تجزیه و تحلیل و پیش بینی بیماری پارکینسون با استفاده از تکنیک های داده کاوی

سپیده قاسمی زرکامی[1] ، سجاد صفیر*[2]

۱-کارشناسی ارشد مهندسی نرم افزار ، دانشگاه علوم تحقیقات گیلان،ایران،

ghasemi.se@centinsur.ir

۲-کارشناسی ارشد مهندسی فناوری اطلاعات ، دانشگاه صنعتی شریف،ایران،

s.safir@centinsur.ir

چکیده

بیماری پارکینسون بیماری است که تشخیص آن به صورت پزشکی بسیار مشکل و هزینه بوده و هر روز محققان در پی این هستند تا یک راه حل برای تشخیص زودهنگام این بیماری بیابند. از آنجا که اغلب این بیماری را توسط نشانه های صوتی بیماران مبتلا به PD[1] مانند کاهش بلندی و وضوح صدا، اختلال در کیفیت صدا شناسایی می کنند، این روش کاربرد زیادی در تشخیص دقیق این بیماری دارد. تحقیقات قبلی بر روی بیماران نشان داده که ۹۰ ٪ از بیماران مبتلا به بیماری پاکینسون یک اختلال صوتی در آنها مشاهده شده است. بنابراین اندازه گیری این علایم صوتی و شناسایی آنها در تشخیص بیماری نقش مهمی ایفا می کند. به این جهت از داده های صوتی در انجام این تحقیق استفاده شد. به دلیل تعداد زیاد بیماران و آزمایش های متعدد هر بیمار،س نیاز به یک ابزار خودکار برای کاوش در میان بیماران پارکینسون احساس می شود. از طرفی از آنجا که تکنیک های داده کاوی برای پیش بینی بیماری های مختلف در زمینه پزشکی نقش مهمی ایفا می کند، در این تحقیق از تکنیک های داده کاوی استفاده شده است. در این تحقیق روش های داده کاوی برای تشخیص بیماری پارکینسون بررسی شدند و در انتها نتایج شبیه‌سازی و تحلیل داده ها در نرم افزار weka ارائه شد. نتایج این تحقیق نشان داد که میزان صحت طبقه بندی شبکه عصبیMLP با ۹۲,۳۰ ٪ بیشترین مقدار صحت را بدست آورده است. پس از آن NaiveBayes و SVM-Smo ۹۱,۲۸ ٪ صحت و J48 با ۸۹,۷۴ ٪ صحت و AdaBoostM1 صحت ۸۸,۲۰ ٪ را بدست آوردند. در انتها DecisionStump با ۸۴,۶۱ ٪ کمترین مقدار صحت را بدست آورد.

کلمات کلیدی: بیماری پارکینسون، داده کاوی، شبکه عصبیMLP ،Weka .

1-مقدمه

پارکینسون یک بیماری عصبی پیش رونده است، که با نشانه های حرکتی و غیر حرکتی مشخص می شود[۱]. بعداز آلزایمر، پارکینسون شایع‌ترین بیماری تحلیل رونده مغز به حساب می‌آید. این بیماری یک بیماری مزمن و تدریجا پیش رونده است که در آن سلول های ترشح‌کننده دوپامین در جسم سیاه مغز کاهش می یابند. علت زمینه ساز برای حدود ۹۵ درصد از افراد که این بیماری در آنها تشخیص داده می شود، ناشناخته باقی می ماند. در دهه ۱۹۶۰ کشف شد که علائم در درجه اول به دلیل عدم وجود یک انتقال دهنده عصبی (دوپامین) به وجود می آیند و در اثر کمبود دوپامین حرکات بدن مختل می شود. ماده دوپامین یک ماده شیمیای است که به وسیله بدن برای کنترل حرکت استفاده می شود. وجود دوپامین کم در سیستم گردش خون باعث می شود فرد سختی در حرکت، لرزش و کرختی در اعضا بدن را احساس کند[۲-۳].

2-داده کاوی

داده کاوی فرآیندی است که به استخراج دانش نهفته در داده ها می پردازد و الگوها و روابط میان داده های موجود در دیتاست را نمایان می کند[7]. به عبارت دیگر داده کاوی یکی از گام های کشف دانش است که الگو های پنهان موجود در داده های وسیع، ناقص و دارای نویز را استخراج می کند[8]. داده کاوی برای نیل به اهداف خود ترکیبی از تحلیل آماری، یادگیری ماشین

[1] Parkinson's Disease

مصنوعی و پایگاه داده را به کار می برد[9]. دسته بندی و خوشه بندی به عنوان عملیات اصلی داده کاوی به شمار می روند. دسته بندی جز یادگیری نظارت شده است که ویژگی های مشترک موجود در میان مجموعه ای از عناصر موجود در دیتاست را یافته و آنها را در دسته های مختلف قرار می دهد[10]. خوشه بندی یک روش یادگیری بدون نظارت است که همانند دسته بندی عناصر مشابه را در یک دسته قرار می دهد اما در دسته بندی دسته ها مشخص و گسسته هستند ولی در خوشه بندی بر چسب دسته ها شناخته می باشد[11]. امروزه داده کاوی پزشکی اغلب در حوزه مسایل دسته بندی قرار می گیرند و در جستجوی بهترین روش برای دسته بندی افراد در دسته های بیمار و سالم هستند. امروزه پژوهشگران در تشخیص و پیش بینی بیماری های مختلفی چون دیابت، سکته، سرطان و بیماری های قلبی، از تکنیک های داده کاوی استفاده می کنند.

۳–روش کار

روش کار کلی بدین صورت است که ابتدا مجموعه داده تبدیل به فایل ARFF می شود. پس از بارگذاری داده ها در Weka دوفیلتر انتخاب ویژگی و Discretize برروی داده های موجود در مجموعه داده های بیماری پارکینسون برای کاهش ویژگی ها اعمال می گردد. سپس شش طبقه بند مورد نظر بر روی ویژگی داده ها باقیمانده پس از مرحله فیلتر کردن اعمال می گردد. درنهایت نتایج طبقه بند ها با یکدیگر مقایسه می شوند. دیاگرام روش کلی کار را در شکل ۱مشاهده می شود.

شکل ۱ : روش کلی کار

۵– مجموعه داده مورد استفاده

داده استفاده شده در این مطالعه از صداهای ضبط شده در دانشگاه آکسفورد به وسیله Max Little است. در این پایگاه، داده ها به وسیله ابزارهای خاص ضبط شده اند. داده ها شامل ۱۹۵ صدای ضبط شده از۳۱ انسان است که سن این افراد در محدوده ۴۶ تا ۸۵ سال بود و ۶ صدای صوتی از هر فرد ضبط شده است. مشخص شد که ۲۳ نفر مبتلا به PD هستند. مجموعه داده ها شامل ۲۳ خصوصیت است. که شامل میانگین فرکانس صوتی ، حداکثر فرکانس صوتی ، حداقل فرکانس صوتی ، تنوع در فرکانس ، اندازه تغییر در دامنه ، نرخ نویز مربوطه به آهنگ صدا ، اندازه گیری غیر خطی فرکانس و ... می باشند

۶– پیش پردازش

استفاده عملی از معیارهای محاسبه شده نیازمند بردار ویژگی های ساخته شده از این معیارها است، که می تواند به صورت متوالی برای کاهش افراد سالم از بیماران استفاده شود. بیشتر انواع طبقه بندها توسط پیش پردازش مقادیر هر معیار با مقیاس شدن مناسب بهتر عمل می کنند [20].در پیش پردازش عمل خواندن و فیلتر کردن داده ها انجام می شود ، در ابتدا باید فایل مجموعه داده به شکل فایل ARFF خواندن در مرحله Preprocess تبدیل شود.

بعد از اینکه فایل مورد نظر بارگذاری شد، weka فیلدها را تشخیص می دهد و در حین بررسی آنها، اطلاعات آماری پایه ای را برای هر فیلد محاسبه می کند. در قسمت پیش پردازش یک سری فیلتر را می توان برای کاهش حجم داده ها و

بهینه کردن کار، بر داده های مورد استفاده اعمال کرد. اولین فیلتری که بر روی داده ها برای کاهش ویژگی ها اعمال می شود فیلتر انتخاب ویژگی است، که موجب میشود تعداد ویژگی ها از ۲۳ ویژگی به ۱۱ویژگی کاهش یافتند . در ادامه فیلتر discretize که با استفاده از آن می توان مقادیر یک صفت پیوسته را به تعداد دلخواه بازه گسسته تبدیل کرد. پس از اعمال فیلتر انتخاب ویژگی، فیلتر discretize برروی داده ها اعمال می گردد.

۷– تکنیک های داده کاوی

۷-۱- روش درخت تصمیم J48:

درخت تصمیم یکی از ابزارهای کار آمد دسته بندی و پیش بینی است. ایجاد درخت سریع بوده و تفسیر قواعد شرطی تولید شده از آن آسان است. از این رو درخت تصمیم از تکنیک های پرکاربرد به شمار می رود[13]. این روش در زمینه داده کاوی پزشکی در تشخیص و پیش بینی بیماریهای مختلف توسط پژوهشگران مورد استفاده قرار گرفته است. درخت تصمیم یک استراتژی بالا به پایین به ایجاد آزمون بر روی هر گروه با مقدار پیوسته و هم از ویژگی های با مقدار گسسته پشتیبانی می کند.

2-7-روش ماشین بردار پشتیبان SMO[1] :

این الگوریتم از روش های یادگیری باناظر است که از آن برای طبقه بندی و رگرسیون استفاده می کنند. این روش از جمله روش های نسبتاً جدیدی است که در سال های اخیر کارایی خوبی نسبت به روش های قدیمی تر برای طبقه بندی از جمله شبکه های عصبی پرسپترون نشان داده است. مبنای کاری دسته بندی کننده ی SVM دسته بندی خطی داده هاست و در تقسیم خطی داده ها سعی بر آن است خطی انتخاب شود که حاشیه اطمینان بیشتری داشته باشد. حل معادله ی پیدا کردن خط بهینه برای داده ها به وسیله روش های QP که روش های شناخته شده ای در حل مسائل محدودیت دار هستند صورت می گیرد. قبل از تقسیم خطی برای اینکه ماشین بتواند داده های با پیچیدگی بالا را دسته بندی کند داده ها را باید به وسیله توابع phi به فضای با ابعاد خیلی بالاتر برد. ماشین بردار پشتیبانSMO ، از الگوریتم یادگیری ماشین است که عملیات تشخیص الگو را بر اساس یادگیری آماری و قاعده حداقل سازی ریسک ساختاری، انجام می دهد[14].

۷-۳- روش دسته بندی بیز ساده :

این روش برای یافتن بهترین دسته بندی ممکن از قاعده بیز ساده استفاده می کند [15]. این قاعده برای ایجاد مدل های با قابلیت پیش بینی مورد استفاده قرار می گیرد و راهکارهایی جهت کاوش و درک داده ها فراهم می کند و با محاسبه ی متغییر هدف و سایر متغیرها از روی شواهد موجود عمل یادگیری را انجام می دهد. براساس قاعده بیز، احتمال اینکه داده X_t متعلق به دسته یC باشد به صورت رابطه ۱ است. دسته بندی بیز احتمال شرطی هر نمونه ی متعلق به هر دسته را محاسبه می کند و براساس آن، نمونه ی مورد بررسی در دسته ای که بالاترین احتمال شرطی را داشته باشد قرار می گیرد.

(۱)

$$P(C|X_t) = \frac{P(C)P(X_t|C)}{P(X_t)}$$

۷-٤-درخت تصمیم گیری :

درخت تصمیم درختی است که در آن نمونه ها را به نحوی دسته بندی می کند که از ریشه به سمت پائین رشد می کنند و در نهایت به گره های برگ می رسد. هر گره داخلی یا غیر برگ با یک ویژگی مشخص می شود، این ویژگی سوالی را در

[1] Sequential Minimal Optimization

رابطه با مثال ورودی مطرح می کند. در هر گره داخلی به تعداد جواب های ممکن با این سوال شاخه وجود دارد که هر یک با مقدار آن جواب مشخص می شود. برگ های این درخت با یک کلاس و یا یک دسته از جواب ها مشخص می شوند[17].

۷-۵-شبکه های عصبی مصنوعی :

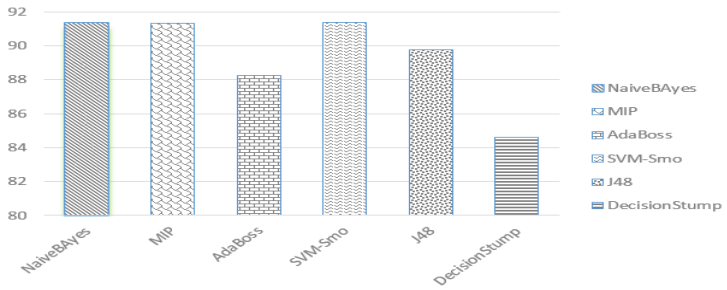
شبکه های عصبی مصنوعی که معمولا به عنوان شبکه های عصبی نام برده می شوند یک الگوی ریاضی مبتنی برسیستم زیستی است. سیستم های عصبی یک الگوریتم برای بهینه سازی و یادگیری آزادانه بر اساس مفاهیم الهام گرفته از تحقیق در ماهیت مغز می باشند. مغز با استفاده از قابلیتی شناخته شده به عنوان نورون اجزا ساختاری خود را سازماندهی می کند، در نتیجه محاسبات معینی را بسیار سریع تر از کامپیوتر دیجیتال انجام می دهد.

۷-٦- طبقه بندAdaBoostm1 :

طبقه بندی AdaBoost1 یک طبقه بند باینری، دو کلاسه، دو بخش است. این طبقه بند برای ارتقاء آموزش دهنده Weka طراحی شده است. طبقه بندAdaBoostm1 یک طبقه بند m کلاس است اما هنوز نیاز به آموزش دهنده Weka برای بدست آوردن صحت بهتر از $\frac{1}{2}$ دارد. خطای آموزشی بدست آمده از فرضیه نهایی تولید شده توسط طبقه بندAdaBoostm1 بسیار کم است اما عیب اصلی آن ناتوانی در کارکردن با فرضیه Weka در خطای بیشتر از $\frac{1}{2}$ است .

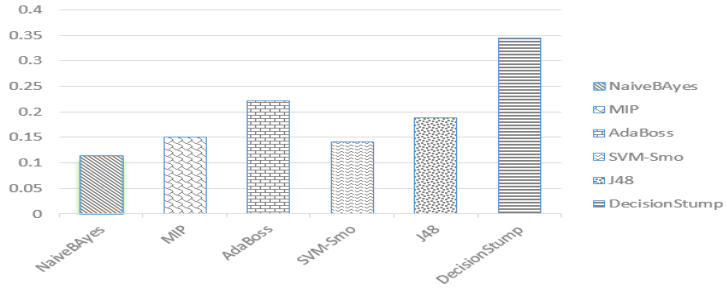
۸ - مقایسه شش طبقه بند موردنظر

Correctly Classified Instances: این پارامتر درصد صحت طبقه بندی نمونه ها را بیان می کند. مقایسه این پارامتر برای شش طبقه بند در شکل ۲ مشاهده می شود.



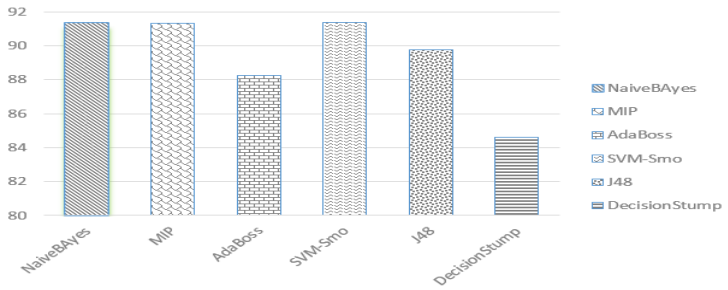
شکل ۲ – مقایسه پارامتر طبقه بندی درست نمونه ها

Incorrectly Classified Instances: این پارامتر درصد عدم صحت طبقه بندی نمونه ها را بیان میکند. مقایسه این پارمتر برای شش طبقه بند در شکل ۳ مشاهده می شود.



شکل ۳ – مقایسه پارامتر طبقه بندی غلط نمونه ها

Kappa statistic: این پارامتر به صورت کلی معیار قرارداد است به طوری که در مورد تغییر قرارداد نرمال سازی می شود. این پارامتر برای دسترسی به کیفیت مورد اطمینان در کارایی بکار می رود.



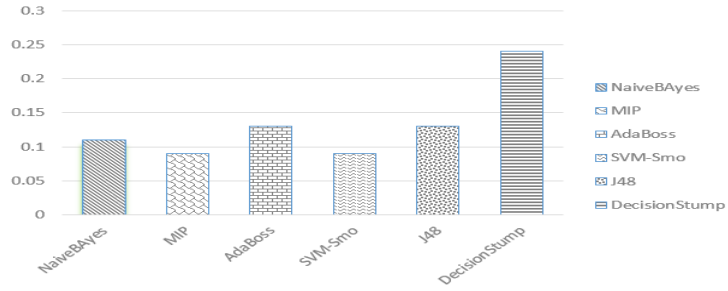
شکل ۴ – مقایسه پارامتر Kappa statistic

Mean absolute error: این پارامتر نزدیکی مقدار بین پیش بین ورودی و خروجی دقیق را محاسبه می کند که از فرمول ۲ بدست می آید. مقایسه این پارامتر برای شش طبقه بند در شکل ۵ مشاهده می شود.

(۲)

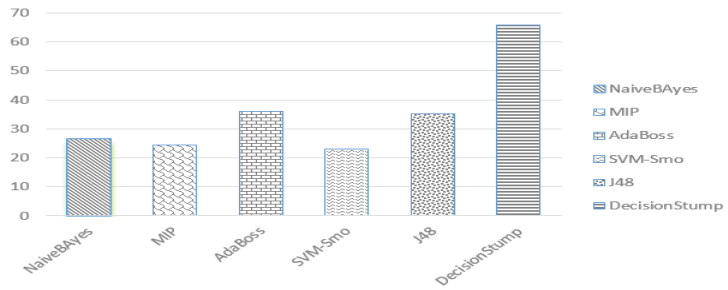
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

که در آن f_i مقدار پیش بینی و y_i مقدار درست هستند .



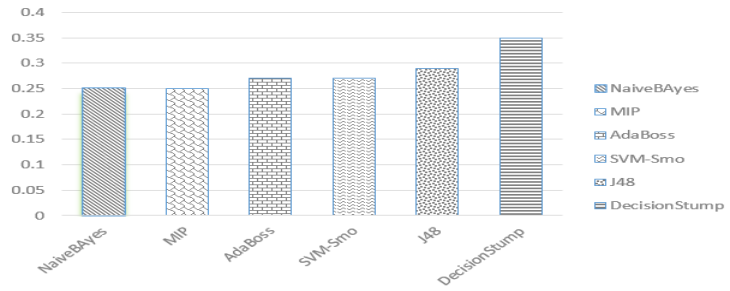
شکل ۵ – مقایسه پارامتر MAE

Root mean squared error : این پارامتر بیشتر در موارد برای محاسبه تفاوت بین مقادیر پیش بینی شده به وسیله مدل با مقدار مشاهده شده استفاده می شود.



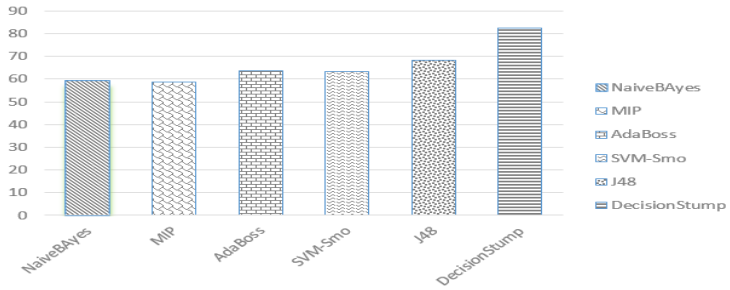
شکل ۶ – مقایسه پارامتر Root mean squared error

Relative absolute error: این پارامتر میانگین مقدار دقیق است. این پارامتر مقدار کل خطا را می گیرد و با تقسیم به وسیله خطای مطلق کل نرمال سازی می کند.



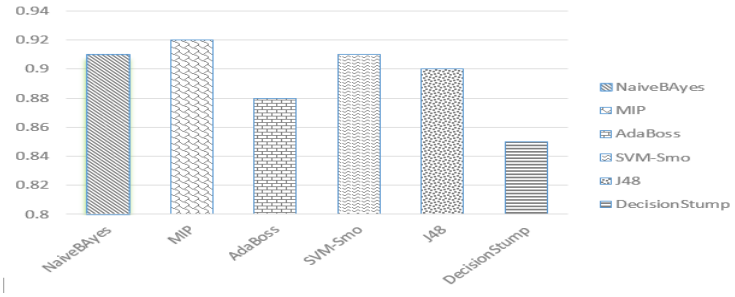
شکل ۷ – مقایسه پارامتر Relative absolute error

Root relative squared error: این پارامتر معیار پیش بینی ساده که میانگین مقدار دقیق را می گیرد، است. مقدار مربع خطای کل را می گیرد و به وسیله تقسیم آن به مرجع خطای کل به وسیله پیش بینی کننده، نرمال سازی می کند.



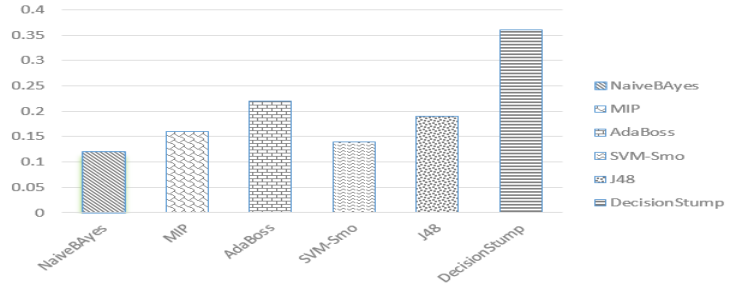
شکل ۸ – مقایسه پارامتر Root relative squared error

Avg TP Rate: نرخ مثبت درست نرخ تعداد بیماران PD پیش بینی شده به کل موارد مثبت است. مقایسه این پارمتر برای شش طبقه بند در شکل ۹ مشاهده می شود.



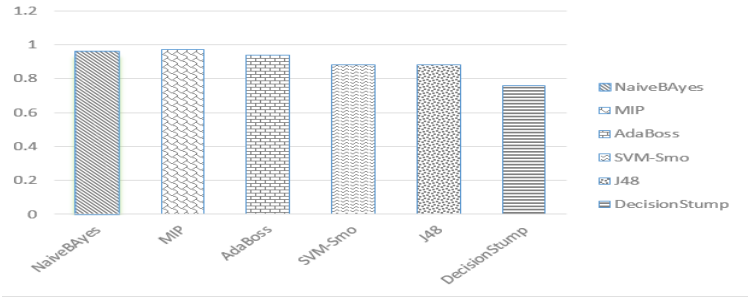
شکل ۹ – مقایسه پارامتر Avg TP Rat

Avg FP Rate: نرخ مثبت غلط نرخ تعداد بیماران سالم پیش بینی شده غلط به عنوان بیمار به کل موارد سالم است. مقایسه این پارمتر برای شش طبقه بند در شکل ۱۰ مشاهده می شود.



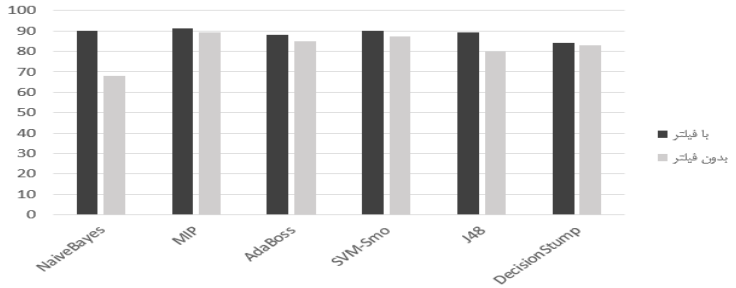
شکل ۱۰ – مقایسه پارامتر Avg FP Rate

Avg ROC Area: این پارامتر معیار صحت پیش بینی مدل منطقی است.



شکل ۱۱- مقایسه پارامتر Avg ROC Area

مقایسه پارامترهای مختلف بدست آمده از این شش طبقه بند بدون استفاده از هیچ یک از فیلترها در شکل ۱۲ مشاهده می شود.



شکل ۱۲ – نتیجه اعمال فیلتر و عدم اعمال فیلتر در طبقه بندی

۹- نتیجه‌گیری

پس از بارگذاری مجموعه داده ها مرحله پیش پردازش را برروی داده ها انجام شد. در انتها طبقه بندهای موردنظر برای تحلیل داده ها معرفی شدند. پس از بارگذاری مجموعه داده ها مرحله پیش پردازش را بر روی داده ها انجام شد. در ادامه نتایج اعمال شش طبقه بند J48 , NaiveBayes , SVM-Smo ,1,درخت تصمیم, شبکه عصبیMLP تحلیل و مقایسه شدند.

نتایج نشان داد که میزان صحت طبقه بندی این طبقه بندها بدین صورت است. شبکه عصبیMLP با ۹۲,۳۰٪ بیشترین مقدار صحت را بدست آورد پس از آن NaiveBayes و SVM-Smo ۹۱,۲۸ ٪ صحت وJ48 با ۸۹,۷۴٪ صحت و AdaBoostM1 ۸۸,۲۰ ٪ صحت را بدست آوردند. در انتها Decision Stump با ۸۴,۶۱ ٪کمترین مقدار صحت را بدست آورد.

مراجع

[1] parkinson australia(http://www.parkinsons.org.au/about-ps/about-pd.htm)

[2] Parkinson's UK (2013) http://www.parkinsons.org.uk/about_parkinsons/what_is_parkinsons.aspx

[3] Sanjivani Bhande , Dr. Ranjan Raut, "A Parkinson Diagnosis using Neural Network: a Survey", International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297:2013).

[4] Oana GEMAN, "A Fuzzy Expert Systems Design for Diagnosis of Parkinson's Disease", Proceedings of the 3rd International Conference on E-Health and Bioengineering - EHB 2011,24th26thNovember,2011, Iași, Romania.

[5] Freddie Astrom , Rasit Koker , "A parallel neural network approach to prediction of Parkinson's Disease", Expert Systems with Applications 38 (2011) 12470–12474.

[6] Dr. R.Geetha Ramani, G.Sivagami, Shomona Gracia jacob " Feature Relevance Analysis and Classification of Parkinson's Disease TeleMonitoring data Through Data Mining" , International Journal of Advanced Research in Computer Science and Software Engineering,vol-2,Issue 3, March 2012.

[7] I.H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann,2005.

[8] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P." From data mining to knowledge discovery in databases," Artificial Intelligence Magazine, 1996.

[9] Venkatadri, M., and Lokanatha, C. R. "A review on data mining from past to the future," International Journal of Computer Applications, 15(7), 19-22,2011.

[10] Chen, M. S., Han, J., and Yu, P. S." Data mining: An overview from a database perspective," IEEE Transactions on Knowledge and Data Engineering,8, 866–883,1996.

[11] F. S. Gharehchopogh, Peyman Mohammadi and Parvin Hakimi. Application of Decision Tree Algorithm for Data Mining in Healthcare Operations: A Case Study. International Journal of Computer Applications 52(6):21-26, August 2012.

[12] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. "Discovering data mining: From concept to implementation,". New Jersey: Prentice Hall,1997.

[13] Idicula-Thomas, S., Kulkarni, A. J., Kulkarni, B. D., Jayaraman, V. K., and Balaji, P. V." A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in escherichia coli," Bioinformatics, 2006.

[14] S. K. Yadev and Pal., S."Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," World of Computer Science and Information Technology (WCSIT), 2012.

[15] Loether, H. J., and McTavish, D. G." Descriptive and inferential statistics: An introduction (4th ed.),"Needham Heights, MA: Allyn and Bacon,1993.

[16] Kohavi, R. The power of decision tables, in: Lavrac, N., Wrobel, S., (Eds.), Machine Learning: Proceedings of the Eighth European Conference on Machine Learning ECML95, Lecture Notes in Artificial Intelligence, Springer Verlag, 914, Berlin, Heidelberg, NY, pp. 174–189, 1995.

[17] Berka P, Rauch J, Zighed DA. Data Mining and Medical Knowledge Management: Cases and Applications.Hershey: Idea Group Inc (IGI); 2009.

[18] Max A. Little, Suitability of Dysphonia Measurements for Telemonitoring of Parkinson’s Disease, IEEE Transactions on Biomedical Engineering, Vol. 56, No. 4, April 2009.