The 9th International Conference on Web Research
ICWR2023
May 3th-4th, 2023, Tehran, Iran

Classification of Web Pages Based on Search Engine Optimization Quality using Random Forest Algorithm

[1]Mohammad Hosein Shariatipour*, [2]Mohammad Rahmanimanesh, [3]Mohammad Javad Parseh

[1]Masters student, Computer Engineering, Semnan University, Semnan
mhshariatipour@semnan.ac.ir

[2]Assistant Professor, Computer Engineering, Semnan University, Semnan
rahmanimanesh@semnan.ac.ir

[3]Instructor, Computer Engineering, Jahrom University, Jahrom
parseh@jahromu.ac.ir

Abstract

In the evening where ranking on the search engine result pages is directly related to increasing the number of visitors and the progress and development of a business, search engine optimization or SEO is a process that helps to gain a higher ranking. Websites can be classified with the help of machine learning techniques based on the quality of setting SEO guidelines. Classification algorithms are combined with each other with the aim of increasing classification accuracy and are used as an ensemble classification model. In this article, we implement an ensemble classification model with the help of a random forest algorithm, which places web pages in one of the predefined classes based on SEO quality. The obtained results show that the accuracy of the constructed model is between 70.50% and 73.17% and is more accurate than previous works in which ensemble classification algorithms were not used. The built model can help developers build automatic software for detecting the SEO quality of web pages.

Keywords: Search Engine Optimization, Machine Learning, Ensemble Learning, Web Page Classification, Random Forest Algorithm.

The 9th International Conference on Web Research
ICWR2023
May 3th-4th, 2023, Tehran, Iran

طبقه‌بندی صفحه‌های وب بر اساس کیفیت بهینه‌سازی موتور جستجو به کمک الگوریتم جنگل تصادفی

محمدحسین شریعتی‌پور*[1]، محمد رحمانی‌منش[2]، محمدجواد پارسه[3]

[1] دانشجوی کارشناسی ارشد، مهندسی کامپیوتر، دانشگاه سمنان، سمنان
mhshariatipour@semnan.ac.ir
[2] استادیار، مهندسی کامپیوتر، دانشگاه سمنان، سمنان
rahmanimanesh@semnan.ac.ir
[3] مربی، مهندسی کامپیوتر، دانشگاه جهرم ، جهرم
parseh@jahromu.ac.ir

چکیده

در عصری که رتبه‌بندی در صفحات نتایج موتور جستجو ارتباط مستقیمی با افزایش تعداد بازدیدکنندگان و درنتیجه آن پیشرفت و توسعه یک کسب‌وکار دارد، بهینه‌سازی موتور جستجو یا سئو فرآیندی است که به کسب رتبه بالاتر کمک می‌کند. وب‌سایت‌ها را می‌توان به کمک تکنیک‌های یادگیری ماشین بر اساس کیفیت تنظیم دستورالعمل‌های سئو طبقه‌بندی کرد. الگوریتم‌های طبقه‌بندی باهدف افزایش دقت طبقه‌بندی با یکدیگر ترکیب می‌شوند و به‌عنوان یک مدل طبقه‌بندی ترکیبی استفاده می‌شوند. در این مقاله یک مدل طبقه‌بندی ترکیبی را به کمک الگوریتم جنگل تصادفی پیاده‌سازی می‌کنیم که صفحات وب را در یکی از طبقه‌بندهای از پیش تعریف‌شده بر اساس کیفیت سئو قرار می‌دهد. نتایج به‌دست‌آمده نشان می‌دهد که دقت مدل ساخته‌شده بین ۷۰/۵۰٪ تا ۷۳/۱۳٪ است و نسبت به کارهای قبلی که در آن‌ها از الگوریتم‌های طبقه‌بندی ترکیبی استفاده‌نشده است دقت بالاتری دارد. مدل ساخته‌شده می‌تواند به توسعه‌دهندگان نرم‌افزار برای ساخت نرم‌افزارهای خودکار تشخیص کیفیت سئو صفحات وب کمک کند.

کلمات کلیدی

بهینه‌سازی موتور جستجو، یادگیری ماشین، یادگیری ترکیبی، طبقه‌بندی صفحات وب، الگوریتم جنگل تصادفی

۱- مقدمه

درحالی‌که دیجیتالی شدن کسب‌وکارها در حال رشد و پیشرفت است، از وب‌سایت‌ها می‌توان به‌عنوان ابزاری برای گسترش آگاهی و توجه به خدمات خود در شبکه اینترنت استفاده کرد [1]. بیش از نیمی از ترافیک یک وب‌سایت توسط موتورهای جستجو به آن هدایت می‌شود و سهم گوگل از بازار موتورهای جستجو تقریباً ۹۲/۴۲٪ است، به همین دلیل رتبه‌بندی در موتور جستجو گوگل از اهمیت ویژه‌ای برخوردار است [2].

اغلب کاربران فقط از صفحاتی که در صفحه اول نتایج موتور جستجو ارائه می‌شود بازدید می‌کنند و نتایج در صفحات بعدی شانس کمتری برای بازدید دارند بنابراین، صفحات وب با رتبه بالا بازدیدکنندگان بیشتری داشته و همین امر، بالا بردن رتبه‌بندی صفحات را به اولویت اصلی مدیران وب تبدیل کرده است [3]. بهینه‌سازی موتور جستجو[1] یا سئو به فعالیت بهینه‌سازی وب‌سایت‌ها و صفحات وب برای کسب رتبه بالاتر در صفحات نتایج موتور جستجو مربوط می‌شود [4]. در عصر حاضر، بهینه‌سازی موتور جستجو برای کسب‌وکارها امری ضروری است تا بتوانند در فضای رقابتی وب باقی بمانند.

مطالعات بسیاری در ارتباط با فاکتورهای بهینه‌سازی موتور جستجو و تأثیر آن بر کسب رتبه بالاتر در نتایج موتور جستجو انجام‌گرفته است [1, 4-6]. به‌طورکلی این فاکتورها به دو دسته درون صفحه و خارج صفحه تقسیم‌بندی می‌شوند. عنوان صفحه، بهینه‌سازی فایل‌های گرافیکی، نام دامنه، استفاده از متاتگ‌ها و کلیه فاکتورهایی که تحت کنترل کامل مدیر وب‌سایت قرار دارند، فاکتورهای درون صفحه و ساخت بک لینک‌ها، فعالیت در رسانه‌های اجتماعی و به‌طورکلی عواملی که مدیر وب‌سایت کنترل کمتری بر روی آن‌ها دارد فاکتورهای خارج از صفحه در فرآیند بهینه‌سازی موتور جستجو محسوب می‌شوند.

امروزه طبقه‌بندی[2] صفحات وب به‌منظور اهداف مختلفی به کار می‌رود، طبقه‌بندی صفحات وب بر اساس کیفیت سئو یکی از موضوعاتی است که اخیراً موردتوجه قرارگرفته است. الگوریتم‌های طبقه‌بندی ویژگی‌های متفاوتی

دارند و عملکرد آن‌ها تحت تأثیر عوامل مختلفی ازجمله مجموعه داده‌ها است. همچنین الگوریتم‌های طبقه‌بندی را می‌توان با یکدیگر ترکیب کرده و به‌عنوان یک مدل طبقه‌بندی ترکیبی از آن استفاده کرد.

در پژوهش پیش رو از فاکتورهای درون صفحه و تکنیک‌های یادگیری ماشین برای ساخت یک مدل طبقه‌بندی استفاده می‌شود که به‌طور خودکار صفحات وب را بر اساس میزان تنظیم دستورالعمل‌های سئو طبقه‌بندی کند. به‌طور ویژه هدف در این پژوهش استفاده از الگوریتم طبقه‌بندی جنگل‌تصادفی[3] به‌عنوان یک الگوریتم یادگیری ترکیبی و مقایسه دقت به‌دست‌آمده با دقت سایر الگوریتم‌های طبقه‌بندی که در پژوهش‌های پیشین استفاده‌شده است خواهد بود. از مدل طبقه‌بندی ترکیبی ایجادشده می‌توان به‌عنوان ابزاری برای تشخیص صفحاتی که نیاز به بهینه‌سازی موتور جستجو دارند استفاده کرد.

ادامه این مقاله به شرح زیر سازمان‌دهی شده است. در بخش ۲ پیشینه پژوهش را بررسی خواهیم کرد. بخش ۳ مفهوم یادگیری‌ترکیبی[4] را تشریح می‌کند. بخش ۴ به معرفی مجموعه داده‌ها می‌پردازد. بخش ۵ روش پیشنهادی را توضیح می‌دهد که چگونه یک مدل طبقه‌بندی ترکیبی مبتنی بر الگوریتم جنگل‌تصادفی پیاده‌سازی کنیم تا به کمک آن صفحات وب را بر اساس کیفیت سئو طبقه‌بندی کند. نتایج ارزیابی نیز در بخش ۶ ارائه می‌شود.

۲- کارهای مرتبط

در [7] نویسندگان عوامل تأثیرگذار سئو بر رتبه‌بندی نتایج موتور جستجو را در پژوهش‌های پیشین مطالعه کرده و ۲۴ عامل را به‌عنوان مهم‌ترین عوامل معرفی کرده‌اند. ازجمله این موارد کیفیت بک لینک، کلمه کلیدی در تگ عنوان سایت، زمان بارگذاری سایت و چگالی کلمات کلیدی است.

استفاده از روش‌های تصمیم‌گیری برای بررسی رابطه متقابل و وزن‌های تأثیرگذار در میان معیارهای سئو در [8] نشان می‌دهد متاتگ‌ها مهم‌ترین معیار در سئو هستند و پس‌ازآن کلمات کلیدی، طراحی وب‌سایت، لینک سازی و رسانه‌های اجتماعی قرار دارند. کارشناسان پیشنهاد می‌کنند که مدیران وب‌سایت‌ها بیشترین تأکید را بر متاتگ‌ها داشته باشند.

طبقه‌بندی صفحات وب برای اهداف مختلف یک زمینه مطالعاتی گسترده در چند سال اخیر بوده است.

وب‌سایت‌های بدافزار و کرک اغلب آدرس اینترنتی خود را برای جلوگیری از شناسایی خودکار تغییر می‌دهند. بااین‌حال، در بسیاری از موارد، این وب‌سایت‌ها طرح‌های بصری خاصی مانند رنگ‌های خاص، فونت‌ها، اشکال و اندازه‌ها را حفظ می‌کنند که می‌تواند معرف عملکرد وب‌سایت به کاربران باشد. نویسندگان در مقاله [9] از ویژگی‌های طراحی بصری و غیر بصری وب‌سایت‌ها برای شناسایی دسته‌بندی وب‌سایت، به‌ویژه وب‌سایت‌های بدافزار و کرک استفاده کردند و وب‌سایت‌های مخرب با نرخ مثبت واقعی نسبتاً بالا و نرخ مثبت کاذب ناچیزی شناسایی شدند. نتایج نشان می‌دهد استفاده از روش‌های طبقه‌بندی که بر ویژگی‌های طراحی متکی است، می‌تواند به‌عنوان یک جایگزین مؤثر برای متن‌کاوی در دسته‌بندی وب‌سایت‌ها باشد که اغلب فرآیندی زمان‌بر و نیاز به محاسبات پیچیده دارد.

عملکرد الگوریتم‌های طبقه‌بندی یادگیری ماشین؛ k نزدیک‌ترین همسایه[5]، ماشین بردار پشتیبان و جنگل‌تصادفی برای تشخیص صفحات فیشینگ در پژوهش [10] ارزیابی‌شده است، این الگوریتم‌ها بر روی مجموعه داده‌های متشکل از ۱۰۰۰ صفحه فیشینگ و ۴۰۰ صفحه قانونی که هر یک از صفحات شامل ۱۴ ویژگی است پیاده‌سازی شد و نتایج نشان داد الگوریتم طبقه‌بندی جنگل‌تصادفی با دقت ۹۸/۳۵٪ نسبت به دو مدل دیگر عملکرد بهتری دارد.

از رویکردهای یادگیری ماشین برای شناسایی کیفیت سئو یک صفحه وب در [11] استفاده شد و ۵ الگوریتم طبقه‌بندی درخت تصمیم، k نزدیک‌ترین همسایه، ماشین بردار پشتیبان، طبقه‌بندی ساده بیز[6] و رگرسیون لجستیک[7] بر روی مجموعه‌ای از ۶۰۰ صفحه وب که کیفیت سئو آن‌ها توسط متخصصان سئو مشخص‌شده بود آموزش داده شدند. ارزیابی‌ها نشان داد این الگوریتم‌های طبقه‌بندی با دقت ۵۴/۶۹٪ تا ۶۹/۶۷٪ نسبت به دقت پایه که مربوط به طبقه‌بند اکثریت (۴۸/۸۳٪) است، دقت بالاتری دارند.

یک مدل طبقه‌بندی صفحات وب بر اساس ترکیبی از طبقه‌بندی کننده‌های متعدد در مقاله [12] پیشنهادشده است که در آن از ساختار درخت مانند تگ‌های HTML برای مشخص کردن ویژگی‌های ساختاری یک صفحه وب استفاده می‌شود، این ویژگی‌های ساختاری به همراه ویژگی‌های متنی صفحه به بردار تبدیل‌شده و باهم ترکیب می‌شوند. ارزیابی مدل پیشنهادی بر روی مجموعه داده‌های عمومی موجود به‌دقت قابل قبولی رسیده است و نتایج اثبات می‌کند دقت طبقه‌بندی صفحات وب با ترکیب چند طبقه‌بند متعدد بهبود می‌یابد.

۳- یادگیری ترکیبی

یادگیری‌ترکیبی یک الگوی موفق از یادگیری ماشین است که مجموعه‌ای از یادگیرندگان را به‌جای استفاده از یک یادگیرنده واحد برای پیش‌بینی ویژگی نمونه جدید ادغام می‌کند. در این ساختار، تمام مقادیر خروجی به‌دست‌آمده از هر یادگیرنده با استفاده از یک مکانیسم رأی‌گیری برای پیش‌بینی برچسب کلاس نهایی ترکیب می‌شود. هدف اصلی یادگیری‌ترکیبی تشکیل یک طبقه‌بندی کننده قوی متشکل از چند یادگیرنده برای به دست آوردن نتایج طبقه‌بندی دقیق‌تر است [13].

۳-۱- Bagging

Bagging یکی از پرکاربردترین روش‌های یادگیری‌ترکیبی است که از روش نمونه‌گیری مجدد Bootsratp استفاده می‌کند. در این روش مجموعه آموزشی اولیه با چند زیرمجموعه از داده‌های اولیه (آزمایش‌های Bootsratp) در طول فرآیند Bagging جایگزین می‌شود. در هر یک از این زیرمجموعه‌ها، برخی از مقادیر داده‌ها می‌توانند مکرر ظاهر شوند این در حالی است که برخی دیگر ممکن است در هیچ‌یک از زیرمجموعه‌ها عضو نباشند. سپس نتیجه نهایی با میانگین‌گیری نتایج هر زیرمجموعه تعیین می‌شود [14].

۱-۱-۳- جنگل‌تصادفی

الگوریتم جنگل‌تصادفی، نمونه‌ای از یک الگوریتم Bagging است که جنگلی با درخت‌های تصمیم‌گیری متعدد ایجاد می‌کند. ایده اصلی الگوریتم جنگل‌تصادفی استفاده از روش نمونه‌برداری Bootsratp برای انجام عملیات نمونه‌برداری با جایگزینی از مجموعه داده اصلی و سپس ساختن درخت تصمیم از زیرمجموعه داده‌های نمونه‌برداری شده و ترکیب درخت‌های تصمیم متعدد در یک جنگل‌تصادفی است [15].

هر درخت تصمیم در الگوریتم جنگل‌تصادفی مطابق شکل(۱) شامل یک دنباله درخت از گره‌های تصمیم است. بر اساس این توالی، درخت به شاخه‌های مختلف تقسیم می‌شود تا به برگ‌های درخت برسد. نتایج پیش‌بینی هر درخت تصمیم از طریق گره‌های برگ مشخص می‌شود و درنهایت خروجی‌های درخت‌های تصمیم چندگانه برای پیش‌بینی ترکیب می‌شوند و بر اساس اکثریت نتایج به‌دست‌آمده از درخت‌های تصمیم چندگانه، پیش‌بینی نهایی جنگل تصادفی نتیجه می‌شود.

وجود درختان تصمیم متفاوت در الگوریتم جنگل‌تصادفی به کاهش مشکل بیش برازش[۸] کمک می‌کند، مشکلی که مخصوص الگوریتم درخت تصمیم است [10].

۴- مجموعه داده‌ها

در کار پیش رو از مجموعه داده‌های [16] جمع‌آوری‌شده در پژوهش [11] استفاده خواهیم کرد. این مجموعه داده شامل ۶۰۰ نمونه تصادفی از صفحات وب است که هر یک از صفحات توسط ۲۱ متغیر مستقل به‌عنوان فاکتورهای مؤثر بر سئو داخلی که در جدول (۱) قابل‌مشاهده است و یک متغیر وابسته به‌عنوان کیفیت سئو صفحه توصیف می‌شود. کیفیت سئو صفحات توسط متخصصان بهینه‌سازی موتور جستجو دریکی از دسته‌های ضعیف، متوسط و قوی قرارگرفته است.

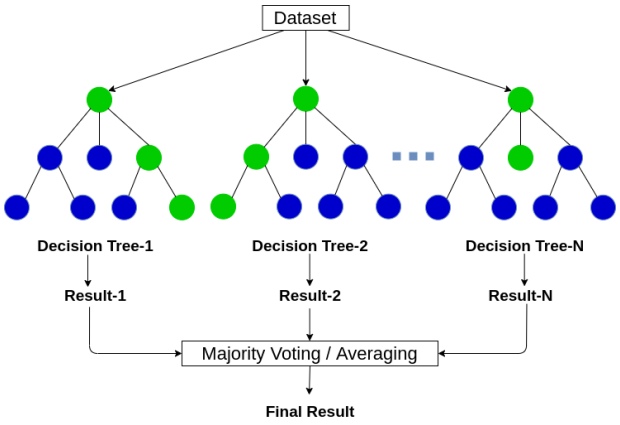
۵- روش پژوهش

هدف از این پژوهش ساخت مدل طبقه‌بندی کننده ترکیبی است که صفحات وب را بر اساس کیفت سئو دریکی از دسته‌های قوی، متوسط و ضعیف قرار دهد. از الگوریتم جنگل‌تصادفی به‌عنوان یک الگوریتم یادگیری‌ترکیبی از نوع Bagging برای ساخت مدل پیشنهادی استفاده خواهیم کرد.

مراحل ساخت مدل طبقه‌بندی و ارزیابی آن در ۴ مرحله دنبال می‌شود.

۱-۵- نرمال‌سازی داده‌ها

نرمال‌سازی معمولاً فرآیندی است که باهدف بهبود دقت و عملکرد مدل‌های طبقه‌بندی بر روی داده‌های عددی انجام می‌شود تا مجموعه داده به یک بازه خاص، معمولاً بین ۰ و ۱ محدود شود. نرمال‌سازی در مواردی که داده‌های عددی دارای محدوده‌های مختلف باشند ضروری است. ازآنجایی‌که کلیه ۲۱ ویژگی مستقل موجود در مجموعه داده‌ها از نوع عددی هستند ابتدا لازم است مقادیر را نرمال کنیم.



شکل(۱): الگوریتم جنگل‌تصادفی

جدول (۱): فاکتورهای مؤثر بر SEO داخلی

عنوان	توضیحات
Tlen	طول محتوا در تگ <title> (تعداد کلمات)
Tkw	فراوانی کلمات کلیدی در تگ <title>
Mlen	طول محتوا در تگ <meta> توضیحات
Mkw	فراوانی کلمات کلیدی در تگ <meta> توضیحات
h1	تعداد تگ <h1>
h1len	میانگین طول محتوا تگ‌های <h1>
h1kw	فراوانی کلمات کلیدی در تگ‌های <h1>
h2	تعداد تگ <h2>
h2len	میانگین طول محتوا تگ‌های <h2>
h2kw	فراوانی کلمات کلیدی در تگ‌های <h2>
h3	تعداد تگ <h2>
h3len	میانگین طول محتوا تگ‌های <h2>
h3kw	فراوانی کلمات کلیدی در تگ‌های <h2>
h3kw	تعداد تگ‌های <img> که حاوی ویژگی alt هستند
altkw	فراوانی کلمات کلیدی در ویژگی alt تگ‌های <img>
linkkw	تعداد فراوانی کلمه کلیدی در متن anchor
linkout	تعداد پیوندهای خروجی
urllen	طول آدرس اینترنتی (url)
urlkw	فراوانی کلمات کلیدی در آدرس اینترنتی (url)
txtlen	طول متن بدنه صفحه
txtkw	تعداد کلمات کلیدی در متن بدنه صفحه

نرمال‌سازی MinMax تکنیکی است که از آن استفاده خواهیم کرد و به کمک آن کلیه مقادیر عددی را محدود به بازه ۰ تا ۱ می‌کنیم. نحوه محاسبه مقدار نرمال متغیر X به کمک فرمول (۱) محاسبه می‌شود.

$$X\_norm = \frac{\mathrm{X} - \mathrm{X\_min}}{X\_\mathrm{max} - X\_min} \quad (۱)$$

که در آن، X_norm نشان‌دهنده مقدار نرمال شده، X_min و X_max به ترتیب حداقل و حداکثر مقادیر داده‌ها هستند و X نشان‌دهنده مقدار اصلی داده است.

۵-۲- تنظیم فراپارامترها

فراپارامترها مقادیری هستند که قبل از شروع فرآیند یادگیری باهدف بهینه‌سازی مدل [17] و تطبیق یک مدل یادگیری ماشین در مسائل مختلف تنظیم می‌شوند [18]. انتخاب فراپارامترهای مناسب برای مدل‌های یادگیری ماشین تأثیر مستقیمی بر عملکرد مدل دارد که اغلب به دانش عمیق از الگوریتم‌های یادگیری ماشین و تکنیک‌های بهینه‌سازی فراپارامترها نیاز دارد.

تعداد درختان (n_estimators) و حداقل نمونه‌های موردنیاز در یک گره‌ی برگ (min_sample_leaf) ازجمله فراپارمترهای پرکاربرد در الگوریتم جنگل‌تصادفی هستند که از آن‌ها برای ساخت مدل طبقه‌بندی خود استفاده خواهیم کرد.

۵-۳- ساخت مدل طبقه‌بندی

در کار پیش رو از الگوریتم جنگل‌تصادفی موجود در کتابخانه Scikit Learn برای ساخت مدل طبقه‌بندی استفاده می‌کنیم، Scikit Learn یک کتابخانه پرکاربرد زبان برنامه‌نویسی پایتون برای اهداف یادگیری ماشین است که ابزارهای کاربردی زیادی به‌منظور یادگیری ماشین و مدل‌سازی آماری داده‌ها همچون طبقه‌بندی و خوشه‌بندی فراهم می‌کند.

مقدار فراپارامتر تعداد درختان را در مدل خود برابر ۱۰۰ و حداقل نمونه‌های موردنیاز در یک گره‌ی برگ را ۱۶ در نظر می‌گیریم، شکل (۲) نحوه تعریف مدل طبقه‌بندی جنگل تصادفی با این مقادیر به زبان پایتون را نشان می‌دهد. مدل تعریف‌شده را با ۲۱ فاکتور مؤثر بر سئو داخلی که در مجموعه داده‌ها وجود دارد آموزش می‌دهیم و انتظار داریم این مدل پس از طی کردن مرحله آموزش بتواند کیفیت سئوی نمونه وب‌سایت‌های جدید را با دقت قابل قبولی پیش‌بینی کند.

۵-۴- ارزیابی مدل طبقه‌بندی

ازآنجایی‌که عوامل مختلفی ازجمله دامنه داده‌ها، ویژگی آن‌ها و فرضیات الگوریتم بر روی الگوریتم‌های طبقه‌بندی تأثیرگذار است انتخاب بهترین الگوریتم طبقه‌بندی برای یک کار خاص نیاز به ارزیابی الگوریتم‌های مختلف دارد. دو تکنیک رایج برای ارزیابی الگوریتم‌های طبقه‌بندی وجود دارد.

```python
from sklearn.ensemble import RandomForestClassifier

clf_model=RandomForestClassifier(
  n_estimators = 100,
  min_samples_leaf = 16
)
```

شکل (۲): تعریف مدل طبقه‌بندی جنگل تصادفی

اعتبارسنجی Hold-out تکنیکی است که در آن مجموعه دادها به دو بخش آموزش و آزمون تقسیم می‌شود. مجموعه آموزش برای آموزش مدل یادگیری و مجموعه آزمون برای اعتبارسنجی عملکرد مدل ایجادشده. اما این تکنیک در مواردی که مجموعه داده کوچک باشد و یا اگر مجموعه داده آموزش انتخاب‌شده معرف خوبی برای کل داده‌ها نباشد ممکن است عملکرد مطلوبی نداشته باشد.

اعتبارسنجی متقاطع K-fold تکنیک دیگری است که در یادگیری ماشین برای ارزیابی عملکرد یک مدل استفاده می‌شود و محدودیت‌های تکنیک قبل را برطرف می‌کند. در این روش مجموعه داده‌ها به k قسمت مساوی تقسیم می‌شود. سپس مدل یادگیری بر روی K-1 قسمت مرحله آموزش را طی می‌کند و پس‌ازآن با یک قسمت باقیمانده اعتبارسنجی مدل سنجیده می‌شود. این فرآیند k بار تکرار می‌شود و هر بار از یک قسمت متفاوت به‌عنوان مجموعه آزمون استفاده می‌کنیم. در آخر با میانگین‌گیری از k مرحله انجام‌شده ارزیابی انجام خواهد شد.

از معیار دقت نیز برای ارزیابی مدل طبقه‌بندی استفاده خواهیم کرد، این معیار نشان می‌دهد چه تعداد از نمونه‌ها در مرحله آزمون به‌درستی طبقه‌بندی‌شده‌اند. برای محاسبه دقت، تعداد داده‌های درست طبقه‌بندی‌شده بر تعداد کل داده‌ها تقسیم می‌شود.

۶- نتایج

ارزیابی مدل طبقه‌بندی توسط هر دو تکنیک Hold-out و K-fold انجام خواهد شد. در تکنیک Hold-out از دوسوم داده‌ها برای مرحله آموزش و از یک‌سوم باقیمانده برای آزمون استفاده می‌شود. در روش K-fold نیز مقدار k را برابر ۱۰ در نظر می‌گیریم.

دقت ۵ الگوریتم طبقه‌بندی درخت تصمیم، k نزدیک‌ترین همسایه، ماشین بردار پشتیبان، طبقه‌بندی ساده بیز و رگرسیون لجستیک که در پژوهش [۱۱] به‌دست‌آمده است به همراه الگوریتم طبقه‌بندی ترکیبی جنگل‌تصادفی ارائه‌شده در این پژوهش به دو روش Hold-out و K-fold به ترتیب در جدول (۲) و جدول (۳) آورده شده.

حداکثر دقت به‌دست‌آمده در پژوهش [11] در حالت Hold-out مربوط به الگوریتم طبقه‌بندی درخت تصمیم با مقدار ۶۵/۶۷٪ و در حالت K-fold حداکثر دقت مربوط به الگوریتم k نزدیک‌ترین همسایه با مقدار ۶۹/۶۷٪ درصد است، این در حالی است که دقت‌های به‌دست‌آمده در مدل پیشنهادی ما به ترتیب برابر ۷۰/۵۰٪ و ۷۳/۱۷٪ است.

جدول (۲): دقت در تکنیک Hold-out

الگوریتم	فراپارامتر	دقت
Decision trees	C = 0.49 min_samples_leaf=16	۶۵٪/۶۷
Naïve Bayes	–	۵۸٪/۷۱
KNN	K = 45, p = 2	۶۵٪/۱۷
SVM	$C = 7.74 \times 10^3$ σ = 0.000464	۶۲٪/۶۸
Logistic regression	C = 2	۶۲٪/۱۹
Random forest	n_estimators=100 min_samples_leaf=16	۷۰٪/۵۰

جدول (۳): دقت در تکنیک K-fold

الگوریتم	فراپارامتر	دقت
Decision trees	C = 0.49 min_samples_leaf=16	۶۷٪/۵۳
Naïve Bayes	–	۵۴٪/۶۹
KNN	K = 45, p = 2	۶۹٪/۶۷
SVM	$C = 7.74 \times 10^3$ σ = 0.000464	۶۶٪/۱۸
Logistic regression	C = 2	۶۲٪/۹۹
Random forest	n_estimators=100 min_samples_leaf=16	۷۳٪/۱۷

از نتایج به‌دست‌آمده می‌توان نتیجه گرفت استفاده از مدل یادگیری‌ترکیبی می‌تواند دقت را در طبقه‌بندی صفحات وب بر اساس کیفیت بهینه‌سازی موتور جستجو بهبود دهد. از مدل طبقه‌بندی ترکیبی ساخته‌شده می‌توان برای ساخت نرم‌افزارهای خودکار پیش‌بینی کیفیت سئو صفحات وب یا به‌عنوان ابزاری برای شناسایی صفحاتی که نیاز به بهبود کیفیت سئو دارند استفاده کرد.

مجموعه داده‌های استفاده‌شده در کار ما شامل ۲۱ عامل مؤثر بر سئو داخلی صفحات بود این درحالی‌که است که عوامل خارجی نیز تأثیر قابل‌توجه‌ای بر کیفیت سئو صفحات دارند بنابراین در کارهای آینده لازم است از ترکیب عوامل داخلی و خارجی برای ارزیابی کیفیت سئو صفحات وب استفاده شود. استفاده از سایر الگوریتم‌های طبقه‌بندی ترکیبی نیز می‌تواند یکی دیگر از کارهای پیش رو باشد.

مراجع

[1] I. Drivas, D. Kouis, D. Kyriaki-Manessi, and G. Giannakopoulos, "Content management systems performance and compliance assessment based on a data-driven search engine optimization methodology," *Information*, vol. 12, no. 7, p. 259, 2021.

[2] M. Vállez and A. Ventura, "Analysis of the SEO visibility of university libraries and how they impact the web visibility of their universities," *The Journal of Academic Librarianship*, vol. 46, no. 4, p. 102171, 2020.

[3] A.-J. Su, Y. C. Hu, A. Kuzmanovic, and C.-K .Koh, "How to improve your search engine ranking: Myths and reality," *ACM Transactions on the Web (TWEB)*, vol. 8, no. 2, pp. 1-25, 2014.

[4] M. Khan and A. Mahmood, "A distinctive approach to obtain higher page rank through search engine optimization," *Sādhanā*, vol. 43, no. 3, p. 43, 2018.

[5] S. An and J. J. Jung, "A heuristic approach on metadata recommendation for search engine optimization," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 3, p. e5407, 2021.

[6] A. Erdmann, R. Arilla ,and J. M. Ponzoa, "Search engine optimization: The long-term strategy of keyword choice," *Journal of Business Research*, vol. 144, pp. 650-662, 2022.

[7] C. Ziakis, M. Vlachopoulou, T. Kyrkoudis, and M. Karagkiozidou, "Important factors for improving Google search rank," *Future internet*, vol. 11, no. 2, p. 32, 2019.

[8] H.-J. Tsuei, W.-H. Tsai, F.-T. Pan, and G.-H. Tzeng, "Improving search engine optimization (SEO) by using hybrid modified MCDM models," *Artificial Intelligence Review*, vol. 53, pp. 1-16.۲۰۲۰ ,

[9] D. Cohen, O. Naim, E. Toch, and I. Ben-Gal, "Website categorization via design attribute learning," *Computers & Security*, vol. 107, p. 102312, 2021.

[10] T. O. Ojewumi, G. Ogunleye, B. Oguntunde, O. Folorunsho, S. Fashoto, and N. Ogbu, "Performance evaluation of machine learning tools for detection of phishing attacks on web pages," *Scientific African*, vol. 16, p. e01165, 2022.

[11] G. Matošević, J. Dobša, and D. Mladenić, "Using machine learning for web page classification in search engine optimization," *Future Internet*, vol. 13, no. 1, p. 9, 2021.

[12] L. Deng, X. Du, and J.-z. Shen, "Web page classification based on heterogeneous features and a combination of multiple classifiers," *Frontiers of Information Technology & Electronic Engineering* ,vol. 21, no. 7, pp. 995-1004, 2020.

[13] P. Y. Taser, "Application of bagging and boosting approaches using decision tree-based algorithms in diabetes risk prediction," in *Proceedings*, 2021, vol. 74, no. 1: MDPI, p. 6 .

[14] A. Aldrees, H. H. Awan, M. F Javed, and A. M. Mohamed, "Prediction of water quality indexes with ensemble learners: Bagging and Boosting," *Process Safety and Environmental Protection*, vol. 168, pp. 344-361, 2022.

[15] Z. Chen, Q. Wu, S. Han, J. Zhang, P. Yang, and X. Liu, "A study on geological structure prediction based on random forest method," *Artificial Intelligence in Geosciences*, vol. 3, pp. 226-236, 2022.

[16] Dataset: https://zenodo.org/record/4416123
[17] D. Sun, H. Wen, D. Wang, and J. Xu, "A random forest model of landslide susceptibility mapping based on hyperparameter optimization using Bayes algorithm," *Geomorphology*, vol. 362, p. 107201, 2020.
[18] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295-316, 2020.

زیرنویس ها

[1] Search engine optimization
[2] Classification
[3] Random Forest
[4] Ensemble Learning
[5] K-Nearest Neighbors
[6] Naive Bayes
[7] Logistic regression
[8] Overfitting