Applied Research in Sustainable Industry Management
Original Research Article/ Vol.۰۱, No.۰۱, Spring ۲۰۲٤, P:۸۳-۱۰۱

Predicting the annual cost of medical insurance using machine learning

Ali Zhaleh Karimi*, Ramin Dalir**

Received: ۲۰۲٤/۰۳/۱۰ Accepted: ۲۰۲٤/۰٦/۱۹

Abstract:
Health insurance is one of the ways to reduce the costs imposed on society. Studying and researching in the field of damages and diseases helps the stakeholders to easily make policies in this regard. The insurance rate is affected by some medical issues. Accurate estimation of individual health care and treatment costs is important for a range of stakeholders and health agencies. Therefore, by predicting medical expenses, both the insured and the insurer can predict the future to some extent and have better options for making decisions. One of the goals of this article is to predict the low, medium or high spending of people for the treatment of the disease and to identify the effective factors in health insurance costs. In this article, the data of the US Census Bureau including ۱۳۳۸ samples with the features of age, gender, body mass index (BMI), smoking, number of dependents, region and annual cost are used. In the proposed method, the data set is first analyzed and reviewed in order to get a general view of it and to identify the influencing factors in the treatment cost. Then, by pre-processing and categorizing costs into low, medium and high, the data is converted into a form suitable for classification. In the next step, classification algorithms are used to learn the category of each of the samples, and by evaluating them, the best algorithm is selected. In the end, with the method of parameter improvement and algorithm parameters adjustment, the performance of the algorithm is improved and the annual cost prediction model is created. Examining the dataset showed that being a smoking, increasing age and being overweight have an effect on treatment costs. The classification results also show that the random forest algorithm has the ability to predict low, medium, and high costs for disease treatment with ۹۱٪ accuracy.

Keywords:
Medical insurance, medical cost, classification, machine learning

* Master's student of artificial intelligence and robotics, Imam Hossein (AS) University, Tehran, Iran (Crossponding Authur), email: azhkarimi@ihu.ac.ir
** PhD Student of artificial intelligence, University of Zanjan, Zanjan, Iran, email: rdalir@ihu.ac.ir

نشریه علمی پژوهش‌های کاربردی در مدیریت صنعت پایدار

پیش‌بینی میزان هزینهٔ سالانهٔ بیمهٔ درمانی با استفاده از یادگیری ماشین

علی ژاله کریمی*، رامین دلیر**

تاریخ دریافت: ۱۴۰۲/۱۲/۲۰ تاریخ پذیرش: ۱۴۰۳/۰۳/۳۰

چکیده

بیمهٔ درمانی، یکی از راهکارهای کاهش هزینه‌های تحمیلی بر افراد جامعه است. مطالعه و بررسی در حوزهٔ خسارات و بیماری‌ها، کمک می‌کند تا ذی‌نفعان به‌راحتی بتوانند دراین‌خصوص سیاست‌گزاری کنند. نرخ بیمه، تحت‌تأثیر برخی مسائل پزشکی است. برآورد دقیق هزینه‌های مراقبت‌های بهداشتی فردی و درمانی، برای طیفی از ذی‌نفعان و آژانس‌های بهداشتی مهم است. ازاین‌رو با پیش‌بینی هزینه‌های درمانی، هم بیمه‌شونده و هم بیمه‌گذار، می‌توانند تا حدودی آینده را پیش‌بینی کنند و گزینه‌های بهتری برای تصمیم‌گیری داشته باشند. پیش‌بینی هزینه‌کرد کم، متوسط یا زیاد افراد برای درمان بیماری و شناسایی عوامل مؤثر در هزینه‌های بیمهٔ درمانی، از اهداف این مقاله است. در این مقاله از داده‌های ادارهٔ سرشماری جمعیت امریکا مشتمل بر ۱۳۳۸ نمونه با ویژگی‌های سن، جنسیت، شاخص تودهٔ بدنی (BMI)، سیگاری‌بودن، تعداد افراد تحت تکفل، منطقه و هزینهٔ سالانه، استفاده شده است. در روش پیشنهادی ابتدا به تحلیل و بررسی مجموعه‌داده پرداخته می‌شود تا یک دید کلی از آن به دست آید و عوامل تأثیرگذار در هزینهٔ درمانی شناسایی شوند. سپس با پیش‌پردازش و دسته‌بندی هزینه‌ها به کم، متوسط و زیاد، داده‌ها به شکل مناسب برای طبقه‌بندی تبدیل می‌شوند. در مرحلهٔ بعد، از الگوریتم‌های طبقه‌بندی برای یادگیری دستهٔ هر کدام از نمونه‌ها استفاده می‌شود و با ارزیابی آن‌ها، بهترین الگوریتم انتخاب می‌شود. در انتها با روش بهبود پارامتر و تنظیم پارامترهای الگوریتم، عملکرد الگوریتم بهبود می‌یابد و مدل پیش‌بینی میزان هزینهٔ سالانه ایجاد می‌شود. بررسی مجموعه‌داده نشان داد که سیگاری‌بودن، افزایش سن و اضافه‌وزن بر روی هزینه‌های درمانی تأثیر گذارند. نتایج طبقه‌بندی نیز بیانگر این است که الگوریتم جنگل تصادفی با دقت ۹۱٪ توانایی پیش‌بینی میزان هزینه‌کرد کم، متوسط و زیاد برای درمان بیماری را دارد.

واژگان کلیدی: بیمهٔ درمانی، هزینهٔ درمانی، طبقه‌بندی، یادگیری ماشین

* دانشجوی کارشناسی ارشد هوش مصنوعی، دانشگاه امام حسین (ع)، تهران ، ایران(نویسنده مسئول)، پست الکترونیکی: azhkarimi@ihu.ac.ir

** دانشجوی دکتری هوش مصنوعی، دانشگاه زنجان، زنجان، ایران ، پست الکترونیکی: rdalir@ihu.ac.ir

۱. مقدمه

بیماری، یکی از رخدادهایی است که انسان در طور زندگی با آن دست و پنجه نرم می‌کند. بیماری، می‌تواند یک سرماخوردگی ساده یا سرطان باشد. هر نوع بیماری به نسبت شدتی که دارد، هزینه‌هایی بر افراد جامعه تحمیل می‌کند که در برخی موارد می‌تولند حجم بسیاری از سرمایهٔ مالی فرد را مصرف کند. برآورد هزینه‌های پزشکی، دشوار است؛ زیرا پرهزینه‌ترین شرایط، نادر و به‌ظاهر تصادفی هستند. با این حال، برخی شرایط برای بخش‌های خاصی از جمعیت، شایع‌تر است. به‌عنوان مثال، احتمال ابتلا به سرطان ریه در میان افراد سیگاری، بیشتر از افراد غیر سیگاری است و احتمال ابتلا به بیماری قلبی در میان افراد چاق، بیشتر است(Lantz, ۲۰۱۹). بیمهٔ درمانی، یکی از راهکارهای کاهش هزینه‌های تحمیلی بر افراد جامعه است. مفهوم محدود هزینه‌های بیمهٔ درمانی را می‌توان به‌عنوان بیمه در برابر هزینه‌های پزشکی تحمیل‌شده در تشخیص و درمان یک بیماری توضیح داد؛ یعنی هزینه‌های بیمهٔ پرداخت‌شده توسط بیمه‌شده برای به‌دست‌آوردن غرامت اقتصادی برای هزینه‌های پزشکی تحمیل‌شده به دلیل بیماری(Dong & Fei, ۲۰۲۱).

بیمه، یکی از ارکان توسعهٔ نظام مراقبت‌های بهداشتی در جهان است. مطالعه و بررسی در حوزهٔ خسارات و بیماری‌ها، کمک می‌کند تا ذی‌نفعان به‌راحتی بتوانند دراین‌خصوص سیاست‌گزاری کنند. ازاین‌رو با پیش‌بینی هزینه‌های درمانی، هم بیمه‌شونده و هم بیمه‌گذار، می‌توانند تا حدودی آینده را پیش‌بینی کنند و گزینه‌های بهتری برای تصمیم‌گیری داشته باشند. نرخ بیمه، تحت‌تأثیر برخی مسائل پزشکی است. برآورد دقیق هزینه‌های مراقبت‌های بهداشتی فردی و درمانی برای طیفی از ذی‌نفعان و آژانس‌های بهداشتی مهم است(Tajaddodi Nodehi et al., ۲۰۲۳).

مؤسسهٔ بین‌المللی تأمین اجتماعی (ISSA) در "گزارش هزینه‌های بیمهٔ پزشکی و بیمهٔ بیماری"، دلایل مشترک زیر را برای افزایش هزینه‌های بیمهٔ پزشکی فهرست کرده است: (۱) رشد جمعیت و پیری؛ (۲) تغییرها در ساختار بیماری؛ (۳) بهبود آموزش فرهنگی؛ (٤) تغییرهای زندگی در محیط و محیط کار؛ (٥) بهبود استانداردهای زندگی؛ (٦) توسعهٔ پزشکی(Dong & Fei, ۲۰۲۱).

در ایران نیز باتوجه‌به افزایش سن جمعیت، تقاضا برای منابع پزشکی در حال افزایش است. تحقیقات اندکی در مورد عوامل افزایش بیمهٔ درمانی و پیش‌بینی هزینه‌های درمانی برای سازمان‌ها و افراد صورت‌گرفته است. پایین‌بودن سهم اعتبارات تخصیصی از محل بودجهٔ

عمومی دولت و منابع پیش‌پرداخت در بیمه‌های اجتماعی درمانی و بالابودن سهم هزینه‌های خانوارها از مجموع هزینه‌های مصرفی بهداشت و درمان کشور، از عوامل محدودکننده در تأمین مالی هزینه‌های مصرفی در بخش سلامت، محسوب می‌شود(Arab et al., ۲۰۲۲).

یکی از راهکارها برای برخورد با کسری منابع مالی، مدیریت و کنترل هزینه‌ها بر اساس پیش‌بینی هزینه‌ها است. در تمامی سازمان‌ها، حجم انبوهی از داده وجود دارد که استفاده و تجزیه‌وتحلیل آن‌ها، می‌تواند کمک شایان توجهی به تصمیم‌گیری مدیران داشته باشد، اما حجم بالای این داده‌ها و تنوع موجود در آنها و ارتباطات زیاد و ناشناخته بین آنها، باعث شده است که ابزارهای دستی و سیستمی معمولی، قادر به استفاده درست از آنها نبوده و بررسی این داده‌ها با روش‌های قدیمی، خارج از توان افراد و ناکارآمد است. پیشرفت فنّاوری و ایجاد فناوری‌های جدید، دغدغه‌های موجود در این زمینه را کاهش داده و با پیشنهاد به‌کارگیری ابزارها و تکنیک‌های جدید، امکان بررسی داده‌های انبوه و کشف دانش از دل این داده‌ها را امکان‌پذیر ساخته است(Arab et al., ۲۰۲۲). یکی از امروزی‌ترین و جدیدترین این فناوری‌ها، هوش مصنوعی و استفاده از الگوریتم‌های یادگیری ماشین است.

در این مقاله از داده‌های آمار جمعیتی ادارهٔ سرشماری ایالات متحده با ویژگی‌های سن، جنسیت، شاخص تودهٔ بدنی (BMI)، سیگاری‌بودن، تعداد افراد تحت تکفل، منطقه و هزینهٔ سالانه، استفاده شده است. این مجموعه‌داده، نمونهٔ خوبی برای ارزیابی توانایی الگوریتم‌های یادگیری ماشین در پیش‌بینی عوامل مؤثر در بیماری و هزینه‌های درمانی است. برای پیش‌بینی هزینهٔ سالانه، در این مقاله از روش طبقه‌بندی استفاده شده است که یکی از تکنیک‌های یادگیری با نظارت است و برای پیش‌بینی کلاس نمونه‌ها، از مجموعه‌داده استفاده می‌شود. به همین منظور ابتدا مجموعه‌داده بررسی و تحلیل شده و به کمک روش‌های پیش‌پردازش و انتخاب ویژگی، مجموعه‌دادهٔ مناسب برای طبقه‌بندی داده‌ها به‌دست‌آمده است. در این روش، سه دسته کم، متوسط و زیاد بر اساس هزینه‌کرد بیمه‌شدگان تعریف شده است و از الگوریتم‌های طبقه‌بندی درخت تصمیم، جنگل تصادفی، رگرسیون لجستیک، ریج، KNN و... برای پیش‌بینی میزان هزینه‌کرد هر یک از بیمه‌شدگان استفاده شده است. همچنین برای افزایش دقت طبقه‌بندی، با استفاده از تکنیک بهبود پارامتر، الگوریتم Grid search به‌کار گرفته شده است.

۲. معرفی مجموعه‌داده

مجموعه‌دادهٔ مورداستفاده یک مجموعه‌دادهٔ شبیه‌سازی‌شده شامل هزینه‌های پزشکی برای بیماران در ایالات متحده است. این داده‌ها با استفاده از آمار جمعیتی ادارهٔ آمار امریکا برای کتاب (Lantz, ۲۰۱۹) تهیه شده است و در پلتفرم Kaggle منتشر شده است[۱] و در حد خود، منعکس‌کنندهٔ شرایط دنیای واقعی است. مجموعه‌دادهٔ مجموعه، شامل ۱۳۳۸ نمونه از ذی‌نفعانی است که در حال حاضر در طرح بیمه ثبت‌نام کرده‌اند، ویژگی‌های سن، جنسیت، شاخص تودهٔ بدنی(BMI)، سیگاری‌بودن، تعداد افراد تحت تکفل ، منطقه و هزینهٔ سالانه که نشان‌دهندهٔ ویژگی‌های بیمار و همچنین کل هزینه‌های پزشکی پرداخت‌شده به طرح برای سال تقویمی مورد توجه قرار گرفته است. توضیح هر یک از ویژگی‌ها در جدول ۱ ارائه شده است. جدول ۲ نیز چند نمونه از داده‌های مجموعه‌داده را نشان می‌دهد. همان‌طور که مشاهده می‌شود ویژگی‌های sex ، smoker و region ، دارای مقادیر غیرعددی هستند و بقیهٔ ویژگی‌ها، مقادیر به‌صورت عددی هستند.

جدول ۱: توضیح ویژگی‌های مجموعه‌داده

ویژگی	توضیح
سن (age)	یک عدد صحیح که سن ذی‌نفع اولیه را نشان می‌دهد (به‌استثنای افراد بالای ٦٤ سال، زیرا عموماً تحت پوشش دولت هستند)
جنسیت (sex)	جنسیت دارندهٔ بیمه‌نامهٔ اعم از مرد یا زن
شاخص تودهٔ بدنی (BMI)	شاخص تودهٔ بدنی (BMI) که این حس را ارائه می‌دهد که یک فرد نسبت به قد خود، چقدر اضافه یا کم‌وزن دارد. BMI برابر است با وزن (به کیلوگرم) تقسیم بر مجذور قد (به متر). BMI ایده‌آل در محدودهٔ ۱۸.۵ تا ۲٤.۹ است.
تعداد افراد تحت تکفل (children)	یک عدد صحیح که نشان‌دهندهٔ تعداد فرزندان / افراد تحت تکفل تحت پوشش طرح بیمه است
سیگاری بودن (smoker)	یک متغیر قطعی بله یا خیر که نشان می‌دهد آیا بیمه‌شده به طور منظم سیگار می‌کشد یا خیر
منطقه (region)	محل سکونت ذی‌نفع در ایالات متحده، به چهار منطقهٔ جغرافیایی تقسیم شده است: شمال شرقی، جنوب شرقی، جنوب غربی یا شمال غربی
هزینهٔ سالانه (charges)	هزینه‌های درمانی هر فرد مشمول طرح بیمه در سال را نشان می‌دهد

[۱] https://www.kaggle.com/datasets/mirichoi۰۲۱۸/insurance/data

جدول ۲: نمونه‌ای از داده‌های مجموعه‌داده

ویژگی‌ها							
#	سن	جنسیت	شاخص تودهٔ بدنی	افراد تحت تکفل	سیگاری بودن	منطقه	هزینهٔ سالانه
۰	۱۹	female	۲۷,۹۰۰	۰	yes	southwest	۱۶۸۸۴,۹۲۴۰۰
۱	۱۸	male	۳۳,۷۷۰	۱	no	southeast	۱۷۲۵,۵۵۲۳۰
۲	۲۸	male	۳۳,۰۰۰	۳	no	southeast	۴۴۴۹,۴۶۲۰۰
۳	۳۳	male	۲۲,۷۰۵	۰	no	northwest	۲۱۹۸۴,۴۷۰۶۱
۴	۳۲	male	۲۸,۸۸۰	۰	no	northwest	۳۸۶۶,۸۵۵۲۰

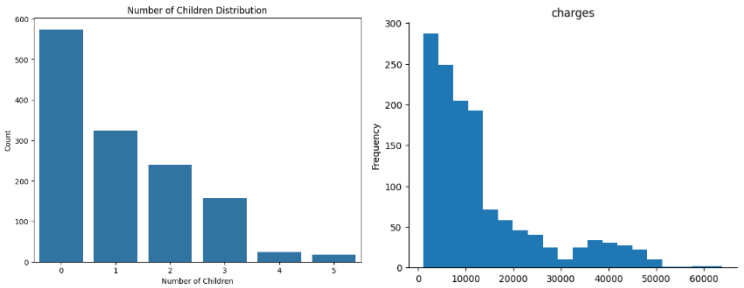
۳. بررسی و تحلیل داده‌ها

مرحلهٔ تحلیل و بررسی مجموعه‌دادهٔ فرایند تحقیق، شامل یک کاوش عمیق در مجموعه‌داده‌ها برای کشف الگوها، روابط و بینش‌هایی است که می‌تواند به آماده‌سازی و تحلیل مدل بعدی کمک کند(Islam et al., ۲۰۲۳).

با بررسی جدول ۳ که تحلیل آماری ویژگی‌های عددی مجموعه‌داده را نشان می‌دهد، مشخص می‌شود که تعداد مقادیر هر یک از ویژگی‌ها ۱۳۳۸ است و مجموعه‌داده دارای دادهٔ ازدست رفته نیست. همچنین بازهٔ مقادیر و میانگین برای هر ویژگی نیز مشخص می‌شود. برهمین اساس میانگین سنی در این مجموعه‌داده حدود ۳۹ سال و از بازهٔ ۱۸ الی ٦٤ سال است. برای ویژگی BMI نیز، بازهٔ از ۱۵/۹٦ الی ۵۳/۱۳ و میانگین برابر ۳۰/٦٦ است که نشان‌دهندهٔ خارج بودن افراد از حالت ایده‌آل و اضافه‌وزن در آن‌ها است. ویژگی تعداد افراد تحت تکفل نیز از بازهٔ ۰ الی ۵ است و میانگین آن، حدود ۱ است. برای ویژگی هزینهٔ سالانه نیز میانگین ۱۳۲۷۰/٤٤ دلار و بازهٔ از ۱۱۲۱/۸۷ الی ٦۳۷۷۰/٤۲ دلار است. نمودار توزیع ویژگی‌های افراد تحت تکفل و هزینهٔ سالانه (شکل۱) نیز نشان می‌دهد که توزیع تعداد افراد تحت تکفل، پله‌ای است و اکثر بیمه‌شوندگان، هزینهٔ سالانه کمتر از ۱۵۰۰۰ دلار پرداخت کرده‌اند.

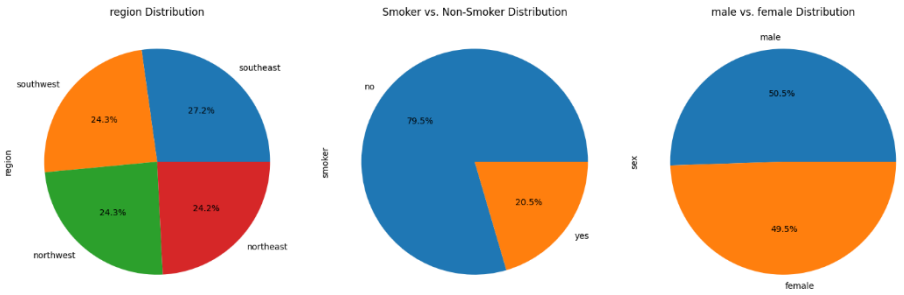
جدول ۳: اطلاعات آماری ویژگی‌های مجموعه‌داده

اطلاعات آماری	ویژگی‌ها			
	سن	شاخص تودهٔ بدنی	افراد تحت تکفل	هزینهٔ سالانه
تعداد	۱۳۳۸.۰۰۰۰۰۰	۱۳۳۸.۰۰۰۰۰۰	۱۳۳۸.۰۰۰۰۰۰	۱۳۳۸.۰۰۰۰۰۰
میانگین	۳۹.۲۰۷۰۲۵	۳۰.۶۶۳۳۹۷	۱.۰۹۴۹۱۸	۱۳۲۷۰.۴۲۲۲۶۵
انحراف معیار	۱۴.۰۴۹۹۶۰	۶.۰۹۸۱۸۷	۱.۲۰۵۴۹۳	۱۲۱۱۰.۰۱۱۲۳۷
کمینه	۱۸.۰۰۰۰۰۰	۱۵.۹۶۰۰۰۰	۰.۰۰۰۰۰۰	۱۱۲۱.۸۷۳۹۰۰
۲۵٪	۲۷.۰۰۰۰۰۰	۲۶.۲۹۶۲۵۰	۰.۰۰۰۰۰۰	۴۷۴۰.۲۸۷۱۵۰
۵۰٪	۳۹.۰۰۰۰۰۰	۳۰.۴۰۰۰۰۰	۱.۰۰۰۰۰۰	۹۳۸۲.۰۳۳۰۰۰
۷۵٪	۵۱.۰۰۰۰۰۰	۳۴.۶۹۳۷۵۰	۲.۰۰۰۰۰۰	۱۶۶۳۹.۹۱۲۵۱۵
بیشینه	۶۴.۰۰۰۰۰۰	۵۳.۱۳۰۰۰۰	۵.۰۰۰۰۰۰	۶۳۷۷۰.۴۲۸۰۱۰



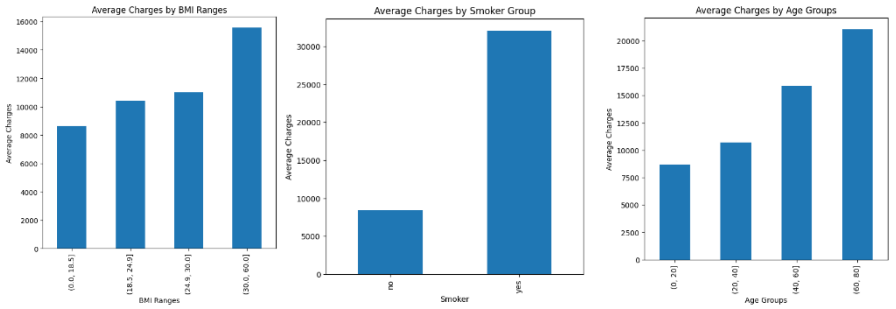
شکل ۱: نمودار توزیع هزینهٔ سالانه(راست) و تعداد افراد تحت تکفل(چپ)

شکل ۲ توزیع داده‌های غیرعددی را نشان می‌دهد. بررسی ویژگی جنسیت نشان می‌دهد که توزیع جنسیت بیمه‌شوندگان زن و مرد، تقریباً یکسان است و در مورد منطقه نیز هر چهار منطقه، تقریباً دارای توزیع یکسانی هستند اما ویژگی سیگاری‌بودن متفاوت از این دو ویژگی است و افراد غیر سیگاری در مقابل افراد سیگاری، بخش قابل‌توجهی (۷۹/۵٪) از بیمه‌شوندگان را تشکیل می‌دهند.

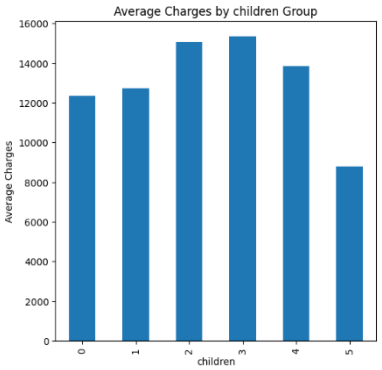


شکل ۲: توزیع ویژگی های جنسیت(راست)، سیگاری‌بودن(وسط) و منطقه(چپ)

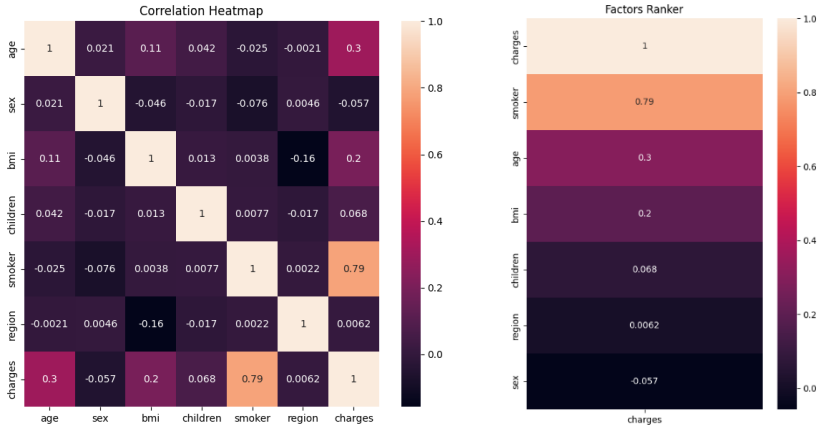
برای تحلیل هزینه‌های پرداختی و عوامل مؤثر در بیماری، سه ویژگی سن، سیگاری‌بودن و BMI، گروه‌بندی شدند تا میانگین هزینه‌کرد هر گروه مشخص شود (شکل ۳). به همین منظور ویژگی سن به گروه‌های (۱۸–۲۰)، (۲۰–۴۰)، (۴۰–۶۰) و (۶۰–۸۰) تقسیم شده و ویژگی BMI به گروه‌های (۱۵–۱۸/۵)، (۱۸/۵–۲۴/۹)، (۲۴/۹–۳۰) و (۳۰–۶۰) تقسیم شده است. همان‌طور که انتظار می‌رفت با افزایش سن و BMI، میزان هزینهٔ سالانهٔ بیمار افزایش می‌یابد. همچنین بررسی ویژگی سیگاری‌بودن، نشان می‌دهد افراد سیگاری، هزینهٔ بسیار زیادی نسبت به افراد غیر سیگاری پرداخت می‌کنند. به‌عبارت‌دیگر عوامل افراد مسن، دارای اضافه‌وزن و سیگاری، بیش‌تر در معرض بیماری قرار دارند و نسبت به بقیهٔ گروه‌ها هزینهٔ بیشتری پرداخت می‌کنند. در مورد بقیهٔ ویژگی‌ها مانند جنسیت، منطقه و افراد تحت تکفل، بررسی‌ها نشان می‌دهد که گروه‌های مختلف، تفاوت قابل توجهی با یکدیگر ندارند. البته در مورد افراد تحت تکفل برخلاف تصور، با افزایش تعداد افراد تحت تکفل، هزینه افزایش نمی‌یابد و در برخی موارد کاهش را نیز تجربه می‌کند (شکل ٤).



شکل ۳: نمودار میانگین هزینه سالانه برای گروه‌های سنی(راست)، سیگاری(وسط) و BMI(چپ)



شکل ٤: نمودار میانگین هزینهٔ سالانه بر اساس تعداد افراد تحت تکفل



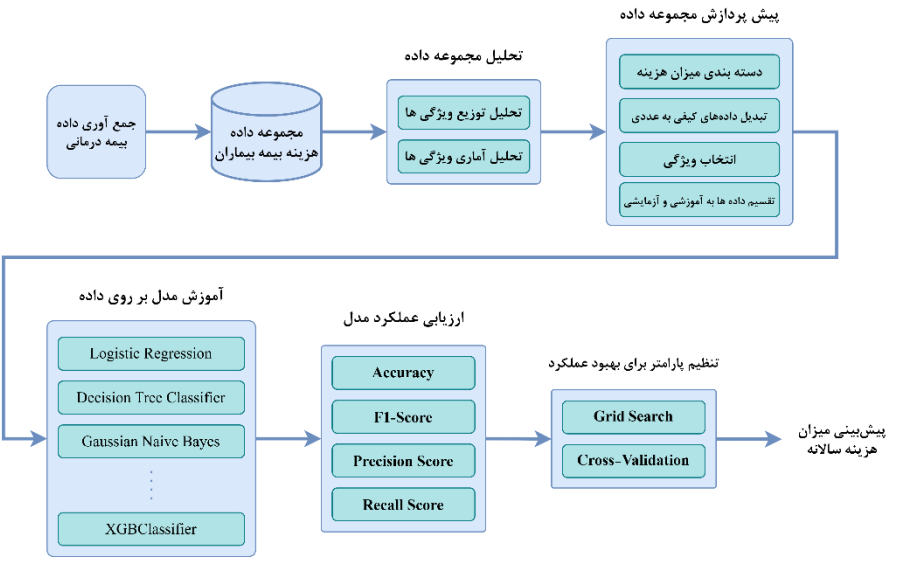
شکل ٥: نقشه حرارتی ویژگی‌های مجموعه‌داده(چپ) و رتبه‌بندی همبستگی ویژگی‌ها بر اساس هزینه سالانه(راست)

نقشهٔ حرارتی حاصل در شکل ۵، یک نمایش بصری از همبستگی بین ویژگی‌ها در مجموعه‌داده ارائه می‌دهد. هر سلول در نقشهٔ حرارتی، باتوجه‌به قدرت و جهت همبستگی، رنگ می‌شود. بررسی نقشهٔ حرارتی، نشان می‌دهد که سیگاری‌بودن، سن و BMI، بیشترین همبستگی و تأثیر را بر روی هزینهٔ سالانه دارند (شکل ۵ رتبه‌بندی). ویژگی‌های تعداد افراد تحت تکفل، منطقه و جنسیت نیز به ترتیب تأثیر مثبت، خنثی و منفی، بر روی هزینهٔ سالانه دارند. در میان ویژگی‌های مجموعه‌داده، می‌توان گفت که سیگاری‌بودن با حدود ۸۰٪ ارتباط مستقیم با هزینه‌های درمانی، با اختلاف، بیشترین عامل افزایش هزینه‌های درمانی است.

به‌طورکلی افراد سیگاریِ مسنِ چاق، مستعدترین افراد به بیماری و پرداخت هزینه‌های بیشتر برای درمان بیماری هستند.

٤. روش پیشنهادی

فرایند پیشنهادی برای حل مسئله این است که پس از تحلیل مجموعه‌داده و به‌دست‌آوردن دید کلی از ویژگی‌ها، به آماده‌سازی و پیش‌پردازش آن‌ها پرداخته می‌شود تا به داده‌های مناسب برای اعمال الگوریتم‌های طبقه‌بندی تبدیل شوند. سپس از ۹ الگوریتم طبقه‌بندی برای پیش‌بینی میزان هزینهٔ سالانه (کم، متوسط و زیاد) استفاده می‌شود و با ارزیابی و مقایسهٔ نتایج هرکدام، بهترین الگوریتم انتخاب می‌شود. در انتها نیز با استفاده از روش Grid Search همراه با cross-validation ، به بهبود پارامترهای الگوریتم منتخب پرداخته می‌شود و در نهایت مدل پیش‌بینی میزان هزینهٔ سالانه برای بیمهٔ درمانی به دست می آید (شکل ٦).



شکل ٦ : روش پیشنهادی برای پیش‌بینی هزینهٔ سالانهٔ بیمهٔ درمانی

٤-١.پیش‌پردازش

پیش‌پردازش مجموعه‌داده، می‌تواند تأثیر بسیار زیادی در نتایج ارزیابی‌ها داشته باشد. در پیش‌پردازش داده، سعی می‌شود به حالت و وضعیت مناسب برای اعمال الگوریتم تبدیل شود. در مورد داده‌های جنسیت، سیگاری‌بودن و منطقه که ماهیتی غیرعددی (کیفی) دارند، تبدیل به

مقادیر عددی ضـــروری اســـت؛ چرا که الگوریتم‌ها بر اســـاس مقادیر عددی، قادر به پیش‌بینی هســـتند. به همین منظور با اســـتفاده از روش‌های جای‌گذاری مقادیر غیرعددی مانند male، female، yes، no و...، به یک عدد نگاشت می‌شوند.

مرحلهٔ بعدی در پیش‌پردازش داده‌ها، دســـته‌بندی میزان هزینهٔ ســـالانه اســـت. این کار، قابلیت طبقه‌بندی را به مجموعه‌داده می‌افزاید. برای دســـته‌بندی هزینهٔ ســـالانه، هزینهٔ ســـالانه با تعداد نمونهٔ مساوی برای هر بازه، بازه‌بندی می‌شود و به هر کدام از بازه‌ها یک عدد اختصاص می‌یابد که به آن کلاس می‌گویند که کلاس ۱، ۲ و ۳به ترتیب نمایندهٔ هزینه‌کرد کم، متوســـط و زیاد هستند. بازهٔ کلاس‌ها به‌صورت کلاس۱ بازهٔ ۱۱۲۱/۸۷ تا ٦۲۵۰/٤۳ دلار، کلاس۲ بازهٔ ٦۲۷۲/٤۷ تا ۱۲۸۱۵/٤٤دلار و کلاس۳ بازهٔ ۱۲۸۲۹/٤۵ تا ٦۳۷۷۰/٤۲ دلار اســـت. جدول ٤، چند نمونه از مجموعه‌داده را پس از پیش‌پردازش نشان می‌دهد.

جدول ٤ : نمونه‌ای از داده‌های مجموعه‌داده بعد از پیش‌پردازش

ویژگی‌ها								
#	سن	جنسیت	شاخص تودهٔ بدنی	افراد تحت تکفل	سیگاری بودن	منطقه	هزینهٔ سالانه	کلاس
۰	۱۹	۱	۲۷.۹۰۰	۰	۱	۰	۱٦۸۸٤.۹۲٤۰۰	۲
۱	۱۸	۰	۳۳.۷۷۰	۱	۰	۱	۱۷۲۵.۵۵۲۳۰	۰
۲	۲۸	۰	۳۳.۰۰۰	۳	۰	۱	٤٤٤۹.٤٦۲۰۰	۰
۳	۳۳	۰	۲۲.۷۰۵	۰	۰	۲	۲۱۹۸٤.٤۷۰٦۱	۲
٤	۳۲	۰	۲۸.۸۸۰	۰	۰	۲	۳۸٦٦.۸۵۵۲۰	۰

در آخرین مرحله نیز نوبت به انتخاب ویژگی و تقســـیم مجموعه‌داده به داده‌های آموزشـــی[۱] و آزمایشی[۲] می‌رسد. باتوجه‌به تحلیل‌های انجام‌شده بر روی مجموعه‌داده و بینش به‌دست‌آمده، همهٔ ویژگی‌ها بر روی هزینهٔ سالانه تأثیرگذارند به‌استثنای ویژگی منطقه با درصد تأثیر ۰/۰۰٦۲ که تقریباً حالت خنثی دارد. بنابراین در مرحلهٔ انتخاب ویژگی، برای کاهش محاســـبات، ویژگی منطقه، حذف می‌شـــود و ویژگی‌های ســـن، جنســـیت، افراد تحت تکفل، ســـیگاری‌بودن و BMI، برای آموزش الگوریتم انتخاب می‌شـــوند. تقســـیم داده‌ها نیز با نســـبت ۸۰ به ۲۰ انجام می‌شـــود؛

[۱] Training Set

[۲] Test Set

یعنی ۸۰ درصد از نمونه‌های مجموعه‌داده برای آموزش الگوریتم استفاده می‌شوند و ۲۰ درصد از آن‌ها برای ارزیابی عملکرد مدل آموزش دیده استفاده می‌شوند.

۲-٤.آموزش مدل بر روی داده‌ها

طبقه‌بندی، یک تکنیک یادگیری تحت نظارت است که تابعی را از مجموعه‌داده‌های آموزشی می‌آموزد که شامل ویژگی‌های ورودی و خروجی طبقه‌بندی شده است. این تابع برای پیش‌بینی برچسب کلاس برای هر بردار ورودی معتبر، استفاده می‌شود. هدف اصلی طبقه‌بندی، استفاده از الگوریتم‌های یادگیری ماشین برای دستیابی به بهترین دقت پیش‌بینی است(Syarif et al., ۲۰۱٦). در این پژوهش، آموزش مدل با استفاده از ۹ الگوریتم طبقه‌بندی صورت می‌گیرد، به این صورت که این الگوریتم‌ها با استفاده از داده‌های آموزشی، آموزش می‌بینند و مدل طبقه‌بندی برای پیش‌بینی دسته(کم، متوسط، زیاد) هزینه ایجاد می‌شوند سپس این مدل‌ها با استفاده از داده‌های آزمایشی، ارزیابی می‌شوند. در جدول ۵، توضیح مختصری برای هریک از الگوریتم‌ها ارائه شده است.

جدول ۵: توضیح الگوریتم‌های طبقه‌بندی

الگوریتم	توضیح
درخت تصمیم[۱]	یک الگوریتم طبقه‌بندی است که از یک مدل درخت‌مانند از تصمیم‌ها و پیامدهای احتمالی آنها استفاده می‌کند. این مجموعه‌داده را بر اساس مقادیر ویژگی به شاخه‌های مختلف تقسیم می‌کند و قوانین تصمیم‌گیری را برای طبقه‌بندی نمونه‌ها ایجاد می‌کند.(Loh, ۲۰۱۱)
جنگل تصادفی[۲]	یک روش یادگیری گروهی است که چندین طبقه‌بندی درخت تصمیم را برای پیش‌بینی ترکیب می‌کند. این یک جنگل از درختان تصمیم را ایجاد می‌کند که در آن هر درخت بر روی یک زیرمجموعهٔ تصادفی از داده‌ها و ویژگی‌های آموزشی، آموزش داده می‌شود. پیش‌بینی نهایی با تجمیع پیش‌بینی‌های همهٔ درخت‌های تصمیم‌گیری فردی انجام می‌شود.(Breiman, ۲۰۰۱)
رگرسیون لجستیک[۳]	یک الگوریتم طبقه‌بندی است که احتمال یک نمونهٔ متعلق به یک کلاس خاص را مدل می‌کند. از یک تابع لجستیک، برای ترسیم ویژگی‌های ورودی به‌احتمال نتیجهٔ باینری استفاده می‌کند.(Hosmer Jr et al., ۲۰۱۳)

[۱] Decision Tree
[۲] Random Forest
[۳] Logistic Regression

Ridge	گونه‌ای از رگرسیون لجستیک است که یک اصطلاح منظم‌سازی به نام منظم‌سازی ریج را در خود جای‌داده است. این منظم‌سازی با افزودن یک عبارت جریمه به تابع ضرر به کاهش بیش از حد برازش کمک می‌کند.(Marquardt & Snee, ۱۹۷۵)
KNN[۱]	یک الگوریتم غیرپارامتری است که نمونه‌ها را بر اساس شباهت آن‌ها با نمونه‌های همسایه در فضای ویژگی، طبقه‌بندی می‌کند. KNN یک نمونهٔ آزمایشی را با درنظرگرفتن کلاس اکثریت k همسایهٔ نزدیکش برچسب‌گذاری می‌کند.(Zhang et al., ۲۰۱۷)
AdaBoost	یک الگوریتم یادگیری گروهی است که چندین طبقه‌بندی ضعیف را برای ایجاد یک طبقه‌بندی قوی ترکیب می‌کند. در هر تکرار، وزن‌های بالاتری را به نمونه‌های طبقه‌بندی‌شدهٔ اشتباه، اختصاص می‌دهد و به طبقه‌بندی‌کننده‌های ضعیف بعدی، اجازه می‌دهد تا روی موارد دشوار تمرکز کنند. (Schapire & Freund, ۲۰۱۳)
گاوسی ساده‌لوح بیز[۲]	یک الگوریتم احتمالی بر اساس قضیهٔ بیز است. فرض می‌کند که ویژگی‌ها باتوجه‌به کلاس، به‌صورت شرطی مستقل هستند و از توزیع گاوسی (عادی) پیروی می‌کنند.(Rish, ۲۰۰۱)
ماشین بردار پشتیبانی[۳]	طبقه‌بندی‌کنندهٔ قدرتمندی است که با ساخت ابرصفحه‌ها در یک فضای ویژگی با ابعاد بالا، نمونه‌ها را از هم جدا می‌کند. هدف SVM، به حداکثر رساندن حاشیهٔ بین کلاس‌های مختلف است که منجر به تعمیم بهتر می‌شود. می‌تواند هر دو وظیفهٔ طبقه‌بندی خطی و غیرخطی را با استفاده از توابع کرنل مختلف انجام دهد.(Cortes & Vapnik, ۱۹۹۵)
XGBClassifier	یک الگوریتم تقویت گرادیان بهینه است که ترکیبی از مدل‌های مبتنی بر درخت و تکنیک‌های منظم‌سازی برای دستیابی به عملکرد پیش‌بینی بالا را استفاده قرار می دهد. از یک الگوریتم بهینه‌سازی جدید و محاسبات موازی، برای افزایش سرعت آموزش و دقت مدل استفاده می‌کند.(Tianqi & Guestrin, ۲۰۱۶)

۳-٤. ارزیابی عملکرد مدل

برای ارزیابی عملکرد مدل‌های طبقه‌بندی حاصل از آموزش الگوریتم‌های طبقه‌بندی، نیاز به استفاده از معیارهای مناسب برای مسائل طبقه‌بندی است. برای مسائل طبقه‌بندی، ارزیابی تبعیض(discrimination evaluation) بهترین راه‌حل (بهینه) در طول آموزش طبقه‌بندی، می‌تواند بر اساس ماتریس سردرگمی تعریف شود. همان‌طور که در جدول ٦ نشان داده شده است. ردیف جدول، کلاس پیش‌بینی‌شده را نشان می‌دهد، در حالی که ستون، کلاس واقعی را نشان می‌دهد. از این ماتریس درهم ریختگی[٤]، tp و tn نشان‌دهندهٔ تعداد نمونه‌های مثبت و منفی هستند که به‌درستی طبقه‌بندی شده‌اند. در همین حال، fp و fn به ترتیب تعداد نمونه‌های منفی و مثبت به‌اشتباه طبقه‌بندی شده را نشان می‌دهند(Hossin & Sulaiman, ۲۰۱۵). از تعمیم

[۱] K-Nearest Neighbors
[۲] Gaussian Naive Bayes
[۳] Support Vector Machine
[٤] confusion matrix

چندکلاسی جدول ٦ ، چهار معیار رایج می‌توانند برای ارزیابی عملکرد طبقه‌بندی‌کننده استفاده شوند، همان‌طور که در جدول ۷ نشان داده شده‌اند. پس از ارزیابی تمام الگوریتم‌ها با معیارهای زیر و مقایسـهٔ نتایج، بهترین الگوریتم به‌عنوان الگوریتم منتخب برای مراحل بعدی انتخاب می‌شود.

جدول ٦: ماتریس درهم ریختگی برای طبقه‌بندی دو کلاسی(Hossin & Sulaiman, ۲۰۱۵)

	در واقعیت مثبت	در واقعیت منفی
پیش‌بینی مثبت	**مثبت صادق[۱] (tp)**	**مثبت کاذب[۲] (fp)**
پیش‌بینی منفی	**منفی کاذب[۳] (fn)**	**منفی صادق[٤] (tn)**

جدول ۷ : معرفی معیارهای ارزیابی مدل‌های طبقه‌بندی‌کننده(Hossin & Sulaiman, ۲۰۱۵)

معیار ارزیابی	فرمول	توضیح
دقت[٥]	$acc = \frac{tp + tn}{tp + fp + tn + fn}$ (۱)	به‌طورکلی، معیار دقت نسبت پیش‌بینی‌های صحیح را بر تعداد کل موارد ارزیابی‌شده اندازه‌گیری می‌کند.
صحت[٦]	$p = \frac{tp}{tp + fp}$ (۲)	صحت برای اندازه‌گیری الگوهای مثبتی که به‌درستی از کل الگوهای پیش‌بینی‌شده در یک کلاس مثبت پیش‌بینی شده‌اند، استفاده می‌شود.
حساسیت[۷]	$r = \frac{tp}{tp + tn}$ (۳)	حساسیت برای اندازه‌گیری کسری از الگوهای مثبت که به‌درستی طبقه‌بندی شده‌اند استفاده می‌شود.
F۱-Score	$F1 = 2 \times \frac{p \times r}{p + r}$ (٤)	این معیار، نشان‌دهندهٔ میانگین هارمونیک بین مقادیر حساسیت و صحت است.

[۱] True Positive
[۲] False Positive
[۳] False Negative
[٤] True Negative
[٥] Accuracy
[٦] Precision
[۷] Recall

۵-٤. بهبود عملکرد

به‌طورکلی، اکثر الگوریتم‌های یادگیری ماشین درصورتی‌که پارامترهای آن‌ها به‌درستی تنظیم نشوند، به نتایج بهینه دست نخواهند یافت. برای ساخت یک مدل طبقه‌بندی با دقت بالا، انتخاب یک الگوریتم قدرتمند یادگیری ماشین و همچنین تنظیم پارامترهای آن، بسیار مهم است. بهینه‌سازی پارامتر اگر به‌صورت دستی انجام شود می‌تواند بسیار وقت‌گیر باشد، به‌خصوص زمانی که الگوریتم یادگیری، دارای پارامترهای زیادی باشد(Syarif et al., ۲۰۱٦).

در این مقاله از روش Grid Search همراه با cross-validation برای بهبود پارامترهای الگوریتم منتخب استفاده شده است. Grid Search در اصل یک جستجوی جامع است که بر اساس زیرمجموعهٔ تعریف‌شدهٔ فضای فراپارامتر است. فراپارامترها با استفاده از مقدار حداقل (کران پایین)، مقدار حداکثر (کران بالایی) و تعداد مراحل، مشخص می‌شوند. سه مقیاس مختلف وجود دارد که می‌توان از آنها استفاده کرد: مقیاس خطی، مقیاس درجهٔ دوم و مقیاس لگاریتمی. در این روش، عملکرد هر ترکیب از پارامترها با استفاده از معیارهای عملکرد، ارزیابی می‌شود(Syarif et al., ۲۰۱٦).

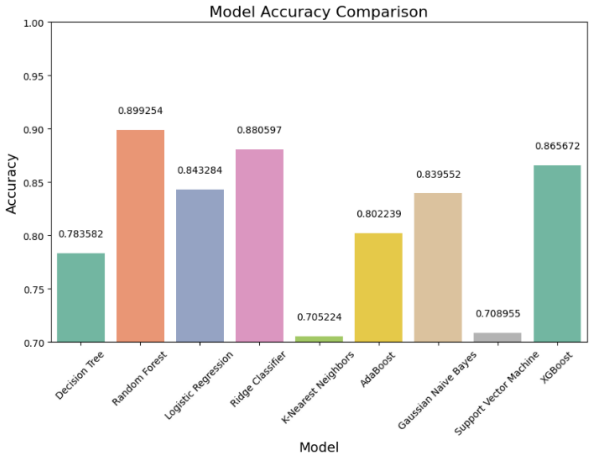
۵. بحث و نتایج

در تمامی مراحل این مقاله، از سرویس COLAB شرکت گوگل با سخت‌افزار بدون GPU و ۱۲.۷ G Ram همراه با Python۳ استفاده شده است. بررسی نتایج تحلیل مجموعه‌داده، نشان داد که عوامل سیگارکشیدن، سن و اضافه‌وزن، اثرات قابل‌توجهی بر روی هزینه‌های درمان دارند که از بین این عوامل، می‌توان سیگارکشیدن و اضافه‌وزن را کنترل نمود. این نتیجه‌گیری برای شرکت‌های بیمه، برای مقایسه و توسعهٔ محصولات بیمه‌ای مؤثرتر و ارائهٔ خدمات بیمه‌ای بهتر، مفید است. علاوه بر این، ممکن است در ارزیابی برنامه‌های مدیریت سلامت جمعیت، مفید باشد.

جدول ۸: نتایج ارزیابی الگوریتم‌های طبقه‌بندی با معیارهای دقت، صحت، حساسیت و F۱-Score (بهترین نتایج پررنگ شده است)

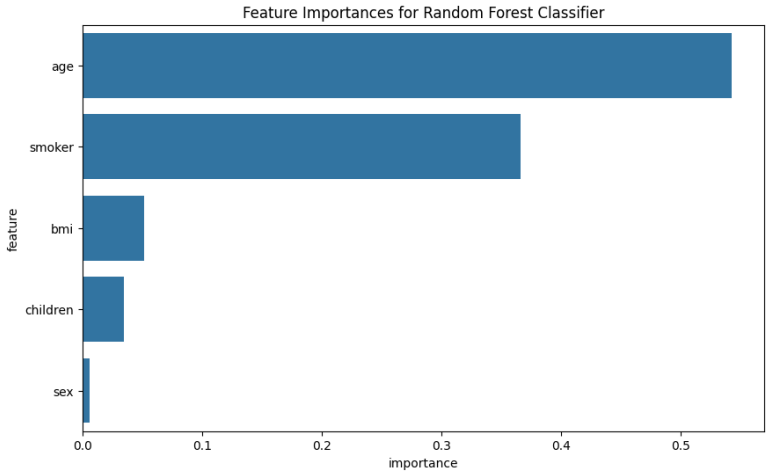
الگوریتم طبقه‌بندی	دقت (acc)	صحت (p)	حساسیت (r)	F۱-Score
درخت تصمیم	۰.۷۸	۰.۷۹	۰.۷۸	۰.۷۸
جنگل تصادفی	**۰.۹۰**	**۰.۹۰**	**۰.۸۹**	**۰.۹۰**
رگرسیون لجستیک	۰.۸٤	۰.۸۸	۰.۸۳	۰.۸٤
Ridge	۰.۸۸	**۰.۹۰**	۰.۸۷	۰.۸۷
KNN[۱]	۰.۷۱	۰.۷۰	۰.٦۹	۰.٦٦
AdaBoost	۰.۸۰	۰.۸۰	۰.۸۰	۰.۸۰
گاوسی ساده‌لوح بیز	۰.۸٤	۰.۸۸	۰.۸۳	۰.۸٤
ماشین بردار پشتیبانی	۰.۷۱	۰.۷٤	۰.٦۹	۰.٦٥
XGBClassifier	۰.۸۷	۰.۸۷	۰.۸٦	۰.۸٦

نتایج ارزیابی الگوریتم‌های طبقه‌بندی در جدول ۸ ارائه شده است. بررسی نتایج ارزیابی نشان می‌دهد که الگوریتم جنگل تصادفی بادقت ۹۰٪، نسبت به دیگر الگوریتم‌ها عملکرد بهتری داشته و در هر چهار معیار ارزیابی، امتیاز بالاتری کسب نموده است. در رتبهٔ بعدی، الگوریتم Ridge قرار دارد که بادقت ۸۸٪، پیش‌بینی صحیح هزینهٔ سالانه را انجام می‌دهد. الگوریتم XGBClassifier نیز بادقت ۸۷٪ در رتبهٔ ۳ قرار می‌گیرد (شکل ۷). بنابراین الگوریتم جنگل تصادفی به‌عنوان الگوریتم منتخب انتخاب می‌شود.



شکل ۷ : نمودار مقایسهٔ دقت مدل‌های طبقه‌بندی

[۱] K-Nearest Neighbors



شکل ۸ : میزان اهمیت ویژگی‌ها برای مدل جنگل تصادفی

برای افزایش توانایی مدل در پیش‌بینی هزینهٔ ســـالانه، از روش بهبود پارامتر Grid Search بر روی الگوریتم منتخب اســـتفاده می‌شـــود. مقادیر پارامتر حاصـــل از اعمال روش Grid Search همراه با cross-validation بر روی الگوریتم جنگل تصـــادفی، برابر با ('bootstrap': False, 'criterion': 'gini', 'max_depth': ۸, 'n_estimators': ۱۰۰) اســـت. ارزیـابی الگوریتم پس از تنظیم مقـادیر پارامترهای به‌دست‌آمده، نشـان می‌دهد، میزان دقت به ۹۱٪، صـحت به ۹۲٪، حسـاسـیت به ۹۰٪ و F۱-Score به ۹۱٪ رسـیده اسـت و عملکرد جنگل تصـادفی در سـه معیار، بهبود یافته اسـت. بررسـی اهمیت ویژگی‌ها برای مدل به‌دست‌آمده نشان می‌دهد که این مدل از سن و سیگاری بودن برای پیش‌بینی دستهٔ هزینه (کم، متوسـط و زیاد) بیشـتر اسـتفاده کرده اسـت (شـکل ۸). به‌عبارت‌دیگر برای مدل منتخب، تأثیرگذارترین عامل در پیش‌بینی دستهٔ هزینه، سن است.

٦. نتیجه‌گیری

بیمهٔ درمانی یکی از راه‌حل‌های افزایش وضـعیت سـلامتی جامعه اسـت؛ چرا که توانایی افراد برای درمان بیماری خود را با پرداخت هزینه، افزایش می‌دهد. از مهم‌ترین چالش‌های صـنعت بیمه، پیش‌بینی هزینه‌های بیمهٔ درمانی افراد اســـت. برای برآورد هزینه‌های بیمه و نرخ‌گذاری بیمه، پیش‌بینی هزینهٔ بیمهٔ درمانی موردنیاز است.

در این مقاله با استفاده از مجموعه‌دادهٔ بیمه‌شوندگان ادارهٔ آمار امریکا، به بررسی عوامل مؤثر بر هزینهٔ درمان پرداخته شـد و با دسـته‌بندی میزان هزینهٔ پرداختی به کم، متوسـط و زیاد، امکان طبقه‌بندی هزینهٔ سـالانه فراهم شـد. مبتنی بر نتایج به‌دسـت‌آمده بر اسـاس این مجموعه‌داده، می‌توان گفت که سـیگارکشـیدن، تأثیر بسـیار زیادی بر هزینه‌های درمانی دارد و افزایش سـن و

اضافه‌وزن (چاقی) به ترتیب بعد از آن قرار دارند. اما جنسیت، منطقه و تعداد افراد تحت تکفل تأثیر زیادی بر هزینه‌های درمانی ندارند. در طبقه‌بندی نیز عملکرد الگوریتم جنگل تصادفی بادقت ۹۱٪ ، نسبت به سایر الگوریتم‌ها بهتر است و دقت مناسبی برای برآورد کلی میزان هزینه‌کرد (کم، متوسط و زیاد) افراد برای درمان بیماری دارد. این نتایج، می‌تواند برای ارزیابی برنامه‌های مدیریت سلامت جمعیت استفاده شود؛ علاوه بر این، این رویکرد می‌تواند به رفاه اجتماعی برای سیاست‌گزاری و تصمیم‌گیری سریع‌تر در پیش‌بینی هزینهٔ بیمهٔ پزشکی، کمک کند.

محققان می‌توانند با جمع‌آوری داده‌های معتبر در حوزهٔ بیمهٔ درمان، زمینه را برای بررسی و تحلیل وضعیت بیمهٔ درمانی در ایران فراهم کنند و به کمک مجموعه‌داده‌های باکیفیت، امکان پیش‌بینی دقیق هزینهٔ درمانی برای بیماران را فراهم نمایند.

تعارض منافع

تعارض منافع وجود ندارد.

کد ارکید

ORCID

Ali Zhaleh Karimi https://orcid.org/۰۰۰۹-۰۰۰۸-۱۱۱۵-٦٣٦٥

Ramin Dalir https://orcid.org/۰۰۰۹-۰۰۰۳-۹۰۰۰-۹۷۷X

منابع

۱. Arab, M., Fathian, M., & Aliahmadi Jeshfaghani, H. (۲۰۲۲). Forecast of Medical Expenses of Iran Health Insurance Organization Using Machine Learning Based Methods. Iranian Journal of Health Insurance, ۰-۰ .
۲. Breiman, L. (۲۰۰۱). Random forests. Machine learning, ٤٥, ٥-۳۲ .
۳. Cortes, C., & Vapnik, V. (۱۹۹٥). Support-vector networks. Machine learning, ۲۰, ۲۷۳-۲۹۷ .
٤. Dong, S., & Fei, D. (۲۰۲۱). Improve the interpretability by decision tree regression: exampled by an insurance dataset. ۲۰۲۱ International Conference on Computer Engineering and Artificial Intelligence (ICCEAI) ,
٥. Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (۲۰۱۳). Applied logistic regression (Vol. ۳۹۸). John Wiley & Sons .

٦. Hossin, M., & Sulaiman, M. N. (٢٠١٥). A review on evaluation metrics for data classification evaluations. International journal of data mining & knowledge management process, ٥(٢), ١ .

٧. Islam, M. A., Nag, A., Chandra, P., Fahim, S. F. A., & Hoque, M. M. (٢٠٢٣). Healthcare Cost Patterns and Prediction: Investigating Personal Datasets Using Data Analytics. Authorea Preprints .

٨. Lantz, B. (٢٠١٩). Machine learning with R: expert techniques for predictive modeling. Packt publishing ltd .

٩. Loh, W. Y. (٢٠١١). Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery, ١(١), ١٤-٢٣ .

١٠. Marquardt, D. W., & Snee, R. D. (١٩٧٥). Ridge regression in practice. The American Statistician, ٢٩(١), ٣-٢٠ .

١١. Rish, I. (٢٠٠١). An empirical study of the naive Bayes classifier. IJCAI ٢٠٠١ workshop on empirical methods in artificial intelligence ,

١٢. Schapire, R. E., & Freund, Y. (٢٠١٣). Boosting: Foundations and algorithms. Kybernetes, ٤٢(١), ١٦٤-١٦٦ .

١٣. Syarif, I., Prugel-Bennett, A., & Wills, G. (٢٠١٦). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. TELKOMNIKA (Telecommunication Computing Electronics and Control), ١٤(٤), ١٥٠٢-١٥٠٩ .

١٤. Tajaddodi Nodehi, M., Hosseini Khatibani, S., Yazdinejad, M., & Zolfi, S. (٢٠٢٣). Predicting people's health insurance costs using machine learning and ensemble learning methods. Iranian Journal of Insurance Research, ١٣(١), ١-١٤. https://doi.org/١٠,٢٢٠٥٦/ijir.٢٠٢٤,٠١,٠١

١٥. Tianqi, C., & Guestrin, C. (٢٠١٦). Xgboost: A scalable tree boosting system In Proceedings of the ٢٢nd acm sigkdd international conference on knowledge discovery and data mining .

١٦. Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (٢٠١٧). Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, ٢٩(٥), ١٧٧٤-١٧٨٥ .