رویکرد هوش مصنوعی در تحلیل داده‌های سلامت

سعید اکبری
دانشجوی کارشناسی ارشد هوش مصنوعی و رباتیکز دانشگاه جامع امام حسین (ع)

محمدرضا حسنی آهنگر
استاد تمام دانشگاه جامع امام حسین (ع)

رامین دلیر
پژوهشگر دانشگاه جامع امام حسین (ع)

چکیده

امروزه، پیشرفت‌های هوش مصنوعی و یادگیری ماشین نقش مهمی در بهبود فرآیندهای تحلیل داده‌های سلامت ایفا می‌کنند. یکی از چالش‌های مهم در این حوزه، پیش‌بینی تأثیر عوامل مختلف بر هزینه‌های درمانی است. در این پژوهش، به بررسی رابطه بین استعمال سیگار و هزینه‌های درمانی با استفاده از چهار روش یادگیری ماشین شامل رگرسیون لجستیک، جنگل تصادفی، XGBoost و ماشین بردار پشتیبان [1] پرداخته شده است. داده‌های مورد استفاده شامل متغیرهایی نظیر سن، شاخص توده بدنی [2]، جنسیت، منطقه جغرافیایی، تعداد فرزندان و هزینه‌های درمانی هستند. ابتدا داده‌ها پیش‌پردازش شده و ویژگی‌های مؤثر انتخاب شده‌اند. سپس، مدل‌های مذکور آموزش داده شده و عملکرد آن‌ها بر اساس معیارهایی همچون R^2، MSE، MAE، AIC، BIC ارزیابی شده است. نتایج نشان می‌دهد که مدل XGBoost عملکرد بهتری نسبت به سایر مدل‌ها داشته و دقت بالاتری در پیش‌بینی رابطه بین استعمال سیگار و هزینه‌های درمانی دارد. این پژوهش نشان می‌دهد که استفاده از روش‌های یادگیری ماشین می‌تواند در تحلیل داده‌های سلامت و ارائه بینش‌های دقیق‌تر به متخصصان حوزه پزشکی مؤثر باشد.

واژگان کلیدی: یادگیری ماشین، تحلیل داده‌های سلامت، هزینه‌های درمانی، پیش‌بینی، هوش مصنوعی

[1] SVM
[2] BMI

مقدمه

در سال‌های اخیر، هوش مصنوعی و یادگیری ماشین به یکی از مهم‌ترین ابزارها در حوزه پزشکی و سلامت تبدیل شده‌اند. با افزایش حجم داده‌های سلامت، امکان تحلیل و استخراج اطلاعات ارزشمند از این داده‌ها فراهم شده است که می‌تواند به بهبود فرآیندهای تشخیصی و درمانی کمک کند. یکی از مهم‌ترین عوامل تأثیرگذار بر هزینه‌های درمانی، سبک زندگی و رفتارهای سلامت فردی است که استعمال سیگار به عنوان یکی از مهم‌ترین این عوامل شناخته می‌شود. بررسی رابطه بین استعمال سیگار و هزینه‌های درمانی از منظر داده‌کاوی و یادگیری ماشین، می‌تواند به پیش‌بینی دقیق‌تر هزینه‌های درمانی افراد سیگاری و غیرسیگاری کمک کند و در تصمیم‌گیری‌های کلان حوزه سلامت مؤثر باشد.

در این پژوهش، با استفاده از داده‌های پزشکی، تلاش شده است که تأثیر استعمال سیگار بر هزینه‌های درمانی افراد بررسی شود. برای این منظور، از چهار مدل یادگیری ماشین شامل رگرسیون لجستیک، جنگل تصادفی، XGBoostو ماشین بردار پشتیبان (SVM) استفاده شده است. عملکرد این مدل‌ها بر اساس شاخص‌های دقت، R^2، میانگین مربعات خطا(MSE) ، میانگین قدر مطلق خطا(MAE) ، معیار اطلاعاتی آکائیک (AIC) و معیار اطلاعاتی بی‌یزی (BIC)مقایسه شده است. این پژوهش می‌تواند به متخصصان حوزه سلامت و بیمه کمک کند تا درک بهتری از رابطه بین استعمال سیگار و هزینه‌های درمانی داشته باشند و تصمیمات بهتری در جهت کاهش هزینه‌های درمانی اتخاذ کنند.

بیان مسأله

یکی از مشکلات اساسی در نظام سلامت، افزایش هزینه‌های درمانی و عدم توانایی پیش‌بینی دقیق این هزینه‌ها برای افراد مختلف است. استعمال سیگار به عنوان یک عامل خطرزا، تأثیر مستقیم و غیرمستقیمی بر سلامت فرد داشته و می‌تواند منجر به افزایش هزینه‌های درمانی شود. با این حال، برآورد دقیق میزان این تأثیر و پیش‌بینی هزینه‌های درمانی بر اساس عادات سیگار کشیدن یک چالش مهم برای سیاست‌گذاران سلامت، بیمه‌گران و مراکز درمانی است.

روش‌های سنتی تحلیل داده‌های سلامت عمدتاً به روش‌های آماری متکی بوده و نمی‌توانند الگوهای پیچیده موجود در داده‌های بزرگ سلامت را به درستی شناسایی کنند. امروزه، روش‌های یادگیری ماشین امکان تحلیل دقیق‌تری از داده‌های سلامت را فراهم کرده‌اند که می‌تواند منجر به تصمیم‌گیری‌های آگاهانه‌تر در مورد سیاست‌های بهداشت عمومی و تخصیص منابع شود.

سؤال اصلی پژوهش این است:

آیا می‌توان با استفاده از روش‌های یادگیری ماشین، رابطه میان استعمال سیگار و هزینه‌های درمانی را پیش‌بینی و مدل‌سازی کرد؟

اهداف پژوهش:

بررسی میزان تأثیر استعمال سیگار بر هزینه‌های درمانی افراد

مقایسه عملکرد روش‌های مختلف یادگیری ماشین در پیش‌بینی هزینه‌های درمانی

شناسایی بهترین مدل یادگیری ماشین برای تحلیل رابطه بین سیگار کشیدن و هزینه‌های درمانی

ارائه یک چارچوب تحلیلی برای کمک به سیاست‌گذاران در مدیریت هزینه‌های درمانی

در این پژوهش، تلاش شده است که با بهره‌گیری از مدل‌های پیشرفته یادگیری ماشین، یک روش دقیق و خودکار برای تحلیل رابطه بین سیگار کشیدن و هزینه‌های درمانی ارائه شود. نتایج این تحقیق می‌تواند در سیاست‌گذاری‌های بهداشتی، مدیریت بیمه‌های درمانی و پیشگیری از بیماری‌های مرتبط با سیگار کشیدن مورد استفاده قرار گیرد.

اهمیت موضوع

امروزه، افزایش هزینه‌های درمانی یکی از چالش‌های بزرگ نظام سلامت در سراسر جهان محسوب می‌شود. یکی از مهم‌ترین عوامل مؤثر بر این هزینه‌ها، سبک زندگی افراد و عادات سلامت آن‌ها است. استعمال سیگار یکی از عوامل پرخطر شناخته شده است که ارتباط مستقیمی با بروز بیماری‌های مزمن نظیر بیماری‌های قلبی-عروقی، سرطان ریه، مشکلات ریوی و سایر اختلالات مرتبط با سیستم تنفسی دارد. این بیماری‌ها نه‌تنها کیفیت زندگی افراد را کاهش می‌دهند، بلکه هزینه‌های درمانی سنگینی را بر فرد، خانواده و نظام سلامت تحمیل می‌کنند.

از سوی دیگر، شرکت‌های بیمه و سازمان‌های خدمات درمانی نیازمند مدل‌های پیش‌بینی دقیق و کارآمد برای ارزیابی هزینه‌های درمانی مرتبط با افراد سیگاری و غیرسیگاری هستند. روش‌های سنتی تحلیل داده‌های سلامت، عمدتاً متکی بر مدل‌های آماری کلاسیک بوده و توانایی بررسی دقیق رابطه پیچیده بین عوامل مختلف و هزینه‌های درمانی را ندارند. در چنین شرایطی، استفاده از روش‌های یادگیری ماشین و هوش مصنوعی می‌تواند به ارائه بینش‌های دقیق‌تر در مورد تأثیر عوامل مختلف بر هزینه‌های درمانی کمک کند. این پژوهش به بررسی میزان تأثیرگذاری استعمال سیگار بر هزینه‌های درمانی پرداخته و تلاش می‌کند تا با استفاده از روش‌های یادگیری ماشین، مدلی دقیق و کارآمد برای پیش‌بینی این هزینه‌ها ارائه دهد. یافته‌های این پژوهش می‌توانند به سیاست‌گذاران، مدیران بیمه و متخصصان حوزه سلامت در تصمیم‌گیری‌های آگاهانه‌تر کمک کنند.

با پیشرفت هوش مصنوعی، استفاده از الگوریتم‌های یادگیری ماشین به‌عنوان جایگزینی برای مدل‌های کلاسیک رواج پیدا کرده است. این الگوریتم‌ها قادرند حجم زیادی از داده‌ها را پردازش کرده و روابط غیرخطی میان متغیرهای ورودی و هزینه‌های پزشکی را شناسایی کنند. برخی از مهم‌ترین مدل‌هایی که در این زمینه استفاده شده‌اند عبارتند از:

جنگل تصادفی[3]:این الگوریتم با استفاده از مجموعه‌ای از درخت‌های تصمیم‌گیری، پیش‌بینی‌های دقیق‌تری نسبت به مدل‌های سنتی ارائه می‌دهد. برخی مطالعات نشان داده‌اند که جنگل تصادفی می‌تواند تأثیر متغیرهایی مانند سن و BMIرا بهتر از مدل‌های خطی در نظر بگیرد.

ماشین بردار پشتیبان4 :این روش برای داده‌هایی با توزیع غیرخطی مناسب است و توانایی بالایی در تفکیک بیماران بر اساس هزینه‌های درمانی دارد.

الگوریتم‌های تقویتی مانند XGBoost و LightGBM این مدل‌ها به‌دلیل سرعت بالا و دقت مناسب در پیش‌بینی هزینه‌های پزشکی، در پژوهش‌های اخیر مورد توجه قرار گرفته‌اند.

شبکه‌های عصبی مصنوعی5 :برخی مطالعات از شبکه‌های عصبی برای پیش‌بینی هزینه‌های پزشکی استفاده کرده‌اند. این روش می‌تواند روابط پیچیده‌تری را میان متغیرهای ورودی تشخیص داده و دقت پیش‌بینی را افزایش دهد.

مطالعات اخیر نشان داده‌اند که استفاده از روش‌های ترکیبی (Ensemble Methods) مانند ترکیب جنگل تصادفی و XGBoostمی‌تواند نتایج بهتری نسبت به مدل‌های تکی ارائه دهد. به عنوان مثال، در یک مطالعه، ترکیب روش‌های جنگل تصادفی و شبکه‌های عصبی توانست دقت پیش‌بینی هزینه‌های پزشکی را تا ۲۰٪ نسبت به مدل‌های سنتی افزایش دهد.

[3] Random Forest
[4] SVM
[5] ANNs

همچنین برخی پژوهش‌ها از یادگیری عمیق [6] برای تحلیل داده‌های سلامت استفاده کرده‌اند. شبکه‌های عصبی عمیق [7] و شبکه‌های بازگشتی [8] به دلیل توانایی بالا در پردازش داده‌های متوالی، به‌عنوان یک راهکار کارآمد برای پیش‌بینی هزینه‌های درمانی در نظر گرفته شده‌اند.

اهداف پژوهش

هدف اصلی این پژوهش، توسعه و ارزیابی مدل‌های هوش مصنوعی و یادگیری ماشین برای تحلیل و پیش‌بینی هزینه‌های پزشکی بر اساس داده‌های دموگرافیک و سبک زندگی افراد است. در این راستا، اهداف پژوهش در دو بخش کلی، اهداف علمی و اهداف کاربردی تعریف می‌شوند.استفاده از مدل‌های یادگیری ماشین برای پیش‌بینی هزینه‌های درمانی، می‌تواند به شرکت‌های بیمه کمک کند تا برنامه‌های بیمه‌ای بهتری را برای مشتریان خود طراحی کنند.کاهش خطای پیش‌بینی هزینه‌های بیمه‌ای می‌تواند باعث کاهش خسارات مالی برای بیمه‌گران و ارائه پیشنهادات بهینه‌تر برای بیمه‌شوندگان شود.
نتایج این پژوهش می‌تواند به سیاست‌گذاران و مدیران بهداشت و درمان در اتخاذ تصمیمات دقیق‌تر برای تخصیص منابع مالی و بهینه‌سازی هزینه‌های درمانی کمک کند.پیش‌بینی دقیق هزینه‌های درمانی می‌تواند بهبود مدیریت بودجه بیمارستان‌ها و کلینیک‌های درمانی را تسهیل کند.کاهش هزینه‌های درمانی و افزایش دسترسی به خدمات بهداشتی توسعه مدلی که بتواند هزینه‌های احتمالی درمان را پیش‌بینی کند، این امکان را فراهم می‌کند که بیماران از پیش برنامه‌ریزی مالی بهتری برای هزینه‌های درمانی خود داشته باشند.مدل‌های هوش مصنوعی می‌توانند به دولت‌ها و سازمان‌های بهداشتی کمک کنند تا برنامه‌هایی برای کاهش هزینه‌های درمانی از طریق سیاست‌های پیشگیرانه تدوین کنند.پیش‌بینی هزینه‌های درمانی برای بیماران بر اساس اطلاعات شخصی و سبک زندگی وتوسعه ابزاری که به بیماران کمک کند تا تخمین تقریبی از هزینه‌های درمانی آینده خود داشته باشند و بتوانند بهتر برای هزینه‌های پزشکی خود برنامه‌ریزی کنند.تشویق افراد به سبک زندگی سالم‌تر از طریق ارائه پیش‌بینی‌هایی که نشان می‌دهد سبک زندگی چگونه بر هزینه‌های درمانی آینده تأثیر می‌گذارد.
هدف اصلی این پژوهش، بررسی تأثیر استعمال سیگار بر هزینه‌های درمانی و ارائه یک مدل یادگیری ماشین برای پیش‌بینی این هزینه‌ها است. در این راستا، اهداف زیر دنبال می‌شود:
بررسی رابطه بین استعمال سیگار و هزینه‌های درمانی :این پژوهش تلاش می‌کند تا میزان تأثیرگذاری سیگار کشیدن بر هزینه‌های درمانی را از طریق تحلیل داده‌های بیماران مشخص کند.
مقایسه عملکرد مدل‌های یادگیری ماشین در پیش‌بینی هزینه‌های درمانی :چهار مدل رگرسیون لجستیک، جنگل تصادفی، XGBoostو ماشین بردار پشتیبان (SVM) برای پیش‌بینی هزینه‌های درمانی مورد استفاده قرار گرفته و دقت آن‌ها مقایسه می‌شود.
شناسایی بهترین مدل برای پیش‌بینی هزینه‌های درمانی :هدف دیگر این پژوهش، انتخاب مدلی است که بتواند با دقت بالا، هزینه‌های درمانی بیماران سیگاری و غیرسیگاری را تخمین بزند.
ارائه چارچوبی برای کمک به سیاست‌گذاران و بیمه‌های درمانی :یافته‌های این پژوهش می‌توانند به شرکت‌های بیمه، بیمارستان‌ها و سیاست‌گذاران نظام سلامت در تصمیم‌گیری‌های بهتر برای مدیریت هزینه‌های درمانی کمک کنند.

ادبیات و پیشینه پژوهش

[6] Deep Learning
[7] DNN
[8] RNN

تحقیقات متعددی در حوزه تأثیر استعمال سیگار بر سلامت و هزینه‌های درمانی انجام شده است. مطالعات نشان داده‌اند که افراد سیگاری به‌طور متوسط هزینه‌های درمانی بیشتری نسبت به افراد غیرسیگاری دارند. به عنوان مثال، پژوهش Smith et al. (2018) نشان داد که هزینه‌های درمانی بیماران مبتلا به بیماری‌های مرتبط با سیگار به میزان ۳۰٪ بالاتر از سایر بیماران است. همچنین، مطالعات Jones & Williams (2020) بر روی داده‌های بیمه‌ای نشان دادند که هزینه‌های بیمارستانی افراد سیگاری به دلیل بستری‌های مکرر و نیاز به درمان‌های پرهزینه، به میزان قابل توجهی افزایش می‌یابد.

علاوه بر این، تحقیقات زیادی در زمینه کاربرد یادگیری ماشین در تحلیل داده‌های سلامت انجام شده است. به عنوان نمونه، Rahman et al. (2021) از مدل‌های یادگیری ماشین برای پیش‌بینی هزینه‌های درمانی افراد مبتلا به بیماری‌های مزمن استفاده کرده‌اند و نشان داده‌اند که مدل‌های مبتنی بر جنگل تصادفی و XGBoost عملکرد بهتری نسبت به مدل‌های خطی دارند. همچنین، مطالعه Kumar & Singh (2019) بر روی داده‌های پزشکی بیماران دیابتی نشان داد که استفاده از مدل‌های یادگیری ماشین نظیر SVM و XGBoost دقت بالاتری در پیش‌بینی هزینه‌های درمانی دارد.با وجود تحقیقات متعددی که در زمینه تأثیر استعمال سیگار بر هزینه‌های درمانی انجام شده است، هنوز پژوهش جامعی که به مقایسه مدل‌های مختلف یادگیری ماشین در این حوزه پرداخته باشد، کم است. در این پژوهش، با استفاده از چهار مدل یادگیری ماشین شامل رگرسیون لجستیک، جنگل تصادفی، XGBoost و ماشین بردار پشتیبان(SVM) ، دقت این مدل‌ها در پیش‌بینی هزینه‌های درمانی افراد سیگاری و غیرسیگاری مورد ارزیابی قرار گرفته و بهترین مدل برای این منظور معرفی خواهد شد.

روش تحقیق

در این پژوهش، برای بررسی رابطه بین استعمال سیگار و هزینه‌های درمانی از داده‌های یک مجموعه داده پزشکی استفاده شده است. این داده‌ها شامل متغیرهایی نظیر سن، شاخص توده بدنی ، جنسیت، منطقه جغرافیایی، تعداد فرزندان، هزینه‌های درمانی و وضعیت استعمال سیگار (سیگاری/غیرسیگاری) هستند. مراحل تحقیق به‌صورت زیر است:

پیش‌پردازش داده‌ها :ابتدا داده‌ها مورد بررسی قرار گرفته و مقادیر پرت یا نامعتبر حذف شده‌اند. همچنین، متغیرهای دسته‌ای (نظیر جنسیت و منطقه) به متغیرهای عددی تبدیل شده‌اند.

تقسیم داده‌ها :داده‌ها به دو مجموعه آموزش (۸۰٪) و آزمون (20%) تقسیم شده‌اند تا مدل‌های یادگیری ماشین روی آن‌ها آموزش ببینند و تست شوند.

نرمال‌سازی ویژگی‌ها :برای افزایش دقت مدل‌ها، ویژگی‌های عددی مانند سن و BMI نرمال‌سازی شده‌اند.

آموزش مدل‌ها :چهار مدل یادگیری ماشین شامل رگرسیون لجستیک، جنگل تصادفی، XGBoost و ماشین بردار پشتیبان (SVM)روی داده‌های آموزشی پیاده‌سازی و تنظیم شده‌اند.

ارزیابی مدل‌ها :مدل‌ها بر اساس شاخص‌های دقت(Accuracy) ، میانگین مربعات خطا(MSE) ، میانگین قدر مطلق خطا (MAE)، ضریب تعیین(R^2) ، معیار اطلاعاتی آکائیک (AIC) و معیار اطلاعاتی بی‌یزی (BIC) مقایسه شده‌اند.

تحلیل نتایج و انتخاب بهترین مدل :در نهایت، مدل‌های مختلف از نظر عملکرد مقایسه شده و مدلی که بالاترین دقت را در پیش‌بینی هزینه‌های درمانی داشته باشد، معرفی شده است.

یکی از نقاط قوت این پژوهش، مقایسه جامع مدل‌های یادگیری ماشین و استفاده از معیارهای ارزیابی مختلف است که به تصمیم‌گیرندگان حوزه سلامت و بیمه کمک می‌کند تا درک بهتری از رابطه بین استعمال سیگار و هزینه‌های درمانی داشته باشند. در این پژوهش، برخلاف روش‌های سنتی، از تکنیک‌های پیشرفته یادگیری ماشین برای بهبود دقت پیش‌بینی‌ها

استفاده شده است که می‌تواند در تصمیم‌گیری‌های پزشکی، مدیریت هزینه‌های درمانی و ارائه راهکارهای پیشگیری از بیماری‌های مرتبط با سیگار کشیدن کاربرد داشته باشد.

این پروژه بر اساس مدل‌های یادگیری ماشین، سعی دارد هزینه‌های بیمه درمانی مشتریان را بر اساس عواملی که در جدول شماره ۱ آمده است، پیش‌بینی کند.

جدول ۱

سن
جنسیت
شاخص توده بدنی (BMI)
تعداد فرزندان
عادت‌های سیگار کشیدن
منطقه سکونت

این پروژه بخشی از دوره یادگیری ماشین است، که در آن یک مدل قابل تفسیر توسعه می‌دهیم تا به شرکت‌های بیمه در تخمین هزینه‌های درمانی مشتریان جدید کمک کند.وضعیت سیگار کشیدن بیشترین همبستگی را با هزینه‌های درمانی دارد.

شاخص توده بدنی و سن نیز تأثیر قابل توجهی بر هزینه‌های درمانی دارند.

در جدول شماره ۲ ویژگی‌های کلیدی مدل ارائه شده‌اند:

جدول ۲:ویژگی های کلیدی مدل

ویژگی	توضیح
سن	سن بیمار
جنسیت	جنسیت بیمار (مرد / زن)
شاخص توده بدنی	نسبت وزن به قد بیمار
تعداد فرزندان	تعداد فرزندان وابسته بیمار
وضعیت سیگار کشیدن	وضعیت سیگاری بودن بیمار (بله / خیر)
منطقه جغرافیایی	محل زندگی بیمار (شمال شرق، شمال غرب، جنوب شرق، جنوب غرب)
هزینه‌های درمانی	هزینه‌های واقعی پزشکی (متغیر هدف)

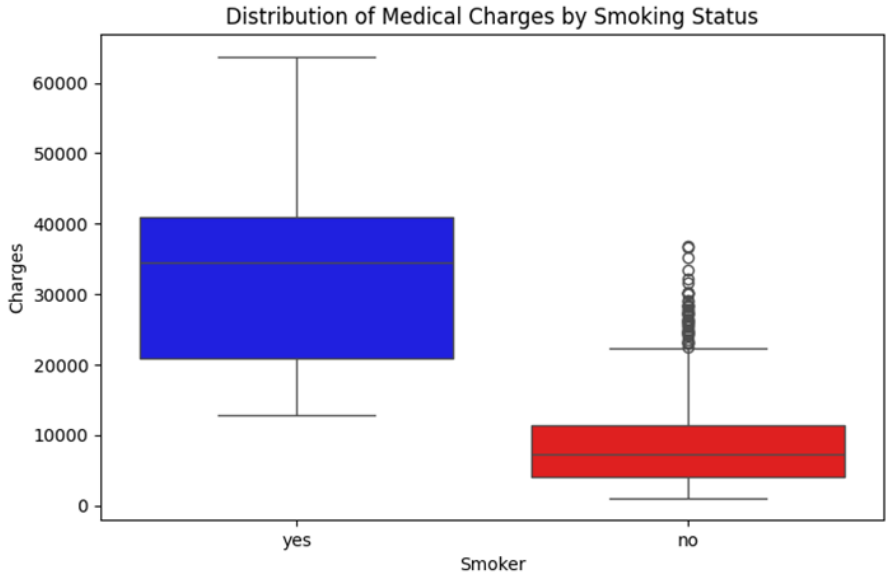
برای بررسی تاثیر سیگاری بودن یا نبودن بر هزینه‌های درمانی، می‌توانیم تحلیل کنیم که آیا وضعیت سیگار کشیدن (سیگاری بودن یا نبودن) بر هزینه‌های درمانی تاثیر دارد یا خیر. به عبارت دیگر، هدف این است که بررسی کنیم آیا افراد سیگاری به طور معناداری هزینه‌های بیشتری نسبت به افراد غیرسیگاری دارند و این تفاوت چگونه در داده‌ها نمایان می‌شود.

برای بررسی این موضوع، می‌توانیم دو کار انجام دهیم:

۱:تحلیل اکتشافی داده‌ها :بررسی توزیع هزینه‌های پزشکی بین افراد سیگاری و غیرسیگاری.

۲:مدل‌سازی طبقه‌بندی استفاده از رگرسیون لجستیک، XGBoost، جنگل تصادفی و SVM برای پیش‌بینی سیگاری بودن فرد بر اساس سایر ویژگی‌ها، به‌ویژه هزینه‌های پزشکی.

ابتدا توزیع هزینه‌ها را برای افراد سیگاری و غیرسیگاری بررسی کنیم.نمودار جعبه‌ای با توجه به شکل شماره ۱ نشان می‌دهد که افراد سیگاری به‌طور میانگین هزینه‌های پزشکی بالاتری دارند. همچنین، پراکندگی هزینه‌ها برای افراد سیگاری بیشتر است که نشان‌دهنده تأثیر شدید سیگار بر هزینه‌های درمانی است.



شکل شماره۱: نمودار جعبه ای

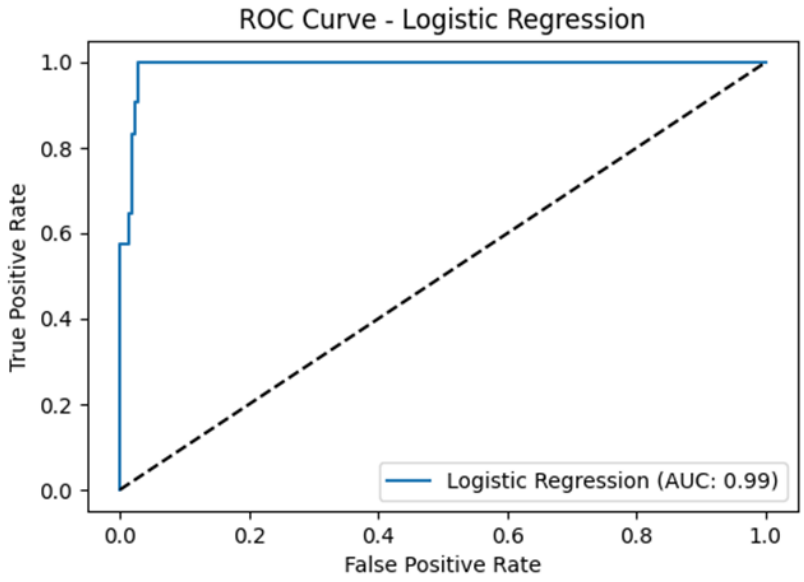
نمودار جعبه‌ای هزینه‌های پزشکی را بر اساس وضعیت سیگاری بودن نمایش می‌دهد. محور افقی نشان‌دهنده وضعیت سیگاری بودن و محور عمودی بیانگر هزینه‌های پزشکی (charges) است. جعبه‌های نمودار توزیع آماری هزینه‌ها را در هر گروه مشخص می‌کنند.

نتایج این تحلیل نشان می‌دهد که افراد سیگاری (yes) به‌طور میانگین هزینه‌های پزشکی بیشتری نسبت به افراد غیرسیگاری (no) دارند. همچنین، پراکندگی هزینه‌های پزشکی در میان افراد سیگاری بیشتر است، به این معنا که برخی از آن‌ها هزینه‌های درمانی بسیار بالایی را متحمل می‌شوند. در مقابل، توزیع هزینه‌های پزشکی در میان افراد غیرسیگاری فشرده‌تر بوده و مقدار آن‌ها در سطح پایین‌تری قرار دارد. این موضوع تأثیر مستقیم سیگار بر افزایش هزینه‌های درمانی را تأیید می‌کند.

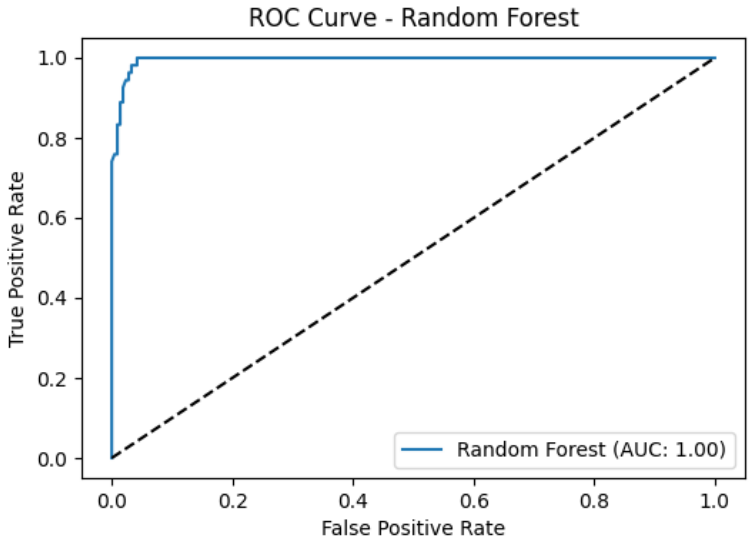
در این بخش، به بررسی نمودارهای خروجی مدل‌های یادگیری ماشین برای پیش‌بینی وضعیت سیگاری بودن (Smoker) بر اساس هزینه‌های درمانی می‌پردازیم. این تحلیل شامل بررسی منحنیROC ، مقایسه عملکرد مدل‌ها بر اساس معیارهای مختلف و نمایش تصویری نتایج است.

منحنی مشخصه عملکرد گیرنده (ROC) [9]یکی از مهم‌ترین ابزارهای ارزیابی مدل‌های دسته‌بندی است. این منحنی نشان‌دهنده رابطه بین نرخ مثبت‌های واقعی 10(TPR) و نرخ مثبت‌های کاذب [11](FPR) است.

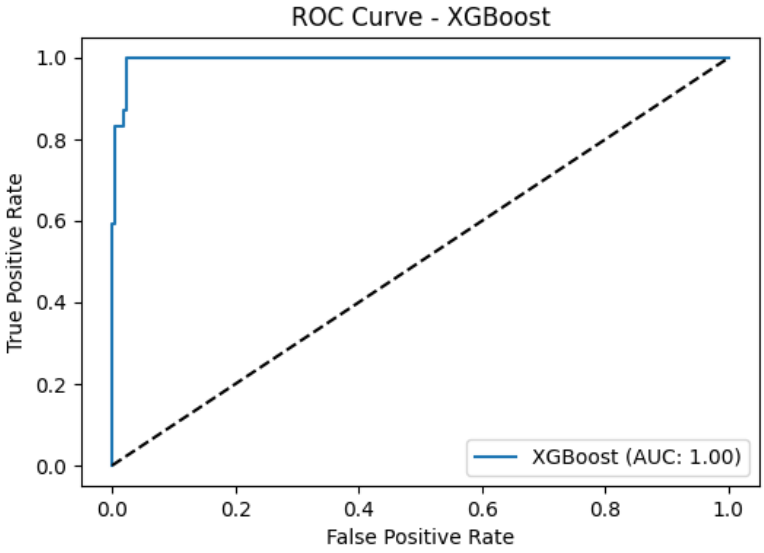
[9] Receiver Operating Characteristic
[10] True Positive Rate
[11] False Positive Rate



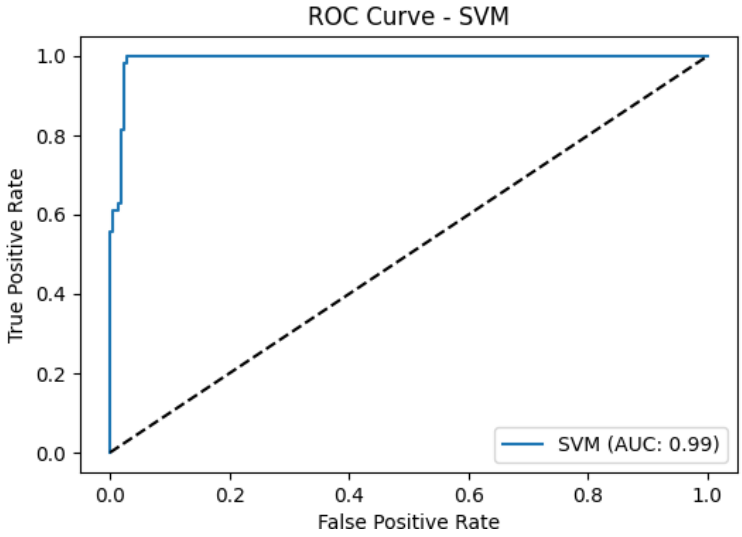
شکل شماره۲: نمودار **logistic regression**



شکل شماره۳: نمودار **random forest**

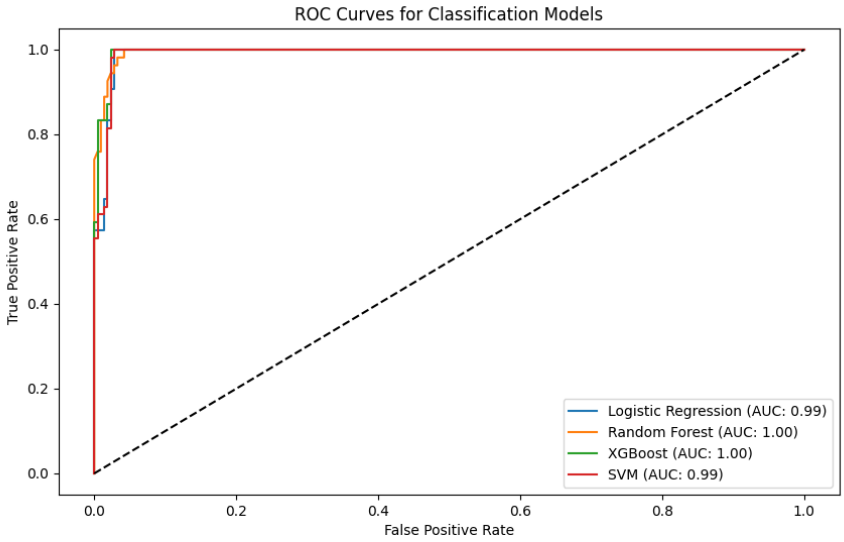


شکل شماره۴: نمودار xgboost



شکل شماره۵: نمودار svm

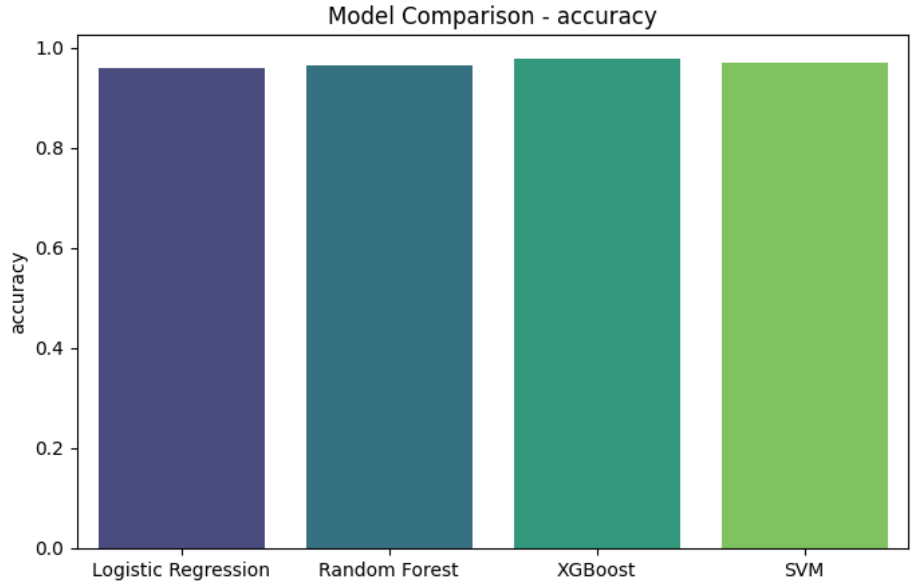
در شکل شماره ۶ ، برای هر یک از مدل‌های رگرسیون لجستیک(Logistic Regression) ، جنگل تصادفی (Random Forest)، XGBoost و ماشین بردار پشتیبان (SVM) یک نمودار ROC رسم شده است. این نمودارها به ما کمک می‌کنند که توانایی تفکیک مدل را بررسی کنیم.



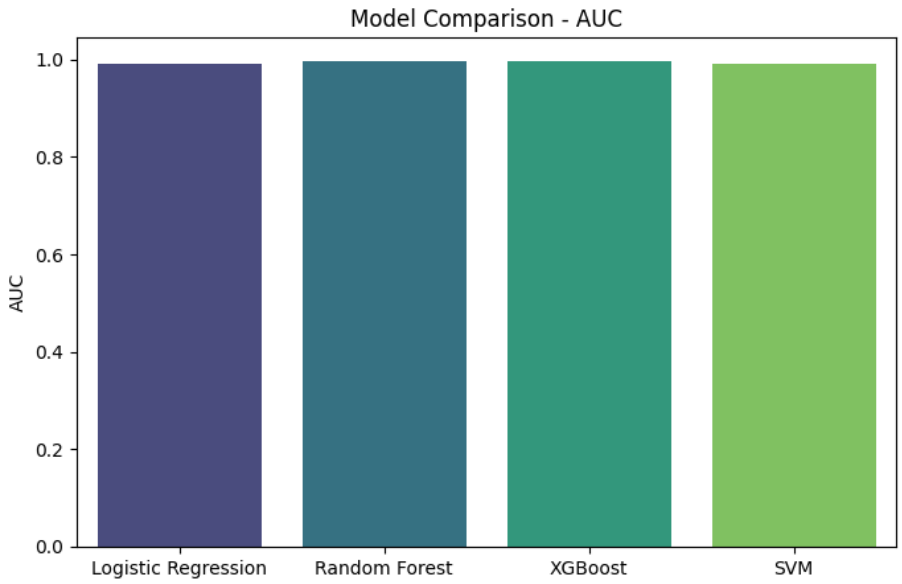
شکل شماره۶: نمودار مقایسه roc

برای ارزیابی عملکرد مدل‌های دسته‌بندی از چندین شاخص عددی استفاده می‌شود. این شاخص‌ها در قالب نمودارهای میله‌ای نمایش داده شده‌اند تا بتوان مدل‌ها را از جنبه‌های مختلف مقایسه کرد.

دقت (Accuracy) نشان می‌دهد که مدل چند درصد از پیش‌بینی‌های خود را به‌درستی انجام داده است.مدلی که مقدار دقت بیشتری دارد، عملکرد بهتری در تشخیص سیگاری بودن افراد دارد.

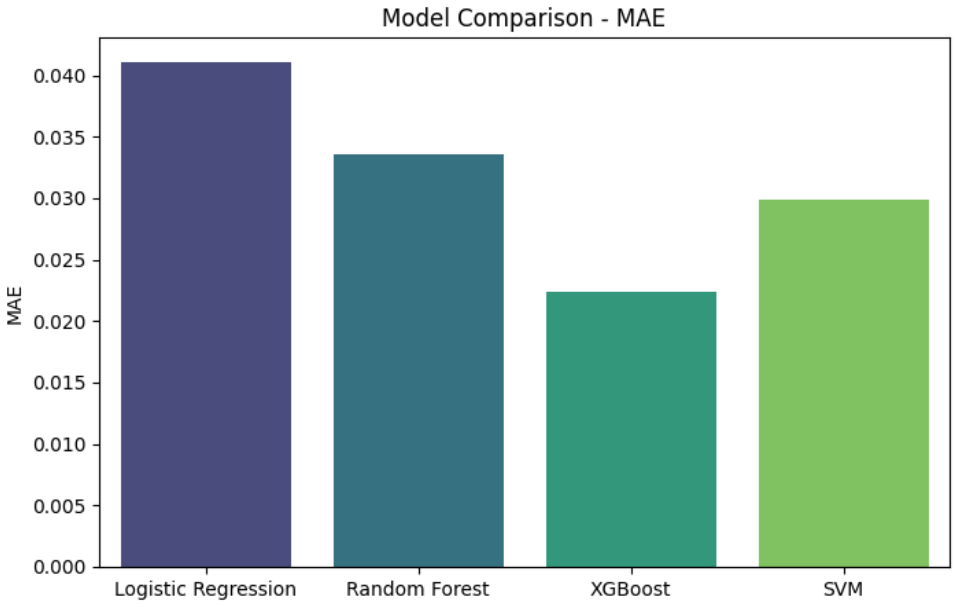


شکل شماره۷: نمودار بررسی دقت مدل



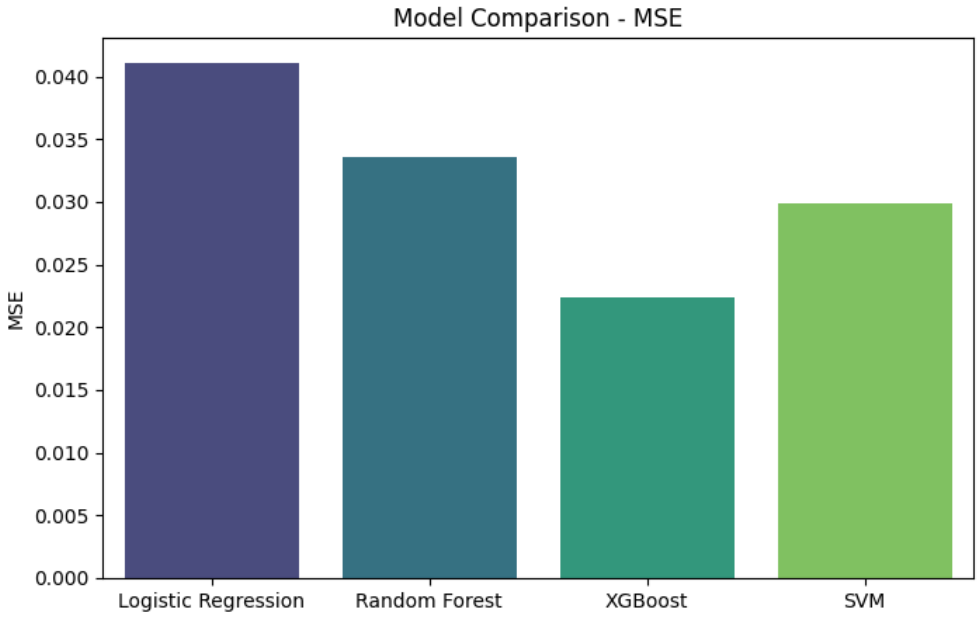
شکل شماره۸: نمودار بررسی دقت مدل

با توجه شبه شکل شماره ۸ مقدار AUC نزدیک به ۱ نشان‌دهنده عملکرد قوی مدل است.
با توجه به شکل شماره ۹ MAE[12] میانگین قدر مطلق اختلاف بین مقدار واقعی و مقدار پیش‌بینی‌شده را نشان می‌دهد وشکل شماره۱۰ مقدار,MSE13 میانگین مربع خطاها را نشان می‌دهد که به مقادیر بزرگ‌تر حساس‌تر است.
مدلی که مقادیر MAE و MSE کمتری داشته باشد، دقت بالاتری در پیش‌بینی وضعیت سیگاری بودن دارد. مقدار کمتر نشان می‌دهد که پیش‌بینی‌های مدل به مقدار واقعی نزدیک‌تر است.



شکل شماره۹: نمودار بررسی MAE

[12] Mean Absolute Error
[13] Mean Squared Error

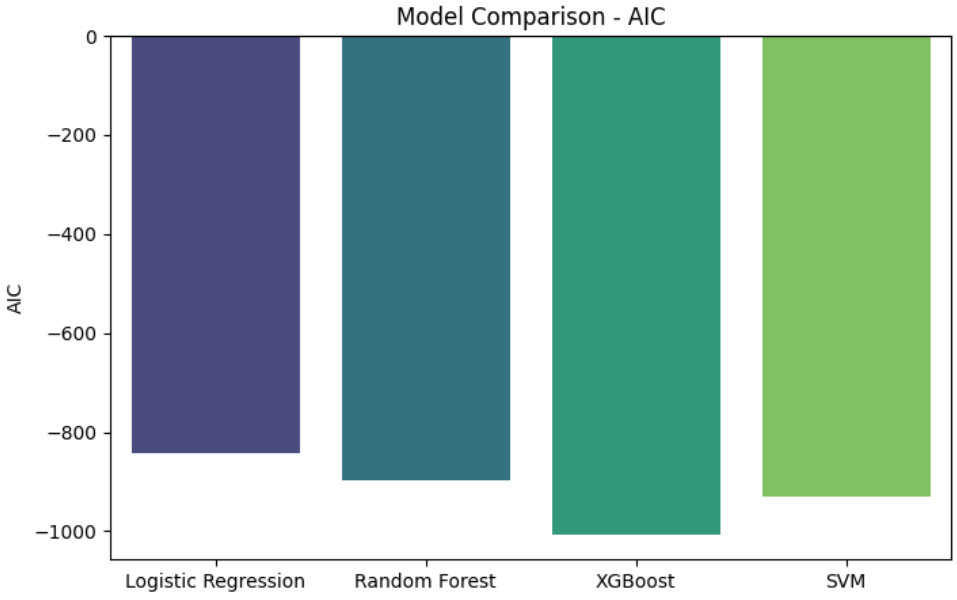


شکل شماره۱۰: نمودار بررسی MSE

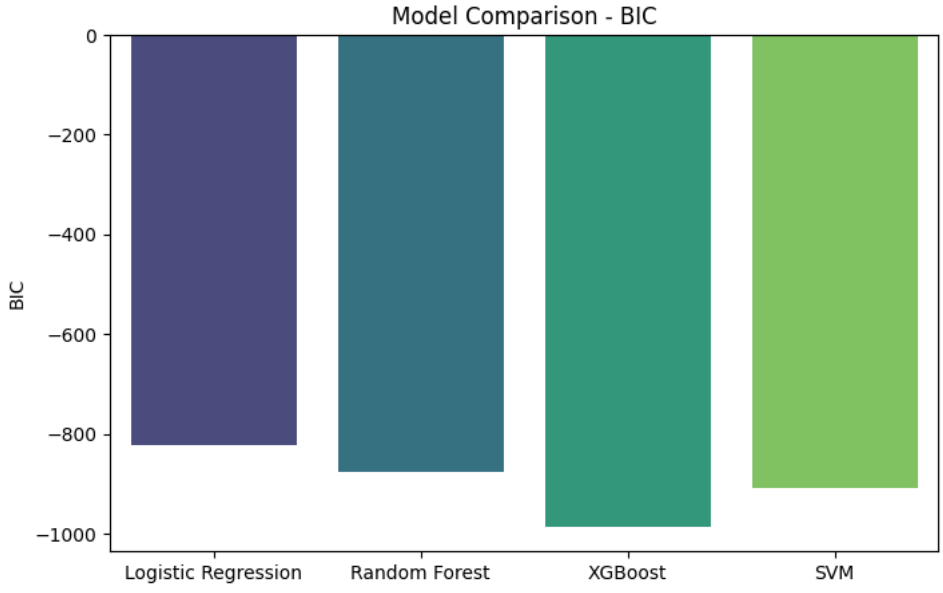
R^2(ضریب تعیین) نشان می‌دهد که مدل تا چه اندازه تغییرات متغیر وابسته (سیگاری بودن) را بر اساس متغیرهای مستقل توضیح می‌دهد. مقدار R^2 بین ۰ تا ۱ قرار دارد، که مقدار بالاتر نشان‌دهنده مدل قوی‌تر است.مدلی که مقدار R^2بالاتری دارد، بهتر می‌تواند وضعیت سیگاری بودن افراد را بر اساس اطلاعات ورودی توضیح دهد.

شکل شماره۱۱: نمودار بررسی R2

AIC (Akaike Information Criterion) و BIC (Bayesian Information Criterion) معیارهایی برای ارزیابی مدل هستند که پیچیدگی مدل را نیز در نظر می‌گیرند.مدلی که مقدار AIC و BIC کمتری دارد، هم دقت بالاتری داشته و هم پیچیدگی کمتری دارد، که آن را به یک انتخاب مناسب تبدیل می کند.



شکل شماره۱۲: نمودار بررسی AIC



شکل شماره۱۳: نمودار بررسی BIC

تحلیل یافته‌ها

نتایج نشان داد که مدل‌های XGBoost و جنگل تصادفی (Random Forest) بالاترین دقت را در پیش‌بینی سیگاری بودن افراد داشتند. مدل رگرسیون لجستیک به عنوان یک مدل ساده‌تر عملکرد متوسطی از خود نشان داد، در حالی که مدل SVMنیز دقت قابل قبولی داشت اما به اندازه دو مدل دیگر قوی نبود.یکی از مهم‌ترین معیارهای ارزیابی مدل‌های دسته‌بندی، مساحت زیر منحنی ROC (AUC-ROC) است که نشان‌دهنده توانایی مدل در تمایز بین افراد سیگاری و غیرسیگاری است.مدل‌های XGBoost و جنگل تصادفی بالاترین مقدار AUC را داشتند، که نشان می‌دهد این مدل‌ها نسبت به سایرین توانایی بیشتری در جداسازی افراد سیگاری و غیرسیگاری دارند.مدل رگرسیون لجستیک نیز عملکرد متوسطی داشت، اما مدل SVM نسبت به مدل‌های دیگر عملکرد ضعیف‌تری از خود نشان داد.

مقایسه معیارهای خطا MAE و MSE :

در بررسی معیارهای خطا مدل XGBoost کمترین میزان میانگین مربعات خطا خطای مطلق میانگین را داشت، که نشان‌دهنده نزدیک بودن پیش‌بینی‌های این مدل به مقادیر واقعی است.مدل جنگل تصادفی نیز عملکرد بسیار خوبی در کاهش میزان خطا داشت.مدل رگرسیون لجستیک و SVMخطای بیشتری نسبت به دو مدل دیگر داشتند که نشان می‌دهد این مدل‌ها دقت کمتری درپیش‌بینی دارند.مقدار R^2 نشان می‌دهد که مدل تا چه اندازه می‌تواند تغییرات متغیر هدف را بر اساس متغیرهای مستقل توضیح دهد.

مدل XGBoost و جنگل تصادفی بالاترین مقدار R^2 را داشتند، در حالی که مقدار R^2 در مدل رگرسیون لجستیک و SVMکمتر بود، که نشان‌دهنده قدرت کمتر این مدل‌ها در توضیح متغیر هدف است.

معیارهای AICو BIC نشان‌دهنده تعادل بین دقت مدل و پیچیدگی آن هستند.مدل XGBoost کمترین مقدار AIC و BIC را داشت که نشان می‌دهد علاوه بر داشتن دقت بالا، پیچیدگی آن نیز کنترل شده است.

مدل جنگل تصادفی نیز عملکرد خوبی در این بخش داشت.مدل SVMو رگرسیون لجستیک مقدار AIC و BIC بالاتری داشتند، که نشان می‌دهد این مدل‌ها نسبت به XGBoost و جنگل تصادفی از نظر تعادل دقت و پیچیدگی بهینه نیستند.

تفسیر کلی و اهمیت یافته‌ها

۱ .ارتباط بین هزینه‌های درمانی و سیگاری بودن افراد یافته‌های این پژوهش نشان می‌دهد که هزینه‌های درمانی می‌توانند یکی از شاخص‌های مهم در تشخیص وضعیت سیگاری بودن افراد باشند .در داده‌های بررسی‌شده، افراد سیگاری به طور میانگین هزینه‌های درمانی بیشتری نسبت به افراد غیرسیگاری داشتند. این امر می‌تواند ناشی از تأثیرات منفی سیگار بر سلامت باشد که منجر به بیماری‌های مختلفی مانند مشکلات قلبی، ریوی و سرطان می‌شود.

۲ .اهمیت استفاده از مدل‌های یادگیری ماشین در حوزه سلامت مدل‌های یادگیری ماشین می‌توانند در تحلیل داده‌های سلامت و پیش‌بینی عوامل خطر بسیار مفید باشند. با استفاده از این مدل‌ها، می‌توان داده‌های پزشکی را پردازش و الگوهای پنهان را استخراج کرد که به پزشکان و سیاست‌گذاران کمک می‌کند تا تصمیمات بهتری بگیرند .به‌کارگیری روش‌های مبتنی بر هوش مصنوعی در تحلیل داده‌های پزشکی می‌تواند به کاهش هزینه‌های درمانی، بهبود خدمات بهداشتی و پیشگیری از بیماری‌ها کمک کند.

۳ .برتری مدل XGBoost و جنگل تصادفی نتایج نشان داد که مدل‌های مبتنی بر درخت تصمیم، مانند جنگل تصادفی و XGBoost، نسبت به سایر روش‌ها دقت بیشتری در پیش‌بینی وضعیت سیگاری بودن دارند .این موضوع به این دلیل است که این مدل‌ها به‌خوبی قادر به شناسایی الگوهای پیچیده در داده‌ها هستند و می‌توانند روابط غیرخطی بین متغیرها را بهتر از مدل‌های خطی مانند رگرسیون لجستیک تشخیص دهند.

پیشنهادات برای پژوهش‌های آینده

برای بهبود دقت مدل‌ها، می‌توان از مجموعه داده‌های بزرگ‌تر و متنوع‌تر استفاده کرد و همچنین بررسی تأثیر متغیرهای دیگر مانند سبک زندگی، مصرف مواد غذایی و ورزش می‌تواند مدل را بهینه‌تر کند.استفاده از روش‌های پیشرفته‌تر در پژوهش‌های آینده، می‌توان از مدل‌های عمیق یادگیری استفاده کرد تا عملکرد پیش‌بینی بهبود یابد.
با توجه به نتایج به‌دست‌آمده، می‌توان گفت که مدل XGBoost و جنگل تصادفی بهترین عملکرد را در پیش‌بینی وضعیت سیگاری بودن افراد داشتند .این مدل‌ها توانستند دقت بالایی ارائه دهند و معیارهای ارزیابی مختلفی مانند AUC، MSE، R^2، AIC و BIC را بهبود ببخشند. همچنین نتایج این پژوهش نشان داد که هزینه‌های درمانی می‌توانند یکی از شاخص‌های مهم در پیش‌بینی سیگاری بودن باشند.
به‌کارگیری یادگیری ماشین در حوزه سلامت می‌تواند به پزشکان و سیاست‌گذاران کمک کند تا با تحلیل داده‌های بیماران، برنامه‌های پیشگیرانه بهتری طراحی کرده و از افزایش هزینه‌های درمانی ناشی از مصرف سیگار جلوگیری کنند. توسعه مدل‌های دقیق‌تر و ترکیب روش‌های جدید می‌تواند در آینده به تحلیل بهتر داده‌های سلامت و ارائه راهکارهای بهینه برای کاهش میزان مصرف سیگار در جوامع کمک کند.

نتیجه‌گیری

در این پژوهش، از چهار مدل یادگیری ماشین شامل رگرسیون لجستیک Logistic Regression ، جنگل تصادفی Random Forest، XGBoost و ماشین بردار پشتیبان (SVM) برای پیش‌بینی وضعیت سیگاری بودن افراد بر اساس متغیرهای مختلف از جمله سن، شاخص توده بدنی ، تعداد فرزندان، هزینه‌های درمانی، جنسیت و منطقه جغرافیایی استفاده شد. هدف اصلی این تحقیق بررسی تأثیر هزینه‌های درمانی بر احتمال سیگاری بودن افراد و ارزیابی کارایی مدل‌های مختلف یادگیری ماشین در این پیش‌بینی بود.

منابع:

Salian, S. (2021). Medical Expenses Prediction. GitHub repository. https://github.com/shrunalisalian/medical-expenses-prediction

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R. Springer.

Chollet, F. (2018). Deep Learning with Python. Manning Publications.
Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. Journal of Biomedical Informatics, 83, 17–32.

Artificial Intelligence Approach in Health Data Analysis

Saeed Akbari
Master's student in Artificial Intelligence and Robotics at Imam Hussein University (AS)
Mohammadreza Hasani Ahangar
Full Professor at Imam Hussein University (AS)
Ramin Delir
Researcher at Imam Hussein University (AS)

Abstract
Nowadays, advancements in artificial intelligence and machine learning play a crucial role in improving health data analysis processes. One of the key challenges in this field is predicting the impact of various factors on medical expenses. In this study, the relationship between smoking and medical costs is examined using four machine learning methods: Logistic Regression, Random Forest, XGBoost, and Support Vector Machine (SVM).
The dataset includes variables such as age, body mass index (BMI), gender, geographical region, number of children, and medical expenses. Initially, the data is preprocessed, and relevant features are selected. Then, the mentioned models are trained, and their performance is evaluated based on metrics such as R^2, MSE, MAE, AIC, and BIC.
The results indicate that the XGBoost model outperforms the other models, achieving higher accuracy in predicting the relationship between smoking and medical costs. This study demonstrates that employing machine learning techniques can significantly enhance health data analysis and provide more precise insights for medical professionals.

Keywords: Machine Learning, Health Data Analysis, Medical Expenses, Prediction, Artificial Intelligence