کلاس‌های مثبت و منفی است، در حالی که AUC نزدیک به ۰.۵ به معنای عملکردی مشابه یک مدل تصادفی است. مدل‌هایی با AUC بالا به‌ویژه در کاربردهایی که شناسایی دقیق موارد مثبت اهمیت زیادی

دارد، مناسب‌تر هستند. این معیار طبق رابطه (۷) و (۸) با روش Trapezoidal محاسبه می‌شود؛ که در آن TPR_i و TPR_{i+1} به ترتیب مقادیر TPR برای دو نقطه متوالی در منحنی ROC هستند و FPR_i و FPR_{i+1} مقادیر FPR برای همان نقاط هستند.

$$(۷)\ FPR = \frac{False\ Positive\ (FP)}{False\ Positive\ (FP) + True\ Negative\ (TN)}$$

$$(۸)\ AUC \approx \sum_{i=1}^{N-1} \frac{TPR_i + TPR_{i+1}}{2} * (FPR_{i+1} - FPR_i)$$

این معیارها به تحلیل دقیق عملکرد مدل‌ها و مقایسه اثربخشی الگوریتم‌های مختلف کمک می‌کنند. در بخش نتایج، مقادیر محاسبه‌شده برای این معیارها ارائه و تحلیل می‌شوند.

۲-۷. ارزیابی و انتخاب بهترین مدل

در این مرحله، مدل‌های مختلف طبقه‌بندی با استفاده از معیارهای ارزیابی مانند دقت کلی، بازخوانی، دقت، امتیاز F1، میانگین کلان، میانگین وزنی و مساحت زیر منحنی مقایسه می‌شوند. این مقایسه به منظور تحلیل عملکرد هر مدل در پیش‌بینی سکته مغزی و شناسایی دقیق‌ترین و کارآمدترین الگوریتم انجام می‌گیرد.

برای ارزیابی بهتر، نمودارهای ROC و Precision-Recall نیز بررسی شده و تحلیل می‌شوند تا عملکرد مدل‌ها از زوایای مختلف مورد سنجش قرار گیرد. مدل‌هایی که در معیارهای کلیدی عملکرد بهتری داشته باشند و در داده‌های نامتوازن بازدهی مناسبی نشان دهند، به‌عنوان مدل‌های برتر شناسایی خواهند شد.

پس از تحلیل نتایج، مدلی که بهترین تعادل را میان دقت و بازخوانی ایجاد کند و بالاترین امتیاز F1 را ارائه دهد، به‌عنوان مناسب‌ترین الگوریتم برای پیش‌بینی سکته مغزی انتخاب خواهد شد. این انتخاب به ما امکان می‌دهد تا در کاربردهای واقعی، مدلی دقیق و قابل‌اعتماد برای پیش‌بینی به‌کار گرفته شود.

۳. نتایج

در این قسمت، انواع خروجی‌ها مورد تجزیه و تحلیل قرار می‌گیرند. ابتدا، نمودارهای ROC، Precision-Recall و Confusion Matrix برای هر یک از شش مدل آموزش‌داده‌شده بررسی می‌شوند. این نمودارها اطلاعات مفیدی درباره توانایی مدل‌ها در تشخیص کلاس‌ها و تحلیل عملکرد کلی ارائه می‌دهند.

مدل‌های آموزشی استفاده‌شده در این پژوهش عبارت‌اند از: رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده، ماشین بردار پشتیبان و K-نزدیک‌ترین همسایه.

۱-۳. نمودار مشخصه عملکرد گیرنده (ROC)

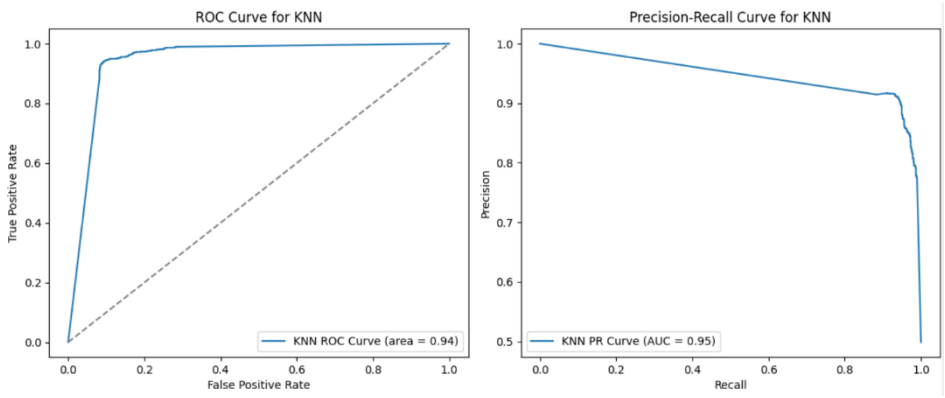
نمودارهای مشخصه عملکرد گیرنده مطابق شکل ۲، توانایی هر یک از شش مدل آموزشی را در تفکیک کلاس‌ها نشان می‌دهند. محور افقی این نمودار نرخ مثبت‌های کاذب (FPR) و محور عمودی آن نرخ مثبت‌های صحیح (TPR) را نشان می‌دهد. هرچه منحنی ROC به گوشه بالا سمت چپ نمودار نزدیک‌تر باشد، مدل عملکرد بهتری دارد.

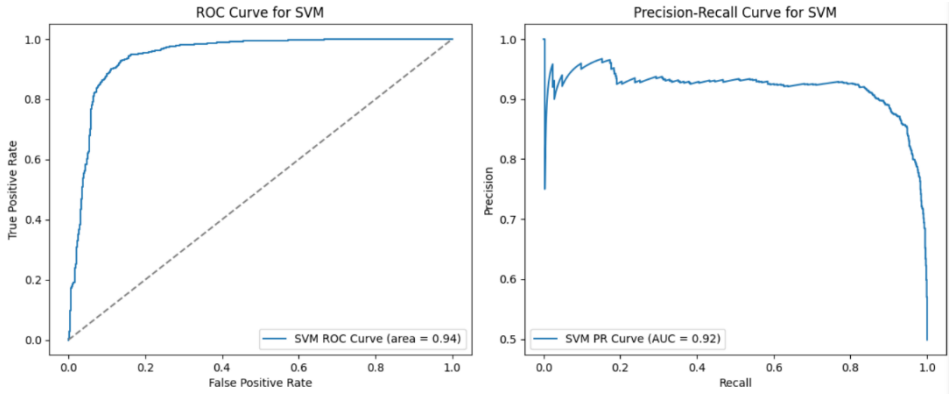
برای تحلیل بهتر، مساحت زیر منحنی (AUC) نیز محاسبه شده است. AUC نزدیک به ۱ نشان‌دهنده توانایی بالای مدل در تفکیک کلاس‌ها و عملکرد مطلوب است، در حالی که AUC نزدیک به ۰.۵ به معنای عملکردی مشابه یک مدل تصادفی است. در این بررسی، مدل جنگل تصادفی با AUC بالاتر از ۰.۹ توانسته عملکرد برتری نشان دهد، در حالی که مدل‌های K-نزدیک‌ترین همسایه و بیز ساده با AUC کمتر از ۰.۷ عملکرد ضعیف‌تری داشته‌اند. منحنی‌هایی که به خط مرجع (خط قطر) نزدیک‌تر باشند، نشان‌دهنده عملکرد ضعیف مدل هستند.

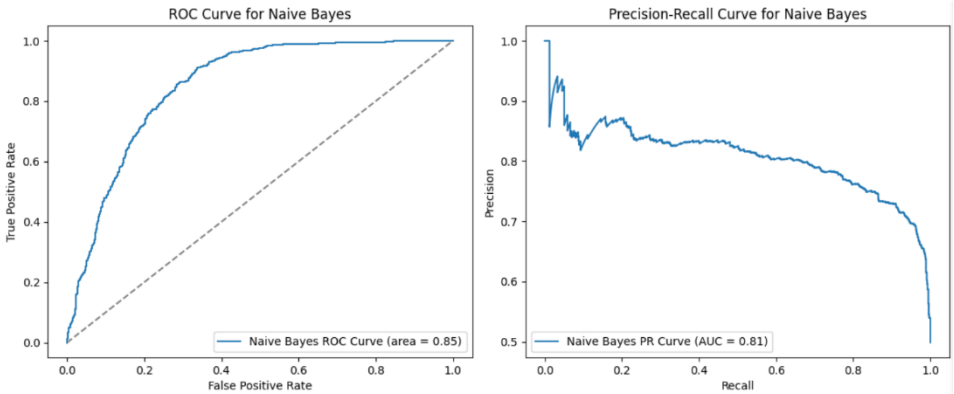
۲-۳. نمودار دقت-بازخوانی (Precision-Recall)

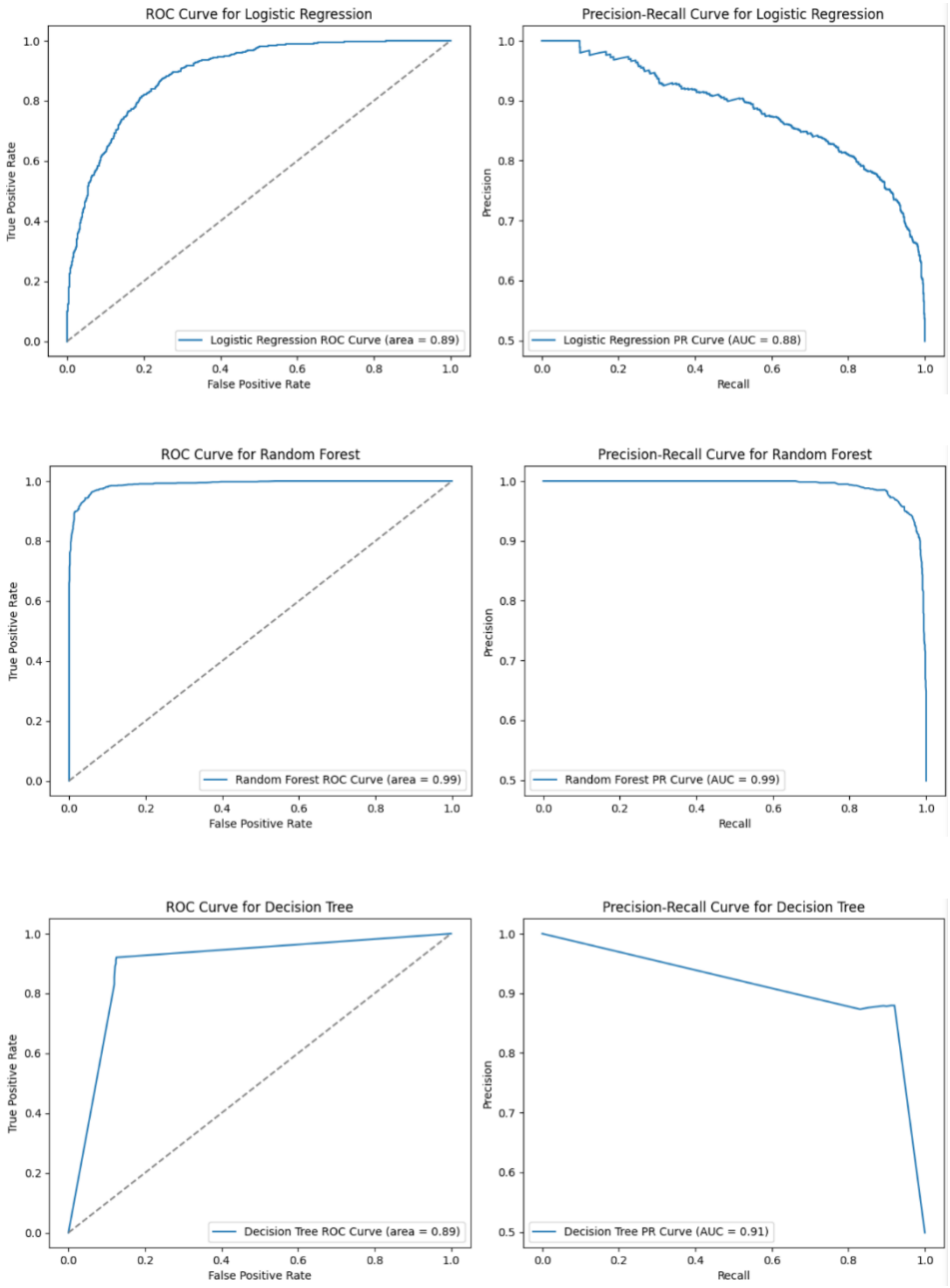
نمودارهای دقت-بازخوانی طبق شکل ۲، نشان‌دهنده رابطه بین دقت و بازخوانی مدل‌ها هستند. محور افقی این نمودار نشان‌دهنده بازخوانی و محور عمودی دقت است. این نمودارها به‌ویژه در تحلیل داده‌های نامتوازن که کلاس اقلیت اهمیت بیشتری دارد، بسیار مفید هستند.

براساس این نمودارها، مدل جنگل تصادفی توانسته است بازخوانی بالا را بدون کاهش دقت ارائه دهد که نشان‌دهنده توانایی بالای این مدل در شناسایی موارد مثبت واقعی است. مدل‌های لجستیک و بیز ساده در این نمودار، بازخوانی و دقت پایین‌تری را نشان داده‌اند که نشان‌دهنده ضعف در شناسایی سکته مغزی است. نموداری مطلوب است که در آن منحنی به گوشه بالا سمت راست نزدیک‌تر باشد و مساحت زیر منحنی بزرگ‌تر باشد.









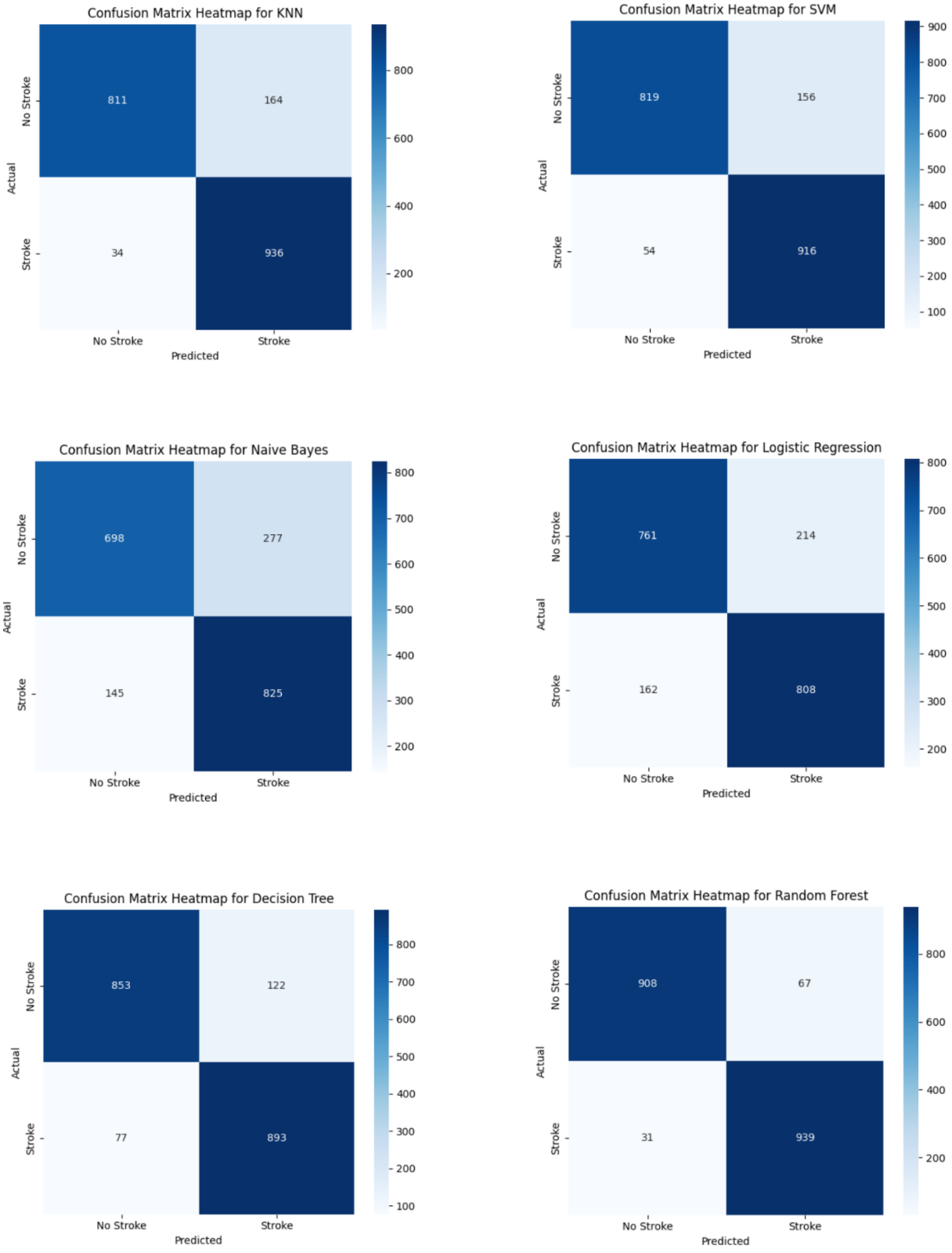
شکل ۲. نمودارهای ROC و Precision-Recall برای ۶ الگوریتم مختلف

۳-۳. ماتریس درهم ریختگی (Confusion Matrix)

ماتریس‌های درهم‌ریختگی برای هر یک از شش مدل آموزشی، تعداد پیش‌بینی‌های درست و نادرست مدل را در دو کلاس مثبت و منفی نشان می‌دهند. محور افقی این ماتریس نشان‌دهنده مقادیر پیش‌بینی‌شده و محور عمودی آن مقادیر واقعی است. مقدارهای موجود در قطر اصلی ماتریس نمایانگر پیش‌بینی‌های صحیح و سایر مقادیر نشان‌دهنده خطاهای مدل هستند.

یک ماتریس درهم‌ریختگی مطلوب باید مقادیر بالایی در قطر اصلی و مقادیر کمی در خارج از آن داشته باشد. براساس این ماتریس‌ها، مدل جنگل تصادفی بیشترین تعداد پیش‌بینی صحیح را ارائه داده است، در حالی که مدل‌های لجستیک و بیز ساده بیشترین خطا را در پیش‌بینی‌ها نشان داده‌اند. این تحلیل نشان می‌دهد که مدل جنگل تصادفی با تعادل بیشتری بین

پیش‌بینی‌های درست مثبت و منفی، عملکرد بهتری داشته است. ماتریس‌هایی که حاوی مقادیر زیاد در خارج از قطر اصلی هستند، نشان‌دهنده ضعف مدل در شناسایی کلاس‌ها می‌باشند.



شکل ۳. نمودارهای ماتریس درهم ریختگی برای ۶ الگوریتم مختلف

۳-۴. معیارهای ارزیابی

برای ارزیابی دقت و عملکرد هر یک از مدل‌های طبقه‌بندی، از معیارهای ارزیابی مختلفی استفاده شده است که شامل دقت کلی، بازخوانی، دقت، امتیاز F1، میانگین کلان و میانگین وزنی می‌باشد. در جدول ۱، مقادیر این معیارها برای هر یک از مدل‌های طبقه‌بندی محاسبه و ثبت شده است.

معیار دقت کلی نسبت پیش‌بینی‌های صحیح به کل پیش‌بینی‌ها را نشان می‌دهد و معمولاً برای داده‌های متوازن کاربرد بیشتری دارد. بازخوانی اهمیت بالایی در شناسایی موارد مثبت واقعی دارد و مدل‌هایی با بازخوانی بالا توانایی بیشتری در کشف موارد مثبت نشان می‌دهند. دقت نشان‌دهنده نسبت پیش‌بینی‌های مثبت درست به کل پیش‌بینی‌های مثبت است و برای داده‌های با اهمیت کلاس مثبت مورد استفاده قرار می‌گیرد. همچنین، امتیاز F1 میانگین هارمونیک دقت و بازخوانی است که توازنی مناسب بین این دو معیار ایجاد می‌کند.

برای داده‌های نامتوازن، میانگین کلان که تمامی کلاس‌ها را به طور مساوی وزن می‌دهد و میانگین وزنی که وزن هر کلاس را براساس تعداد نمونه‌های آن تنظیم می‌کند، بسیار مفید هستند. هرچه مقادیر این معیارها به مقدار یک نزدیک‌تر باشد، نشان‌دهنده عملکرد بهتر مدل است.

بر اساس نتایج به‌دست‌آمده، مدل جنگل تصادفی با بالاترین دقت، بازخوانی و امتیاز F1 بهترین عملکرد را در پیش‌بینی سکته مغزی ارائه داده است. در مقابل، مدل‌های لجستیک و بیز ساده عملکرد ضعیف‌تری داشته‌اند و در شناسایی کلاس‌های مثبت خطای بیشتری داشته‌اند. بنابراین، مدل جنگل تصادفی به‌عنوان مناسب‌ترین مدل برای پیش‌بینی سکته مغزی انتخاب شده است.

جدول ۱. مقادیر معیارها برای ۶ الگوریتم مختلف

Algorithm	Precision (0)	Precision (1)	Recall (0)	Recall (1)	F1-Score (0)	F1-Score (1)
Random Forest	0.97	0.93	0.93	0.97	0.95	0.95
SVM	0.94	0.85	0.84	0.94	0.89	0.90
KNN	0.96	0.85	0.83	0.96	0.89	0.90
Decision Tree	0.92	0.88	0.87	0.92	0.90	0.90
Logistic Regression	0.82	0.79	0.78	0.83	0.80	0.81
Naïve Bayes	0.83	0.75	0.72	0.85	0.77	0.80

۳-۵. مقایسه مدل‌های مختلف

در این بخش، با توجه به اینکه در داده‌های متوازن، معیار Accuracy انتخاب مناسبی می‌باشد، زیرا توانایی مدل در پیش‌بینی‌های کلی را به خوبی منعکس می‌کند. اما در داده‌های نامتوازن، معیار AUC ترجیح داده می‌شود زیرا توانایی مدل در تفکیک کلاس‌ها را به دقت ارزیابی می‌کند و تأثیر توزیع نامتوازن داده‌ها را کاهش می‌دهد و به‌طور کلی در مواردی که حساسیت در شناسایی موارد مثبت اهمیت دارد، معیار دقیق‌تر و قابل‌اعتمادتری محسوب می‌شود، مدل‌های مختلف بر اساس دو معیار اصلی، یعنی AUC و Accuracy، از مقدار بیشتر به کمتر در جداول ۲ و ۳ مرتب شدند و برای ارزیابی عملکردشان مورد بررسی قرار گرفتند. که مدل Random Forest بهترین و مدل Naïve Bayes بدترین مقدار را داشتند. بنابراین با توجه به نتایج، بهترین مدل جهت پیش بینی وقوع سکته مغزی مدل Random Forest انتخاب می‌شود.

جدول ۲. رتبه‌بندی الگوریتم‌ها بر اساس معیار Accuracy

Rank	Algorithm	Accuracy
1	Random Forest	0.9496
2	KNN	0.8982
3	Decision Tree	0.8977
4	SVM	0.8920
5	Logistic	0.8067
6	Naïve Bayes	0.7830

جدول ۳. رتبه‌بندی الگوریتم‌ها بر اساس معیار AUC

Rank	Algorithm	AUC
1	Random Forest	0.9902
2	KNN	0.9737
3	SVM	0.9445
4	Decision Tree	0.8946
5	Logistic	0.8917
6	Naïve Bayes	0.8521

۴. نتیجه‌گیری

در این پژوهش، از داده‌های مربوط به سکته مغزی که شامل مجموعه‌ای متوازن‌سازی‌شده از نمونه‌های مثبت و منفی است، برای پیش‌بینی وقوع سکته بهره گرفته‌ایم. مدل‌های طبقه‌بندی ارزیابی‌شده شامل رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده، ماشین بردار پشتیبان و K-نزدیک‌ترین همسایه بوده‌اند. نتایج نشان دادند که جنگل تصادفی با بالاترین دقت، بازخوانی و امتیاز F1 عملکرد بهتری نسبت به سایر مدل‌ها داشته و توانسته است تعادل مطلوبی بین شناسایی موارد مثبت و کاهش خطاهای پیش‌بینی ایجاد کند.

علاوه بر این، نیز مدل K-نزدیک‌ترین همسایه نتایج قابل‌قبولی ارائه داد و از دقت و بازخوانی مناسبی برخوردار بود. در مقابل، مدل‌های لجستیک و بیز ساده در شناسایی موارد مثبت خطای بیشتری داشته‌اند.

این تحلیل‌ها نشان می‌دهند که جنگل تصادفی به دلیل توانایی بالا در مدیریت داده‌های پیچیده و غیرخطی، بهترین گزینه برای پیش‌بینی سکته مغزی بوده و استفاده از آن در کاربردهای عملی می‌تواند نتایج دقیقی ارائه دهد.

منابع

مهدی پور، یوسف، ابراهیمی، سعید، کریمی، افسانه، علی پور، جهان پور، خمرنیا، محمد، و سیاسر، فاطمه. (۱۳۹۵). ارائه مدل پیش بینی سکته مغزی با استفاده از الگوریتم داده کاوی. علوم پزشکی صدرا، ۴(۴)، ۲۵۵-۲۶۵.
https://sid.ir/paper/238865/fa

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. Remote Sensing, 13(22), 4712.

Zhou, X. H., Obuchowski, N. A., & McClish, D. K. (2014). Statistical methods in diagnostic medicine. John Wiley & Sons.

Rajula, H. S. R., Verlato, G., Manchia, M., Antonucci, N., & Fanos, V. (2020). Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina, 56(9), 455.

LaValley, M. P. (2008). Logistic regression. Circulation, 117(18), 2395-2399.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

Webb, G. I., Keogh, E., & Miikkulainen, R. (2010). Naïve Bayes. Encyclopedia of machine learning, 15(1), 713-714.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46).

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101-121). Academic Press.

Peterson, L. E. (2009). K-nearest neighbor. Scholarpedia, 4(2), 1883.

King, R. D., Feng, C., & Sutherland, A. (1995). Statlog: comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence an International Journal, 9(3), 289-333.

Holloway, R. G., Benesch, C. G., Burgin, W. S., & Zentner, J. B. (2005). Prognosis and decision making in severe stroke. Jama, 294(6), 725-733.

Bonkhoff, A. K., & Grefkes, C. (2022). Precision medicine in stroke: towards personalized outcome predictions using artificial intelligence. Brain, 145(2), 457-475.

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data