مقاله پژوهشی

پیش‌بینی نوع بیماری مالاریا با استفاده از الگوریتم درخت تصمیم تقسیم و رگرسیون

مریم عاشوری[1*]، فاطمه حمزه‌وی[2]

۱-کارشناس ارشد مهندسی فناوری اطلاعات، مربی دانشکده فنی و مهندسی، مجتمع آموزش عالی سراوان، سراوان، ایران، مجتمع آموزش عالی سراوان ۲-کارشناس ارشد مهندسی کشاورزی، مربی دانشکده کشاورزی، مجتمع آموزش سراوان، سراوان، ایران

***نویسنده مسئول:** سیستان و بلوچستان، سراوان، بلوار پاسداران، مجتمع آموزش عالی سراوان، دانشکده فنی و مهندسی، گروه مهندسی فناوری اطلاعات، ۰۹۱۵۵۳۳۸۷۲۱، نمابر ۰۵۴۳۷۶۳۳۰۰۹۰،

پست الکترونیک: mashoori@saravan.ac.ir

دریافت:۹۷/۱/۲۸ پذیرش:۹۷/۵/۱۰

چکیده

مقدمه: *مالاریا یک بیماری عفونی است که سالانه ۲۰۰ تا ۳۰۰ میلیون نفر را گرفتار می‌سازد. ویژگی‌های محیطی چون بارش، درجه حرارت و رطوبت از جمله عوامل تاثیرگذار بر توزیع جغرافیایی و شیوع این بیماری هستند. همچنین، شرایط محیطی نیز بر فعالیت و وفور پشه ناقل بیماری مالاریا تاثیرگذار است. پژوهش حاضر با هدف ارائه مدلی برای پیش‌بینی نوع مالاریا صورت گرفت.*

روش کار: *پژوهش حاضر به روش توصیفی- مقطعی صورت گرفت. جامعه‌ی پژوهش متشکل از ۲۸۵ نفر مراجعه‌کننده به مرکز بهداشت سراوان از مرداد ۸۸ تا آذر ۹۵ بود. جهت تحلیل داده‌ها از نرم‌افزار کلمنتاین ۱۲ (Clementine12.0) استفاده شد. برای مدل‌سازی از الگوریتم‌های درخت تصمیم تقسیم و رگرسیون، آشکارساز تعامل خودکار مجذور کای، سی پنج و شبکه عصبی استفاده شد.*

یافته‌ها: *مقدار صحت به‌دست آمده از اجرای الگوریتم‌های درخت تصمیم تقسیم و رگرسیون، آشکارساز تعامل خودکار مجذور کای، سی پنج و شبکه عصبی به ترتیب ۰/۷۲۱۷، ۰/۶۶۹۸، ۰/۶۸۴۰ و ۰/۶۵۵۷ بود. مقادیر به‌دست آمده برای شاخص‌های حساسیت، شفافیت، صحت، دقت و ارزش اخباری منفی و سطح زیر منحنی مشخصه‌ی عملکرد سیستم برای مدل درخت تصمیم تقسیم و رگرسیون نشان‌دهنده‌ی عملکرد بهتر این الگوریتم نسبت به سایر الگوریتم‌ها بود. مقادیر حساسیت و سطح زیر منحنی برای مدل درخت تصمیم تقسیم و رگرسیون ۰/۵۷۸۷ و ۰/۶۶ بود.*

نتیجه‌گیری: *استفاده از روش‌های نوین داده‌کاوی برای تحلیل داده‌های مالاریا، نحوه‌ی نگرش مراکز تحقیقات مالاریا را نسبت به شناسایی نوع مالاریا با توجه به گونه حشره تغییر می‌دهد. شناسایی سریع‌تر و دقیق‌تر نوع مالاریا به تشخیص صحیح روش درمان مناسب کمک نموده، منجر به ارتقای عملکرد سازمان‌های سلامت می‌گردد.*

گل‌واژه‌گان: *مالاریا، درخت تصمیم، شبکه عصبی، منحنی ROC*

مقدمه

مالاریا از نظر بهداشتی-درمانی از اهمیت جهانی برخوردار است. مالاریا یکی از شش بیماری مطرح در برنامه‌ی پژوهش بیماری‌های گرمسیری سازمان جهانی بهداشت است. در ایران، علی‌رغم سرمایه‌گذاری‌ها و تلاش‌های زیادی که جهت فرونشاندن و کنترل این بیماری اعمال گردیده، بخش قابل‌ملاحظه‌ای از کشور هنوز درگیر مالاریا است و این بیماری همچنان یکی از مشکلات بهداشتی-درمانی درجه اول استان‌های جنوب و جنوب شرقی از جمله هرمزگان، سیستان و بلوچستان و قسمتی از نواحی گرمسیری استان کرمان (کهنوج و جیرفت) به‌شمار می‌رود[۱]. مالاریا نوعی بیماری انگلی است و عامل آن تک یاخته انگلی از جنس پلاسمودیوم است[۲].

مالاریای انسانی به‌وسیله‌ی یکی از چهارگونه پلاسمودیوم ویواکس، پلاسمودیوم فالسیپاروم، پلاسمودیوم اوال و پلاسمودیوم مالاریه ایجاد می‌شود و عامل بیماری با گزش چندین‌گونه از پشه آنوفل، به‌عنوان ناقل واسطه، به انسان منتقل می‌شود. بیش از ۳۰ گونه مختلف از پشه آنوفل انگل را انتقال می‌دهند که در ایران از ۲۵ گونه یافت شده، فقط ۸ گونه ناقل بیماری هستند[۳].

پلاسمودیوم قسمتی از چرخه‌ی زندگی خود را که مرحله جنسی نام دارد در بدن پشه و مرحله دیگر، که غیر جنسی است را در بدن انسان می‌گذراند. شیوع بیماری مالاریا به اکولوژی و بیولوژی پشه آنوفل بستگی دارد که مراحل لاروی خود را در آب سپری می‌کند. آب‌های ذخیره‌شده

در استخرها و حوض‌ها در مناطق شهری و آب‌های ذخیره شده برای کشاورزی از جمله زیستگاه‌های مناسب برای تکثیر این حشره در ایران هستند. بیماری مالاریا یکی از مهم‌ترین عوامل مرگ و میر بشری است [۴-۵]؛ در ایران، پلاسمودیوم ویواکس شایع‌ترین عامل مالاریا است[۱].

درجه حرارت، بارش و رطوبت نسبی از جمله عوامل اقلیمی تاثیرگذار بر توزیع جغرافیایی بیماری مالاریا به‌شمار می‌روند. این عوامل اقلیمی، هم در رشد و تکثیر پشه آنوفل و هم در فعالیت انگل پلاسمودیوم نقش تعیین‌کننده‌ای دارند. طول دوره اسپوروگونی پلاسمودیوم به درجه حرارت محیط و گونه‌ی آنوفل ناقل بستگی دارد[۳،۶]. حداقل حرارت برای فعالیت انگل پلاسمودیوم ۱۹-۱۴ درجه سلسیوس است؛ این آستانه برای گونه ویواکس کمتر از گونه فلسیپاروم است، در حالی‌که آستانه حداقل دمایی برای فعالیت پشه آنوفل ۸ تا ۱۰ درجه سلسیوس می‌باشد. پژوهش‌گران نشان دادند که آستانه‌ی دمای بهینه برای رشد و فعالیت پشه آنوفل ۲۵ تا ۲۷ درجه سلسیوس و برای انگل پلاسمودیوم ۲۱ تا ۲۷ درجه سلسیوس است [۳].

به‌دنبال تاثیرپذیری شدید بیماری مالاریا از عوامل محیطی و اقلیم شناختی، سازمان‌های جهانی از جمله سازمان جهانی بهداشت همواره به دنبال ابزارها و روش‌هایی برای شناسایی، پالایش، تحلیل و مدیریت عوامل محیطی و اقلیم‌شناختی موثر بر شیوع و گسترش این بیماری بوده‌اند[۳]. فرآیند کاوش و تحلیل حجم بزرگی از داده‌ها به کمک کامپیوتر به منظور استخراج دانش معنادار، داده‌کاوی یا کشف دانش نام دارد. هدف داده‌کاوی، پیش‌بینی ناشناخته‌ها یا مقادیر آینده متغیرها با استفاده از بعضی مقادیر شناخته‌شده به‌کمک روش‌های پیش‌بینی و شناسایی الگوهای توصیف داده به‌صورت قابل فهم با استفاده از روش‌های توصیفی است[۷]. پژوهش حاضر، علم داده‌کاوی را برای پالایش داده‌های موجود و شناسایی الگویی جهت پیش‌بینی نوع مالاریا بر اساس عوامل اقلیمی دما و بارش، مورد استفاده قرار داد. از آن‌جایی‌که مالاریای نوع ویواکس به دلیل داشتن هیپنوزوئیت می‌تواند عود کند، [۸] شناسایی درست نوع مالاریا و درمان کامل آن از اهمیت ویژه‌ای برخوردار است.

روش کار

مطالعه‌ی حاضر از نوع توصیفی- مقطعی است. جامعه‌ی پژوهش متشکل از ۲۸۵ نفر مراجعه‌کننده به مرکز بهداشت سراوان از مرداد ۸۸ تا آذر ۹۵ بود که با مراجعه مستقیم پژوهش‌گر و به‌صورت فایل اکسل تهیه گردید. محتوای داده‌ها مورد تایید متخصصان حوزه مربوطه قرار گرفت. به‌منظور رعایت ملاحظات اخلاقی، نوع مطالعه و هدف آن برای پرسنل مرکز بهداشت توضیح و به آن‌ها اطمینان داده شد که داده‌های گردآوری‌شده به‌صورت محرمانه محفوظ خواهد ماند. به‌منظور آماده‌سازی داده‌ها، رکوردهایی که دارای مقادیر از دست رفته بودند یا هیچ وجه تشابهی با سایر داده‌ها نداشتند، حذف گردیدند. تعداد رکوردهای مورد بررسی به ۲۸۰ رسید.

مدل‌سازی با استفاده از نرم‌افزار کلمنتاین ۱۲ انجام شد. روش کار در پژوهش حاضر، داده‌کاوی پیش‌بینانه بود. در الگوریتم‌های پیش‌بینی، هدف پیش‌بینی یک ویژگی خاص بر مبنای ویژگی‌های دیگر است. در ویژگی پیش‌بینی‌شونده، یک متغیر وابسته و بقیه متغیرها مستقل نامیده می‌شوند[۹]. الگوریتم‌های درخت تصمیم و شبکه عصبی برای مدل‌سازی استفاده شد. در تولید مدل، ابتدا داده‌های تحت بررسی به‌طور تصادفی به دو مجموعه مستقل آموزش و آزمایش تقسیم گردید. داده‌های بخش آموزش (۸۰ درصد)، مدل را تولید نمود و داده‌های بخش آزمایش (۲۰ درصد)، مدل تولید شده را آزمایش کرده، برچسب مربوط به رکوردهای مذکور را تعیین نمود. برای آموزش درخت تصمیم، یک متغیر رشته‌ای (طبقه‌ای/ اسمی) متغییر خروجی (نوع مالاریا) بود و یک یا تعداد بیشتری متغییر ورودی وجود داشت. پارامترهای سن، جنس، دما و میزان بارش به‌منظور پیش‌بینی نوع مالاریا مورد استفاده قرار گرفت. از آن‌جایی‌که سه عامل اقلیمی درجه حرارت ماهانه، بارش و نسبت اختلاف رطوبت هوا تعیین‌کننده‌ی طول دوره فعالیت، رشد و تکثیر پشه آنوفل و انگل پلاسمودیوم است، [۳] به‌منظور مدل‌سازی در مطالعه حاضر، متغییرهای دما و بارش مورد استفاده قرار گرفتند.

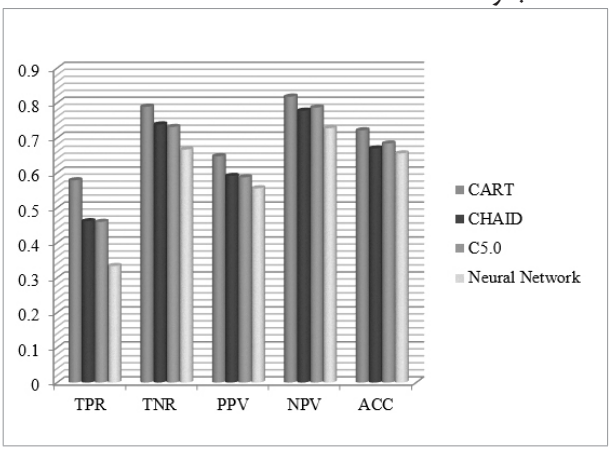
برای مدل‌سازی از الگوریتم‌های درخت تصمیم تقسیم و رگرسیون[1]، آشکارساز تعامل خودکار مجذور کای[2]، سی پنج [3] و شبکه عصبی[4] استفاده گردید. درخت تصمیم، آشکارساز تعامل خودکار و مجذور کای برای حل مسائل دسته‌بندی کاربرد دارند؛ در حالی‌که درخت‌های تقسیم و رگرسیون و سی ۴/۵ [5] در حل هر دو نوع مسائل رگرسیون و دسته‌بندی مورد استفاده قرار می‌گیرند[۱۰]. سی پنج نسخه جدید سی ۴/۵ است که نسبت به سی ۴/۵ هنگام تولید مجموعه قوانین، حافظه‌ی کمتری را اشغال می‌کند[۱۱]. گزینه آزمایشی انتخاب شده برای الگوریتم‌های درخت تصمیم، اعتبارسنجی متقاطع با ده تکرار[6] بود؛ زیرا آزمایش‌ها نشان داده‌اند که بهترین انتخاب برای به‌دست آوردن دقیق‌ترین تخمین، اعتبارسنجی متقاطع ده است[۹]. درصد نویز[7] قابل‌انتظار

[1] Classification and Regression Tree: C&R Tree or CART
[2] Chi-Squared Automatic Interaction Detector: CHAID
[3] C5.0
[4] Neural Network
[5] C4. 5
[6] 10-Fold Cross-Validation
[7] Noise

برای این الگوریتم‌ها، صفر و تعداد سطوح زیر ریشه، پنج انتخاب گردید. شبکه عصبی مصنوعی نیز از یک لایه مخفی با ۲۰ نرون[8]، حالت آموزش دسته‌ای و الگوریتم‌های سریع با پایداری ۲۰۰ بهره می‌برد. برای ارزیابی مدل‌ها و انتخاب بهترین مدل، از شاخص‌های حساسیت[9]، شفافیت[10]، دقت یا ارزش اخباری مثبت[11]، ارزش اخباری منفی[12] و صحت[13] استفاده گردید[۱۲].

یافته‌ها

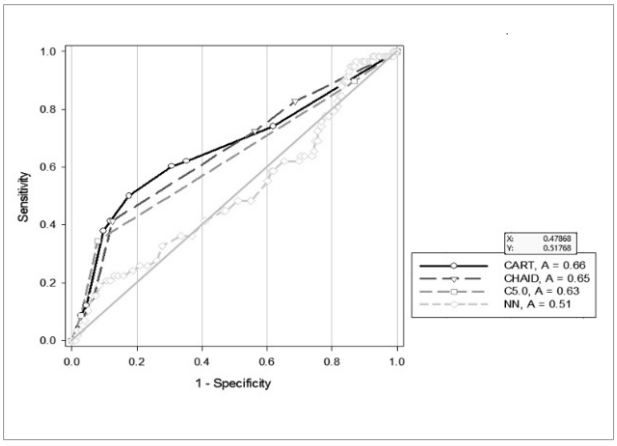
یافته‌ها نشان داد که میانگین سن افراد ۲۷/۰۵ سال بود. همچنین، ۸۱/۰۸ درصد مبتلایان مرد و ۱۸/۹۲ درصد زن بودند. به‌علاوه، ۲۰/۷۱ درصد بیماران مبتلا به نوع فالسیپاروم، ۱۶/۴۳ درصد مبتلا به نوع میکس و ۶۲/۸۶ درصد مبتلا به نوع ویواکس بودند. در همین رابطه، ۷۴/۲۸ درصد مبتلایان ایرانی، ۱۶/۴۲ درصد پاکستانی، ۸/۹۲ درصد افغانی و ۰/۳۸ درصد ترک بودند. همچنین، ۳۱/۷۸ درصد افراد دارای شغل آزاد، ۲۵/۷۱ درصد کارگر، ۱۲/۸۵ درصد خانه‌دار، ۸/۹۲ درصد کودک و ۲۰/۷۴ درصد دارای سایر مشاغل بودند. مقادیر شاخص‌های ارائه شده در نمودار ۱ (ترسیم شده در نرم افزار اکسل) روی مجموعه داده‌های آزمایش نشان داد که درخت تقسیم و رگرسیون بهترین مدل را تولید نموده است. شاخص‌های حساسیت، شفافیت، صحت، دقت و ارزش اخباری منفی برای این مدل دارای بیشترین مقدار بود. مقادیر این شاخص‌ها هرچه بیشتر باشد نشان‌دهنده این است که طبقه‌بند مورد استفاده، نمونه‌های بیشتری را در جای درست خود قرارداده است. از میان الگوریتم‌های مورد استفاده، بهترین نتایج مربوط به الگوریتم تقسیم و رگرسیون با دقت ۰/۶۴۷۶، صحت ۰/۷۲۱۷، شفافیت ۰/۷۸۹۳ و حساسیت ۰/۵۷۸۷ بود.



نمودار ۱: مقادیرشاخص‌ها برای مدل‌های تولید شده

[8] Neuron
[9] True Positive Rate or Sensitivity: TPR
[10] True Negative Rate or Specificity: TNR
[11] Positive Predictive Value or Precision: PPV
[12] Negative Predictive Value: NPV
[13] Accuracy: ACC

در تحلیل داده‌های پزشکی، استفاده از معیارهای شفافیت و حساسیت برای ارزیابی مدل پیش‌بینی کننده مناسب است[۱۳]. همچنین جهت ارزیابی کارایی دسته بندها، از روش موثر و شناخته شده‌ی منحنی مشخصه عملکرد سیستم[14] که نتایج آن متغیری در مقیاس رتبه‌ای و یا کمی است، استفاده شد[۱۴]. این معیار برای مقایسه‌ی گرافیکی مدل‌های دسته‌بندی استفاده می‌شود و نتیجه حاصل از آن، منحنی‌هایی است که به کمک آن‌ها به‌راحتی می‌توان مدل مناسب را تشخیص داد. منحنی ایجاد شده، مبادله‌ی میان حساسیت (محورy) و شفافیت-۱ (محور x) است[۱۵]. بیشینه شدن حساسیت به برخی مقادیر بزرگ y و بیشینه شدن شفافیت به یک مقدار کوچک x روی منحنی مربوط است. یکی از معیارهای مهم صحت در تست‌های کلینیکی، سطح زیر منحنی ROC است. اگر سطح زیر منحنی برابر ۱ شود، آزمایش انجام شده ۱۰۰درصد صحیح است؛ زیرا هر دو مقدار حساسیت و شفافیت ۱ هستند و داده‌های برچسب منفی که به نادرست مثبت و داده‌های برچسب مثبتی که به نادرست منفی دسته‌بندی شده‌اند، وجود ندارند[۱۶]. نمودار ۲، ترسیم شده در نرم‌افزار سیگما پلات، منحنی ROC را برای مدل‌های تولید شده نشان داد و برتری الگوریتم تقسیم و رگرسیون را به لحاظ بیشتر بودن سطح زیر منحنی تایید نمود.



نمودار ۲: منحنی ROC برای مدل‌های تولید شده

پارامترهای دما، بارش و جنس به ترتیب با مقادیر ۰/۴۴، ۰/۳۶۱ و ۰/۱۱۴ به‌عنوان تاثیرگذارترین پارامترها در مدل‌سازی توسط درخت تصمیم تقسیم و رگرسیون انتخاب گردیدند. قوانین تولیدشده نیز پارامترهای تاثیرگذار روی نوع مالاریا را شناسایی و تایید نمود. قوانین ایجادشده برای یک نمونه جدید با ویژگی‌های مشخص، می‌توانند نوع مالاریا را پیش‌بینی نمایند. جدول ۱ قوانین استخراج شده از درخت تصمیم تقسیم و رگرسیون را نمایش می‌دهد.

[14] Receiver Operating Characteristic: ROC

جدول ۱: نمونه‌ای از قواعد استخراج شده از درخت تصمیم تقسیم و رگرسیون

ردیف	قانون	صحت قانون
۱	اگر دما>۱۷/۹۵ و دما ≥۱۹/۶ باشد، آنگاه نوع مالاریا فالسیپارم است.	۰/۶۰۰
۲	اگر دما>۱۹/۶ و دما ≥۲۹/۴ و بارش >۰/۲۵ و بارش ≥۲۲/۳ و سن >۳۴/۵ باشد، آنگاه نوع مالاریا میکس (عفونت توام ویواکس و فالسیپاروم) است.	۰/۶۰۰
۳	اگر دما>۲۹/۴ و دما ≥۳۰/۹۵ و بارش ≥۲۲/۳ و سن >۲۰/۵ باشد، آنگاه نوع مالاریا میکس است.	۰/۶۶۷
۴	اگر دما>۱۹/۶ و بارش>۲۲/۳ باشد، آنگاه نوع مالاریا میکس است.	۰/۴۵۵
۵	اگر دما ≥۱۷/۹۵ باشد، آنگاه نوع مالاریا ویواکس است.	۰/۷۵۰
۶	اگر دما>۱۹/۶ و دما ≥۲۹/۴ و بارش ≥۲۲/۳ و سن ≥۳۴/۵ باشد، آنگاه نوع مالاریا ویواکس است.	۰/۷۸۳
۷	اگر دما>۱۹/۶ و دما ≥۲۹/۴ و بارش ≥۰/۲۵ و سن <۳۴/۵ باشد، آنگاه نوع مالاریا ویواکس است.	۰/۶۹۲
۸	اگر دما>۲۹/۴ و دما ≥۳۰/۹۵ و بارش ≥۲۲/۳ و سن ≥۲۰/۵ باشد، آنگاه نوع مالاریا ویواکس است.	۰/۸۷۵
۹	اگر دما>۳۰/۹۵ و بارش ≥۲۲/۳ باشد، آنگاه نوع مالاریا ویواکس است.	۰/۷۵۶

بحث

اهمیت مبارزه و کنترل بیماری مالاریا سبب شد تا سازمان ملل متحد کاهش ۵۰ درصدی موارد بیماری را جزء اهداف خود تا سال ۲۰۱۵ قرار دهد[۱۷]. به‌دلیل تاثیرپذیری شدید بیماری مالاریا از عوامل محیطی و اقلیم شناختی، سازمان جهانی بهداشت به دنبال ابزارها و روش‌هایی برای شناسایی، پالایش، تحلیل و مدیریت عوامل محیطی و اقلیم‌شناختی موثر بر شیوع و گسترش این بیماری است[۳]؛ از این‌رو، پژوهش حاضر از علم داده‌کاوی برای مدل‌سازی داده‌های موجود جهت پیش‌بینی تاثیر دما و بارش بر نوع مالاریا استفاده نمود.

یافته‌ها نشان داد ۸۱/۰۸ درصد مبتلایان مرد و ۶۲/۸۶ درصد بیماران مبتلا به نوع ویواکس بودند. طولابی و همکاران در پژوهش خود دریافتند که ۹۱/۰۸ درصد مبتلایان به مالاریا در شهر بم به پلاسمودیوم ویواکس مبتلا بوده و نیز ۶۶/۱ درصد مبتلایان مرد بودند. پژوهش‌های دیگری که در آفریقا، بحرین، شرق مدیترانه و استان‌های همدان، چهارمحال و بختیاری و مازندران انجام شد، درصد بالاتر ابتلا به نوع ویواکس را تایید نمود. همچنین، پژوهش‌های انجام شده در استان‌های همدان، اردبیل و چهارمحال و بختیاری، درصد بالاتر ابتلای مردان به مالاریا را تایید می‌نماید[۸].

یافته‌های پژوهش نشان داد از میان الگوریتم‌های مورد استفاده، بهترین نتایج از الگوریتم تقسیم و رگرسیون با صحت ۰/۷۲۱۷، دقت ۰/۶۴۷۶، شفافیت ۰/۷۸۹۳، حساسیت ۰/۵۷۸۷ و ارزش اخباری منفی ۰/۸۱۸۱ حاصل گردید. مقدار ۰/۶۶ برای سطح زیر منحنی ROC در نمودار ۲، برتری مدل تقسیم و رگرسیون را بر سایر مدل‌ها تایید نمود. قواعد استخراج‌شده از این الگوریتم نشان داد که دمای پایین‌تر، شرایط مساعدتری برای شیوع مالاریای نوع ویواکس فراهم می‌نماید؛ به‌طوری‌که حلیمی و همکاران حداقل درجه حرارت برای فعالیت انگل پلاسمودیوم را ۱۹-۱۴ درجه سلسیوس بیان نموده و این آستانه را برای گونه ویواکس کمتر از گونه فلسیپاروم دانستند. ایشان نشان دادند که مهم‌ترین عامل در تبیین بروز سالیانه بیماری مالاریا در استان سیستان و بلوچستان بارش است که در دوره فعالیت پشه آنوفل نقش مهمی دارد و درجه حرارت و رطوبت نسبی، اولویت‌های دوم و سوم بروز سالیانه‌ی این بیماری است[۳].

احمدی از روش‌های شبکه عصبی و سیستم استنتاج فازی عصبی تطبیقی برای استخراج الگوی همبستگی مکانی و زمانی داده‌ها و گستره مکانی و زمانی بیماری مالاریا استفاده نمود. وی به داده‌کاوی مکانی بیماری پرداخت و بدین منظور از میانگین‌گیری نرخ وقوع در یک دوره چهارساله، میانگین دما، رطوبت و بارش چهارساله در کنار سایر عوارض مکانی در استان سیستان و بلوچستان استفاده نمود. سپس با داشتن نرخ ابتلای ماهیانه بیماری در استان هرمزگان، به بررسی زمانی بیماری و داده‌کاوی زمانی مالاریا پرداخت[۱۸]. ولی پور و همکاران با استفاده از سامانه‌ی اطلاعات مکانی و فرآیند تحلیل سلسله مراتبی، شیوع بیماری مالاریا را مدل‌سازی نمودند. ایشان از دو شاخص تعداد موارد مثبت بیماری در یک سال بر جمعیت متوسط همان سال[15] و شاخص تراکم بیماری (تعداد موارد مثبت بیماری بر مساحت منطقه) برای ارزیابی نتایج استفاده نموده و نشان دادند که در مناطق با پتانسیل بالاتر برای شیوع بیماری، مقدار شاخص API بیشتر است[۱۹]. مطالعات انجام شده تا کنون تمرکز بیشتری بر بروز و شیوع بیماری مالاریا داشته‌اند و تاثیر عوامل اقلیمی بر نوع بیماری کمتر مورد توجه قرار گرفته است. از محدودیت‌های مطالعه حاضر می‌توان به عدم وجود پارامتر اقلیمی نسبت اختلاف رطوبت هوا در مجموعه داده‌ی تحت بررسی جهت کشف روابط پنهانی این پارامتر با نوع بیماری مالاریا اشاره نمود.

[15] API

نتیجه‌گیری

با توجه به نتایج پژوهش، وجود رویکردی سازمان‌دهی شده در مراکز درمانی جهت پیش‌بینی نوع مالاریا به‌منظور کمک به کارکنان بخش جهت افزایش صحت تشخیص نوع بیماری و ارائه روش درمان مناسب، امری ضروری است. نقطه قابل بهبود در این حیطه، پیش‌بینی دقیق درمان مناسب به‌جهت حفظ رفاه حال بیماران و کاهش مرگ و میر آن‌ها و ایجاد سیستم‌های تصمیم‌یار جهت کمک به تشخیص اولیه نوع مالاریا می‌باشد؛ زیرا مالاریای نوع ویواکس به دلیل داشتن هیپنوزوئیت می‌تواند عود کند. به‌همین علت، شناسایی و درمان مبتلایان این نوع مالاریا بسیار ضروری است. از آن‌جایی که نتایج این تحقیق وابسته به داده‌های مرکز بهداشت سراوان است، پیشنهاد می‌شود برای بررسی بیشتر تاثیر عوامل اقلیمی در نوع مالاریا، در مطالعات بعدی از داده‌های مراکز بهداشت سایر شهرهای درگیر با بیماری مالاریا استفاده گردد.

کاربرد در تصمیم‌های مرتبط با سیاست‌گذاری در نظام سلامت

۱-پیش از انجام پژوهش، زیان‌های مالی و جانی بیماری مالاریا در منطقه مرزی سراوان همچنان قابل‌مشاهده بود و لازم است با توجه به نوع مالاریا، درمان مناسبی انتخاب گردد؛ زیرا به عنوان مثال، مالاریای نوع ویواکس به دلیل داشتن هیپنوزوئیت می‌تواند عود کند[۸].

۲-پژوهش حاضر با بررسی متغیرهای اقلیمی می‌تواند نوع بیماری مالاریا را به سرعت تشخیص دهد تا درمان مناسب انتخاب گردد. به‌عنوان مثال، قواعد جدول ۱ نشان داد که دمای پایین‌تر، شرایط مساعدتری برای شیوع مالاریای نوع ویواکس فراهم می‌نماید؛ به‌طوری‌که حلیمی و همکاران حداقل درجه حرارت برای فعالیت انگل پلاسمودیوم را ۱۹-۱۴ درجه سلسیوس بیان نموده و این آستانه را برای گونه ویواکس کمتر از گونه فلسیپاروم دانستند[۳]. قانون شماره ۵ در مقابل قانون شماره ۱ موید این مطلب است.

۳-با توجه به شرایط اقلیمی مناطق درگیر با بیماری مالاریا، سیاست‌گذاران نظام سلامت باید با اعمال راهبردهای مدیریتی نظیر سمپاشی، قرنطینه نمودن بیماران و...، شرایط مناسب جهت جلوگیری از رشد و شیوع پشه آنوفل را فراهم نمایند.

تشکر و قدردانی

از کارکنان محترم مرکز بهداشت سراوان جهت همکاری در تهیه داده قدردانی می‌گردد.

References

1- Karimazar M, Khazan H, Azar Gashb E. Comparing the prevalence of malaria in thalassemic individuals and non-thalassemic ones in Iranshahr. J Ardabil Univ Med Sci. 2004; 4(3): 56-61. (In Persian)

2- Azadbakht M, Azadbakht M. Five prevalent antiprotozoal herbal drugs. J Mazandaran Univ Med Sci. 2008; 18(67): 118-32. (In Persian)

3- Halimi M, Delavari M, Takhtardeshir A. Survey of climatic condition of Malaria disease outbreak in Iran using GIS. Journal of School of Public Health and Institue of Public Health Research. 2013; 10(3): 41-52. (In Persian)

4- Blencowe H, Cousens S, Jassir F, Chou D, Mathers C, Hogan D, et al. National, regional, and worldwide estimates of stillbirth rates in 2015, with trends from 2000: a systematic analysis. The Lancet Global Health. 2016; 4(2): 98-108.

5- Hay S, Okiro E, Gething P, Patil A, Tatem A, Guerra C, et al. Estimating the global clinical burden of Plasmodium falciparum malaria in 2007. PLoS Med. 2010; 7(6): pp. 1-22.

6- Salehi M, Salehifard O, Soleimani-Ahmadi M. Environmental factors effective on malaria prevalence in Rudan county during 2003 to 2011. Journal of Preventive Medicine. 2015; 1(2): 13-21. (In Persian)

7- Ashoori M, Mohammadi S, Hossieny Eivary H. Exploring blood donors' status through clustering: a method to improve the quality of services in blood transfusion centers. Journal of Knowledge & Health. 2016; 11(4): 73-82. (In Persian)

8- Toolabi A, Kermanizadeh A, Nikonahad A, Miri N, Khabiri F. Spatial analysis of malaria disease reports using Geographic Information System (GIS) in Bam, 2004-2014. J Rafsanjan Univ Med Sci. 2016; 15(3): 331-42. (In Persian)

9- Ashoori M. A model to predict hemodialysis buffer type using data mining techniques. Journal of Health Administration. 2017; 20(67): 99-110. (In Persian)

10- Pham H. Springer Handbook of Engineering Statistics. Germany: Springer; 2006: 651-6.

11- Pandya R, Pandya J. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications. 2015; 117(16): 18-21.

12- Han J, Kamber M. Data Mining: Concepts and Techniques. 2nd ed. USA: Elsevier Inc; 2006: 383-464.

13- Yeh J, Wu T, Tsao C. Using data mining techniques to predict hospitalization of hemodialysis patients. Decision Support Systems. 2011; 50(2): 439-448.

14- Sadat-Hashemi S, Ghorbani R, Kavehie B. Analyzing receiver operating characteristic curves to compare medical diagnostic tests. Koomesh. 2005; 6(2): 145-150. (In Persain)

15- Lowry R. Simple ROC curve znalysis vassarwtats: website for statistical computation [Internet]. 2001-2019. Avilable from: http://vassarstats.net. Accessed February 7, 2019.

16- SigmaPlot 12.5 User's Guide. United States: Systat Software; 2013: 371-373.

17- Soofi K, Khanjani N, Kamyabi F. Study of malaria in-

fection trend and the role of preventive interventions on malaria incidence in Sarbaz city, Sistan and Baluchestan province. Journal of Preventive Medicine. 1394; 2(3): 56-66. (In Persian)
18- Ahmadi A. Malaria disease located-time data mining [MS.c thesis]. Tehran: K.N.Toosi University of Technology; 1392: 1. (In Persian)
19- Valipoor A.A, Al Sheykh A.A, Garegozloo A, Khierkhah M.M. Malaria incidence modeling using Geographic Information System (GIS) and Analytical Hierarchy process(AHP): case study Hormozgan province. Proceedings of the Geomatics; 2011 May 15-19. Tehran. Iran. (In Persian)

Predicting the Type of Malaria Using Classification and Regression Decision Trees

Maryam Ashoori[1 *], Fatemeh Hamzavi[2]

[1]*School of Technical and Engineering, Higher Educational Complex of Saravan, Saravan, Iran*
[2]*School of Agriculture, Higher Educational Complex of Saravan, Saravan, Iran*

Abstract

Background: Malaria is an infectious disease infecting 200 - 300 million people annually. Environmental factors such as precipitation, temperature, and humidity can affect its geographical distribution and prevalence. The environmental factors are also effective in the abundance and activity of malaria vectors. The present study aimed at presenting a model to predict the type of malaria.

Methods: This cross-sectional study was conducted using the data of 285 people referring to a health center in Saravan from June 2009 to December 2016. Clementine 12.0 was used for data analysis. The modeling was done using classification and regression decision trees, chi-squared automatic interaction detector, C 5.0, and neural network algorithms.

Results: The accuracy of classification and regression decision trees, chi-squared automatic interaction detector, C5.0, and neural network was 0.7217, 0.6698, 0.6840, and 0.6557, respectively. Classification and regression decision trees performed better than the other algorithms in terms of sensitivity, specificity, accuracy, precision, negative predictive value, and area under the ROC curve. The sensitivity and area under the ROC curve were 0.5787 and 0.66 for classification and regression decision trees.

Conclusions: Applying data mining methods for the analysis of malaria's data can change the current attitude toward malaria type determination. Faster and more precise identification of malaria type helps determine the proper cure and improve the performance of health organizations.

Keywords: Malaria; Decision Tree; Neural Network; ROC Curve

Please cite this article as follows:

Ashoori M, Hamzavi F. Predicting Malaria Type Using Classification & Regression Decision Tree Algorithm. Hakim Health Sys Res 2019; 22(1): 75- 81.

***Corresponding Author:** M.Sc. of Information Technology Engineering, School of Technical and Engineering, Higher Educational Complex of Saravan, Pasdaran Blvd., Saravan, Iran. Tel: +98-9155338721, Fax: +98-5437630090, Email: mashoori@saravan.ac.ir