پیش‌بینی ریزش کارمندان با استفاده از الگوریتم‌های یادگیری گروهی مبتنی‌بر درخت تصمیم

سیده محبوبه مزارعی[1*]، جعفر پورامینی[2]

[1] مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور مرکز بین‌الملل عسلویه، ایران

[2] مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور، تهران، ایران

چکیده:

یکی از مهم‌ترین نگرانی‌های مدیران، ترک خدمت کارکنان کلیدی است؛ زیرا سازمان با ازدست‌دادن نیروهای ارزشمند خود، متحمل ازدست‌دادن دانش و تجربیاتی می‌شود که طی سال‌ها تلاش به‌دست‌آمده‌است؛ بنابراین، پیش‌بینی ریزش کارکنان به مدیران منابع انسانی در استخدام نیروهای ماندگار و حفظ و نگهداری آنها کمک می‌کند. یکی از ابزارهای کارآمد دراین‌خصوص استفاده از روش‌های مختلف داده‌کاوی است. تعداد کم نمونه‌ها و نامتوازن‌بودن داده‌های ریزش کارکنان و تنظیم ابرشاخص‌ها از جمله مشکلات استفاده از روش‌های داده‌کاوی برای پیش‌بینی ریزش کارکنان است. هدف این پژوهش، ارائهٔ روش‌های مناسب کاهش ویژگی و پیش‌پردازش داده‌ها به همراه ارائهٔ راهکار برای تنظیم مناسب ابرشاخص‌ها برای پیش‌بینی ریزش کارکنان با استفاده از روش‌های مختلف داده‌کاوی و الگوریتم‌های یادگیری ماشین تجمعی است. باتوجه‌به نامتوازن‌بودن داده‌ها، از روش‌های کم نمونه‌گیری تصادفی و ترکیب آن با بیش نمونه‌گیری تصادفی برای متوازن‌سازی داده‌ها با نسبت‌های متفاوت استفاده شد. باتوجه‌به مثبت‌بودن همهٔ داده‌ها از روش کاهش ابعاد تجزیهٔ ماتریس نامنفی (NMF) استفاده و با استفاده از روش‌های جستجوی ابرشاخص‌ها، مقادیر بهینهٔ آن‌ها برای الگوریتم‌های پیشنهادی، تعیین شد. برای ارزیابی روش پیشنهادی از مجموعه‌داده‌های استاندارد با اندازه‌های مختلف استفاده شده‌است. نتایج روش پیشنهادی نیز با نتایج حاصل از سایر روش‌های مطرح در این حوزه مانند KNN، AdaBoost، DT وSVC مقایسه شده‌است. نتایج به‌دست‌آمده نشان می‌دهند که الگوی پیشنهادی این پژوهش نسبت به پژوهش‌هایی که در گذشته روی همین داده‌ها صورت‌گرفته، دارای دقت پیش‌بینی بهتری است. طبق بررسی‌های انجام‌شده در این پژوهش که با استفاده از یک روش ترکیبی انتخاب ویژگی انجام شده، ویژگی‌های «سن»، «درآمد ماهیانه»، «نرخ روزانه»، «اضافه‌کاری» و «تعداد شرکت‌هایی که کارمند در آنها کار کرده»، بیش‌ترین تأثیر را بر ریزش کارکنان داشته‌اند.

واژگان کلیدی: داده‌کاوی، مدیریت منابع انسانی، یادگیری تجمعی، ریزش کارکنان

Predicting Employee Turnover Using Tree-Based Ensemble Learning Algorithms

Seyede mahboobe mazarei[1*], Jafar pooramini[2]

[1]Computer Engineering and Information Technology, Payam Noor University, Asalouye International Center, Iran

[2]Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran

Abstract

Key employee's turnover is one of the most important concerns of Human Resource Managers (HRM); Since the organization by losing its valuable staff, suffers from the loss of skills and experience gained over the years, predicting employee turnover helps HRMs to hire and retain permanent employees. One of the effective tools in this regard is the use of different data mining methods. Many researchers have done research in this field. This study reviewes recently published articles based on machine learning models, using Kaggle Human Resource (HR) databases [1-5] to compare them with this proposed models. In the article [9], the authors have selected 11 of the most important features by collecting common features from previous articles and filtering them using feature review and selection

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۰/۸/۱۰ • تاریخ پذیرش: ۱۴۰۲/۴/۱۷ • تاریخ انتشار: ۱۴۰۲/۱۰/۲۴ • نوع مطالعه: کاربردی

[DOI: 10.61186/jsdp.20.3.73]

algorithms. After converting non-numerical variables to numerical and normalizing the data in the range [0,1], those attrition prediction approach is based on machine, deep and ensemble learning models and is experimented on a large-sized and a medium-sized simulated HR datasets, and then a real small-sized dataset from a total of 450 responses. Those approach achieves higher Accuracy (0.96, 0.98 and 0.99 respectively) for the three datasets, when compared previous solutions. In 2021, authors examined the relationship between features using Pearson correlation coefficient and selected 11 features with the highest correlation coefficient. Then used from six different machine learning algorithms including Random Forest (RF), Logistic Regression (LR), …, to predict employee turnover. The highest accuracy they obtained, was 0.85 for RF [3]. In the article[1], the authors used two IBM datasets and a database containing HR information from a regional bank in the USA to predict employees turnover. After cleaning and preprocessing the data, the performance of 10 different machine learning algorithms such as Decision Tree (DT), RF, LR, Neural Network, …, was evaluated using ROC criteria on 10 small, medium, and large subsets of randomly selected, unassigned primary datasets. The average accuracy of algorithms is 0.83 in small datasets, 0.81 in medium datasets and 0.86 in large datasets. The authors of the paper [4] used three main experiments on IBM Watson simulated datasets to predict employees turnover. The first experiment involved training the original class-imbalanced dataset with the following machine learning models: support vector machine with several kernel functions, random forest and K-nearest neighbour (KNN). The second experiment focused on using adaptive synthetic (ADASYN) approach to overcome class imbalance, then retraining on the new dataset using the abovementioned machine learning models. As a result, training an ADASYN-balanced dataset with KNN (K = 3) achieved the highest performance, with 0.93 F1-score. This turnover prediction approach is based on tree-based ensemble learning models and is experimented on a large-sized standard simulated HR dataset (hr_data), including 15,000 samples with 10 features and a medium-sized (IBM) including 1470 samples with 34 features. The employees turnover rate in the IBM is 16.1% and in the hr_data is 23.8%, so datasets are unbalanced. To balance the data, the random-under-sampling technique and its combination of random-over-sampling with a ratio of 0.5965 for the IBM and 0.6558 for the hr_data has been used. In the preprocessing stage, Features with zero variance and samples containing the missing value were also removed. Then, categorical (non-numeric) values were converted to binary fields and then All features were scaled using data normalization in [0,1]. In order to reduce the feature dimensions in the IBM dataset, we used the "Non-negative Matrix Factorization" (NMF) technique (n_components=17, max_iter=500) and For initialization, non-negative singular value analysis method with zeros filled with X value has been used. After reviewing and cleaning the data, in the processing stage, six different classification algorithms, including KNN (k=1), RF (number of trees= 1500), DT, ExtraTreesClassifier (number of trees= 1000) and Support Vector Classifier were training on 70% of data. The optimal value of the hyperparameters for the algorithms, was set using RandomizedSearchCV and GridSearchCV techniques. In order to investigate the effect of balancing and Dimensionality Reduction on the performance of models, experiments were performed in 3 stages (befor balancing, after balancing befor Dimensionality Reduction, after balancing and Dimensionality Reduction) on 30% of the remaining data. The results shown in Table (2-4) indicate that this proposed model, which uses tree-based optimized ensemble learning algorithms with data balancing and NMF dimensionality reduction method, increases the f1score of turnover prediction. In the hr_data dataset, the best f1score for the RandomForest algorithm was 99.52% and for the IBM HR dataset, the best f1score for the ExtraTreesClassifier algorithm was 95.82%, which is higher than previous research. Table 5 compares the results of previous research with this research. Since, the prediction of employee attrition will not be enough without finding the characteristics that affect it, therefore, after building models and evaluating their performance, using a combined feature selection method by averaging the results of the single-variable feature selection method called "SelectKBest", and A wrapper feature selection method called "Recursive feature elimination" (RFE) with four learning algorithms RF, DT, ExtraTreesClassifier and AdaBoost, the most effective features were selected. SelectKBest combines the chi2 univariate statistical test with the selection of K features based on the statistical result between the features and the target variable. Also, in the RFE method, machine learning algorithms are used to remove the least important features after recursive training, so that finally the number of features reaches the set number (17 features in this article). The performance results of the models based on the selected features are shown in Table 6. The most effective characteristics are "age", "daily rate", "over time", "NumCompaniesWorked" and, "monthly income" .

Keywords: **data mining, human resource management, ensemble learning, employee turnover**

۱- مقدمه

تجزیه و تحلیل منابع انسانی نقش مهمی در هر جنبه‌ای از عملکرد منابع انسانی در سازمان‌ها، از جمله استخدام، آموزش و توسعه، حفظ، تعامل و جبران خسارت ایفا می‌کند. کارکنان به‌طور گسترده‌ای به‌عنوان دارایی ضروری و ارزشمند یک سازمان شناخته می‌شوند. تخصص، توانمندی و تجربهٔ آنها برای موفقیت سازمان مهم است. در

[DOI: 10.61186/jsdp.20.3.73]

این زمینه، ریزش کارکنان توجه بسیاری از کارشناسان منابع انسانی را به خود جلب‌کرده‌است، زیرا به‌طور گسترده به‌عنوان یک شاخص مهم رقابت‌پذیری درنظرگرفته‌می‌شود [۱]. ریزش کارکنان به فرآیند ازدست‌دادن کارمندان و جایگزینی آنها با استخدام‌های جدید اشاره دارد [۲]، که سازمان‌ها را در سراسر صنایع تحت‌تأثیر قرار می‌دهد و موضوع پژوهش‌ها و تحلیل‌های گسترده‌ای در سال‌های اخیر بوده‌است. تأثیر ریزش کارکنان بر بهره‌وری، روحیه و موفقیت کلی سازمان، آن را به موضوعی حیاتی تبدیل می‌کند که نیازمند توجه و تحلیل است [۳, ۴, ۶]. نرخ بالای ریزش کارکنان، هزینه‌های قابل‌توجهی را برای سازمان‌های مرتبط با سرمایهٔ انسانی، استخدام، آموزش و توسعهٔ کارکنان جدید متحمل می‌شود [۵]. بنابراین، درک علل ریزش کارکنان و توسعهٔ راهبردهای مؤثر برای کاهش اثرات منفی آن، برای موفقیت بلندمدت و رقابت‌پذیری سازمان از اهمیت بالایی برخوردار است. به‌دست‌آوردن درکی کامل از علل اساسی این پدیده برای مدیریت مؤثر ریزش کارکنان ضروری است. این امر مستلزم آن است که سازمان‌ها در تجزیه و تحلیل عمیق فرهنگ محل کار، شیوه‌های رهبری و سطوح رضایت کارکنان خود شرکت کنند [۶]. با انجام این کار، سازمان‌ها می‌توانند راهبردهای هدفمندی را برای کاهش ریزش کارکنان و اثرات منفی آن بر بهره‌وری، روحیه و موفقیت کلی سازمان توسعه دهند. سازمان‌هایی که رویکردی فعالانه برای مدیریت ریزش کارکنان دارند از ثبات نیروی کار بهبودیافته، کاهش هزینه‌های مرتبط با استخدام و آموزش و افزایش رقابت درازمدت بهره‌مند می‌شوند. هدف از انجام این پژوهش، استفاده از روش‌های داده‌کاوی و مقایسهٔ الگوریتم‌های مختلف یادگیری ماشین بر روی داده‌های استاندارد منابع انسانی و رسیدن به بالاترین سطح دقت برای پیش‌بینی ریزش کارکنان است تا بتوانیم از این طریق به مدیران منابع انسانی سازمان‌ها کمک کنیم هزینه‌های ناشی از ریزش کارکنان را کاهش دهند و تمهیدات لازم را برای استخدام نیروهای ماندگار در سازمان به کار گیرند و از خطر ازدست‌دادن آگاهی در امان بمانند. همچنین، با پیش‌بینی ریزش کارکنان، تمهیدات لازم را برای جبران آن درنظربگیرند. داده‌کاوی، تنها دارای یک استاندارد رسمی در جهان است که **CRISP-DM** نام دارد. الگوی **CRISP-DM**، یک الگوی استاندارد برای اجرای فرایندهای داده‌کاوی است که توسط بزرگان و رهبران صنعت، با همکاری کاربران مجرب داده‌کاوی و تولیدکنندگان ابزارهای نرم‌افزاری داده‌کاوی ایجاد شده‌است. الگوی **CRISP-DM** فعالیت‌های داده‌کاوی را به شش مرحله دسته‌بندی می‌کند که هریک وظایف مختلفی دارند: فهم تجاری، فهم داده، آماده‌سازی داده، الگوسازی، ارزیابی و پیاده‌سازی.

در این مقاله، از الگوریتم‌های یادگیری ماشین[1] مختلف، برای پیش‌بینی ریزش کارکنان، استفاده شد. به‌منظور بهبود دقت پیش‌بینی، پیش‌پردازش‌های متفاوتی بر روی مجموعه‌داده‌های استاندارد، انجام شده‌است. نتایج به‌دست‌آمده نشان می‌دهد که روش پیشنهادی این پژوهش که استفاده از الگوریتم‌های بهینه‌شده یادگیری تجمعی[2] جنگل تصادفی (**RF**)[3] و **ExtraTrees** است، به‌همراه استفاده از روش کاهش ابعاد «تجزیهٔ ماتریس نامنفی»[4] (**NMF**)، متوازن‌سازی داده‌ها و تنظیم صحیح ابرشاخص‌ها[5]، باعث افزایش دقت پیش‌بینی می‌شود. جنگل تصادفی یک روش یادگیری تجمعی مبتنی‌بر درخت است که در آن درختان زیادی با استفاده از زیرمجموعه‌های کوچکی از داده‌ها (با جایگزینی) که بوت‌استرپ[6] نیز نامیده می‌شوند، ساخته‌می‌شوند. در نهایت، از رأی اکثریت برای تعیین پیش‌بینی الگو استفاده می‌شود. این مفهوم به‌عنوان **Bagging**[7] شناخته می‌شود. جنگل‌های تصادفی با درخت‌های استاندارد متفاوت هستند، زیرا هر گره با استفاده از بهترین‌ها در میان زیرمجموعه‌ای از پیش‌بینی‌کننده‌ها که به‌طور تصادفی در آن گره انتخاب شده‌اند، تقسیم می‌شود، که آن را در برابر بیش‌برازش[8] مقاوم می‌کند [۲].

الگوریتم **ExtraTrees**، نوعی روش یادگیری تجمعی است که تعدادی درخت تصمیم تصادفی‌شده (معروف به درخت‌های اضافی) را در زیرنمونه‌های مختلف مجموعه‌داده برازش می‌دهد و از میانگین‌گیری برای بهبود دقت پیش‌بینی و مدیریّت بیش‌برازش استفاده می‌کند. تفاوت آن با جنگل تصادفی این است که جنگل تصادفی تقسیم بهینه را انتخاب می‌کند، درحالی‌که **ExtraTrees** آن را به‌طور تصادفی انتخاب می‌کند. بااین‌حال، پس‌از

[1] Machine Learning
[2] Ensembel Learning
[3] Random Forest
[4] Non-negative matrix actorization
[5] Hyper Parameters
[6] Bootstrap
[7] Bootstrap Aggregation
[8] Over-fitting

[DOI: 10.61186/jsdp.20.3.73]

انتخاب نقاط تقسیم، این دو الگوریتم بهترین را از بین همهٔ زیرمجموعه‌ویژگی‌ها انتخاب می‌کنند.

این مقاله به شرح زیر تنظیم شده‌است: در بخش دو، پیشینهٔ پژوهش مطالعه شده‌است. در بخش سه، روش پیشنهادی بررسی و ارزیابی شده و مهم‌ترین ابرشاخص‌ها تعیین شده‌اند. در بخش چهار نتایج پژوهش نشان داده شده‌است. بخش پنج به بحث و بررسی نتایج این پژوهش و مقایسهٔ آنها با نتایج به‌دست‌آمده در مقالات قبل پرداخته و نتیجه‌گیری از پژوهش در بخش آخر آمده‌است.

۲- پیشینهٔ پژوهش

در زمینهٔ مدیریت منابع انسانی و روش‌های پیش‌بینی ریزش کارکنان، پژوهش‌های داخلی و خارجی فراوانی صورت‌گرفته و از روش‌های مختلفی برای این کار استفاده شده‌است. یکی از این روش‌ها استفاده از روش‌های مختلف داده‌کاوی و الگوریتم‌های یادگیری ماشین است که پیشتر توسط پژوهشگران مختلف مطالعه شده‌است. ازجمله، نویسندگان مقالهٔ [۷] قابلیت‌های پیش‌بینی هفت الگوریتم مختلف یادگیری ماشین، از جمله الگوریتم‌های به‌تازگی توسعه‌یافته **XGB**[1] را در مورد ریزش کارمندان مقایسه کردند. به‌طور مشابه، *سیکارودی و همکارانش* [۸] شبیه‌سازی‌هایی را برای پیش‌بینی ریزش کارمندان با استفاده از ده الگوریتم داده‌کاوی مختلف، از جمله آزمایش روی انواع مختلف شبکه‌های عصبی و روش‌های قاعدهٔ القایی، انجام دادند.

دولت‌آبادی و کی‌نیا [۹]، چهار الگوی پیش‌بینی ریزش مشتری و کارمند، از جمله درخت تصمیم[2] (**DT**)، بیز ساده[3] (**NB**)، ماشین بردار پشتیبان[4] (**SVM**)و شبکهٔ عصبی[5] (**NN**) را بررسی کردند. الگوریتم **SVM**، به بالاترین دقت پیش‌بینی (۹۹/۸۳) دست یافت. پس‌از **SVM**ها، روش **NB** با ۹۲/۳۷ درصد بالاترین دقت پیش‌بینی را داشت.

محسن سرداری زارچی و همکارانش [۱۰] در سال ۱۳۹۷، در مقاله‌ای، با استفاده از مجموعه‌دادهٔ به‌دست‌آمده از نظرات کارکنان شرکت بهره‌برداری نفت و گاز کارون، مبنی‌بر رضایت‌مندی و تمایل به ترک سازمان، با استفاده از شبکه‌های عصبی مصنوعی به‌عنوان طبقه‌بند و الگوریتم تکاملی ژنتیک چندمنظوره برای انتخاب ویژگی‌های مؤثر، یک سامانهٔ خبره با دقت ۰/۸۸ طراحی کردند.

مادارا پرت[6] و همکارانش [۳] در سال ۲۰۲۱، ارتباط بین ویژگی‌ها را با استفاده از ضریب همبستگی پیرسون بررسی کرده و ۱۱ ویژگی با بیشترین ضریب همبستگی را انتخاب نموده‌اند. سپس، از شش الگوریتم مختلف یادگیری ماشین از جمله **RF**، **DT** رگرسیون لجستیک[7] (**LR**)، بیز سادهٔ گوسی (**GNB**)[8]، **k**-نزدیکترین همسایه (**KNN**)[9] و **SVM**، برای پیش‌بینی ریزش کارکنان استفاده کرده‌اند. بالاترین دقت پیش‌بینی در این پژوهش، مربوط به **RF** و برابر با ۰/۸۵ است.

نویسندگان در مقالهٔ [۲]، یک رویکرد تجزیه و تحلیل افراد را برای پیش‌بینی ریزش کارکنان پیشنهاد کرده‌اند، که با تمرکز بر کیفیت داده‌ها به‌جای کمیت، از یک دادهٔ حجیم به یک بافت دادهٔ عمیق تغییر می‌کند. در این روش، نخست، با جمع‌آوری ویژگی‌های رایج از مقالات (یک پژوهش اکتشافی) و فیلترکردن آنها با استفاده از الگوریتم‌های بررسی و انتخاب ویژگی، ۱۱ مورد از مهم‌ترین ویژگی‌ها را انتخاب کرده‌اند. در مرحلهٔ پیش‌پردازش، از روش رمزگذاری **One-hot**، برای تبدیل متغیرهای غیرعددی به متغیرهای عددی استفاده شده‌است. برای پیش‌بینی ریزش کارکنان از الگوهای یادگیری ماشین **DT**، **LR** و **SVM**، یادگیری تجمعی **RF**، **XGB**[10]، **VotingClassifier** بر اساس رأی‌گیری اکثریت[11] و یک الگوی پشته‌ای مبتنی‌بر شبکهٔ عصبی شامل شبکه‌های **DNN**[12]، **CNN**[13] و **LSTM**[14] استفاده شده‌است. الگوها را با استفاده از مجموعه‌داده‌های شبیه‌سازی‌شدهٔ منابع انسانی با اندازهٔ بزرگ (**hr_data**)، متوسط (**IBM**) و سپس یک مجموعه‌داده با اندازهٔ کوچک واقعی با ۴۵۰ نمونه، آزمایش شدند و در نهایت، به‌ترتیب به دقت‌های (۰/۹۸، ۰/۹۶ و ۰/۹۹) برای سه مجموعه‌داده دست‌یافته‌اند که نسبت به ۱۳ مقاله‌ای که در پژوهش خود بررسی کرده‌اند، دقت‌های بسیار خوبی است. بهترین دقت در این مقاله مربوط به الگوریتم یادگیری تجمعی **VotingClassifier** است.

[1] Extreme Gradient Boosting
[2] Decision Tree
[3] Naïve Bayes
[4] Support Vector Machines
[5] Neural Network

[6] Madara Pratt
[7] Logistic Regression
[8] Gaussian Naive Bayes
[9] K Nearest Neighbor
[10] Extreme Gradient Boosting
[11] Majority voting
[12] Deep Neural Networks
[13] Convolutional Neural Networks
[14] Long Short-Term Memory Networks

[DOI: 10.61186/jsdp.20.3.73]

نویسندهٔ مقالهٔ [۵]، از ترکیب هوش مصنوعی، داده‌کاوی، ریاضی و آمار، برای به‌دست‌آوردن درک و بینش بهتر از عملکرد فرایند کسب‌وکار استفاده می‌کند و با پیشنهاد نوعی سامانهٔ تحلیلی پیش‌بینی‌کننده که بر اساس بیز ساده استوار است، پیش‌بینی می‌کند که آیا کارکنان، سازمان را زودتر از موعد ترک می‌کنند یا خیر.

کای و همکاران [۱۱] روش جدیدی به نام جاسازی گراف دوبخشی پویا[1] (**DBGE**) برای تعیین چگونگی پیش‌بینی ریزش کارکنان پیشنهاد کرد. این روش نمایش‌های برداری کم‌بعدی را برای نمودارهای دو بخشی پویا آموزش داد. آزمایش‌ها با مجموعه داده‌های دنیای واقعی از یکی از بزرگ‌ترین شبکه‌های اجتماعی حرفه‌ای برخط چین نشان داد که ویژگی‌های آموخته‌شده به‌وسیلهٔ **DBGE** ریزش کارکنان را بسیار دقیق‌تر پیش‌بینی می‌کنند.

نویسندگان در مقالهٔ [۱]، از دو مجموعه‌دادهٔ متفاوت برای پیش‌بینی ریزش کارکنان استفاده کرده‌اند. نخستین مجموعه‌داده شامل اطلاعات منابع انسانی یک بانک منطقه‌ای در ایالات متحده آمریکاست که از سال ۲۰۱۳ تا ۲۰۱۶ جمع‌آوری شده‌است و شامل ۱۴۳۲۲ ورودی کارمند و ۲۴ ویژگی است. مجموعه‌دادهٔ دوم یک مجموعه‌دادهٔ شبیه‌سازی‌شده است که توسط شرکت **IBM** ایجاد شده و برای تسهیل تجزیه و تحلیل دقیق‌تر در این پژوهش گنجانده شده‌است. برای آماده‌سازی داده‌ها، همهٔ ویژگی‌هایی را که دارای مقادیر ثابت بودند، حذف کرده‌اند. در برخورد با داده‌های گم‌شده داده‌های عددی را با مقدار میانگین و داده‌های غیرعددی را با مقدار میانهٔ آن ویژگی برای همهٔ ورودی‌ها، جایگزین کرده‌اند. داده‌های با مقادیر غیرعددی را با روش رمزگذاری برچسب[2]، به مقادیر عددی تبدیل کرده‌اند. سپس، عملکرد ده الگوریتم مختلف یادگیری ماشین مانند **KNN**، **DT**، **RF**، **GBT**[3]، **XGB**، **LR**، **SVM**، **NN**، **LDA**[4] و **NB** را، بر روی ده زیرمجموعهٔ کوچک، متوسط و بزرگ از مجموعه‌داده‌های اولیه که به‌طور تصادفی و بدون جای‌گذاری، انتخاب کرده‌اند، با استفاده از معیار **ROC**[5] ارزیابی کردند. در این مقاله، دقت الگوریتم‌ها به‌طور میانگین، در مجموعه‌داده‌های کوچک برابر ۰/۸۳، در مجموعه‌داده‌های متوسط برابر با ۰/۸۱ و در مجموعه‌داده‌های بزرگ برابر با ۰/۸۶ به‌دست‌آمده‌است.

جین و همکاران [۱۲] ریزش کارکنان را با استفاده از الگوهای یادگیری ماشین **SVM**، **DT** و **RF** و مجموعه‌داده‌های سامانهٔ اطلاعات منابع انسانی با بیش از چهارده‌هزار رکورد و ده ویژگی، پیش‌بینی کرد. بر اساس یافته‌های این پژوهش، عملکرد **RF** از سایر طبقه‌بندی‌کننده‌ها بهتر و دقیق‌تر بود.

نویسندگان مقالهٔ [۱۳] از مجموعه‌داده‌های نظرسنجی بزرگ و واقعی از سراسر اروپا برای شناسایی بهترین روش‌های یادگیری ماشین برای پیش‌بینی ریزش کارکنان استفاده کردند. **KNN**، **DT**، **LR**، **RF**، **XGBoost**، **LightGBM (LGBM)** و معماری یادگیری جدولی عمیق (**TabNet**)، برخی از روش‌های استفاده‌شده بودند. نتایج نشان داد که **LR** و **LGBM** دقیق‌ترین الگوها برای پیش‌بینی این که کدام کارمندان ترک کار می‌کنند، بودند.

نویسندگان مقالهٔ [۴]، از سه آزمایش اصلی بر روی مجموعه‌دادهٔ شبیه‌سازی‌شدهٔ **IBM Watson**، برای پیش‌بینی ریزش کارکنان استفاده کرده‌اند. اولین آزمایش شامل آموزش مجموعه‌دادهٔ اصلی ردهٔ نامتوازن با الگوهای یادگیری ماشینی **SVM** با چند تابع هسته[6]، **RF** و **KNN** بود. در آزمایش دوم بر روی استفاده از رویکرد مصنوعی تطبیقی[7] (**ADASYN**) برای متوازن‌سازی داده‌ها متمرکز شدند و در آزمایش سوم، از نمونه‌برداری دستی از داده‌ها برای متوازن‌سازی داده‌ها استفاده کردند. در نتیجه، بهترین عملکرد (**f1score**)، مربوط به الگوریتم **KNN** (**K=3**) در مجموعه‌داده متوازن‌شده با روش **ADASYN**، برابر با ۰/۹۳ بود. همچنین، دقت الگوریتم **RF** بعداز انتخاب ۱۲ ویژگی از کل ۲۹ ویژگی، برابر با ۰/۹۱ به‌دست آمده‌است. در جدول (۱) مقایسه‌ای بین پژوهش‌های ارائه‌شده در این بخش نمایش داده شده‌است.

در این پژوهش، به مقالاتی که به‌تازگی منتشر شده و مبتنی‌بر الگوهای یادگیری ماشینی و عمیق هستند، و از مجموعه‌داده‌های منابع انسانی شبیه‌سازی‌شده **Kaggle** استفاده کرده‌اند [۱-۴]، استناد شده‌است. دلیل این انتخاب، وجود نتایج آزمایش‌های مربوط به دقت الگوهای پیش‌بینی برای این مجموعه‌داده‌های استاندارد است تا بتوان آن‌ها را با الگوهای پیشنهادی در این پژوهش مقایسه کرد. در مقایسه با مطالعات انجام‌گرفته در گذشته، نوآوری‌های این پژوهش شامل موارد زیر است:

[1] Dynamic Bipartite Graph Embedding
[2] Label encoding
[3] Gradient Boosting Trees
[4] Linear Discriminant Analysis
[5] Receiver Operating Characteristic

[6] Kernel
[7] Adaptive Synthetic

[DOI: 10.61186/jsdp.20.3.73]

- ترکیب روش کم نمونه‌گیری تصادفی از داده‌های اکثریت و بیش‌نمونه‌گیری تصادفی از داده‌های اقلیت، با نسبت‌های متفاوت، برای متوازن‌سازی داده‌ها
- حذف ویژگی‌های دارای مقادیر گم‌شده و ویژگی‌های دارای مقادیر ثابت
- استفاده از یک روش ترکیبی انتخاب ویژگی، با میانگین‌گیری از نتایج حاصل از پنج روش متفاوت انتخاب ویژگی، به‌منظور جستجوی مؤثرترین ویژگی‌ها بر ریزش کارکنان
- استفاده از روش کاهش ابعاد تجزیهٔ ماتریس غیرمنفی، برای پیش‌گیری از پیچیدگی الگوها و افزایش عملکرد آن‌ها. از «مقداردهی اولیهٔ غیرمنفی تجزیه مقدار مفرد دوگانه[1]» (**NNDSVD**) استفاده شده‌است که مقادیر صفر با مقدار میانگین **X** پر شده‌اند.

استفاده از الگوریتم‌های تجمعی مبتنی‌بر درخت بهینه‌شده به‌وسیلهٔ تنظیم ابرشاخص‌ها با استفاده از روش‌های مختلف جستجوی ابرشاخص‌ها.

۳- روش پیشنهادی

روش پیشنهادی برای پیش‌بینی ریزش کارکنان، شامل مراحل مختلف است که بر اساس الگوی **CRISP-DM**، در شکل (۱) نشان داده شده‌است.

باتوجه‌به اهمیت پیش‌پردازش در کارایی الگو، در مرحلهٔ پیش‌پردازش روش‌هایی برای مشکلات داده‌های گم‌شده، تبدیل داده‌های غیرعددی به داده‌های عددی، حل مشکل عدم توازن داده‌ها، تغییر مقیاس داده‌ها با استفاده از عادی‌سازی در بازهٔ [۰،۱]، و کاهش ابعاد مجموعه‌داده با روش تجزیهٔ ماتریس غیرمنفی پیشنهاد شده‌است.

در مرحلهٔ بعد الگویی با استفاده از دو الگوریتم تجمعی **RF** و **ExtraTrees** بهینه‌شده به‌وسیلهٔ تنظیم ابرشاخص‌ها، برای دسته‌بندی داده‌ها پیشنهاد شده‌است.

در نهایت، نتایج به‌دست‌آمده از ارزیابی الگو، با نتایج پیش‌بینی چهار الگوریتم یادگیری **DT, KNN, SVC** و **AdaBoost** مقایسه شدند.

و در آخر به‌منظور شناسایی مؤثرترین ویژگی‌ها بر ریزش کارکنان، از یک روش ترکیبی انتخاب ویژگی استفاده شد که در آن از نتایج حاصل از پنج روش انتخاب ویژگی متفاوت، میانگین‌گیری و ویژگی‌های برتر انتخاب شد. در ادامه هر یک از موارد ذکر شده با جزئیات تشریح می‌شود.

۳-۱ مجموعه‌داده

باتوجه‌به‌این‌که پایگاه‌های دادهٔ منابع انسانی، به‌ویژه در سازمان‌های ایران، به‌طور معمول، ناقص و اغلب شامل داده‌های کم‌ارزش برای تجزیه و تحلیل هستند، بنابراین، در این پژوهش، به‌جای استفاده از داده‌های واقعی مربوط به سازمان‌های ایران، از دو مجموعه‌دادهٔ شبیه‌سازی‌شده استاندارد که از پایگاه **Kaggle.com** استخراج شده‌اند، استفاده شده‌است. مجموعه‌دادهٔ منابع انسانی شرکت **IBM (HR IBM)**، شامل ۱۴۷۰ نمونه با ۳۴ ویژگی ورودی و مجموعه‌دادهٔ **hr_data**، یک مجموعه‌دادهٔ شبیه‌سازی‌شده است که شامل پانزده‌هزار نمونه با ده ویژگی ورودی است. متغیرهای مستقل آن‌ها در جدول (۱) نمایش داده‌شده‌اند. متغیر هدف، ترک خدمت کارکنان **(Left)** است.

۳-۲ آماده‌سازی و پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها به‌طور معمول، در مطالعات پیش‌بینی ریزش کارکنان انجام می‌شود، زیرا مجموعه‌داده‌ها اغلب حاوی ورودی‌های گم‌شده[2]، درجات مختلف نوفه و تفاوت‌های اساسی در مقیاس[3] در هر ویژگی هستند. در این پژوهش، از روش‌های پیش‌پردازش داده که در ادامه معرفی می‌شوند، جهت تولید داده مناسب برای الگوریتم‌های یادگیری ماشین، استفاده شده‌است. در مجموعه‌دادهٔ **hr_data**، مقدار ویژگی سیگاری‌بودن **(is_smoker)** برای بیشتر نمونه‌ها ثبت نشده‌است؛ بنابراین، این ویژگی، به دلیل فراوانی داده‌های گم‌شده و نیز تمامی نمونه‌هایی که دارای دادهٔ گم‌شده بودند، حذف شدند. با بررسی مجموعه‌دادهٔ **IBM**، مشخص شد که این مجموعه‌داده، دادهٔ گم‌شده‌ای ندارد. در مرحلهٔ بعد، به این دلیل که بعضی از الگوریتم‌ها، قادر نیستند به طور مستقیم داده‌های غیرعددی را پردازش کنند، در هر دو مجموعه‌داده، مقادیر و ویژگی‌های غیرعددی، مانند وضعیت تأهل **(MaritalStatus)** به مقادیر عددی تبدیل شدند. در مجموعه‌داده‌های موردمطالعه در این پژوهش، ویژگی‌ها مقیاس‌های متفاوتی دارند. به‌طور مثال، مقدار ویژگی **MonthlyIncome** بین ۱۰۰۹ تا ۱۹۹۹۹ است، درحالی‌که مقدار ویژگی **Age** بین ۱۸ تا ۶۰ تغییر می‌کند؛

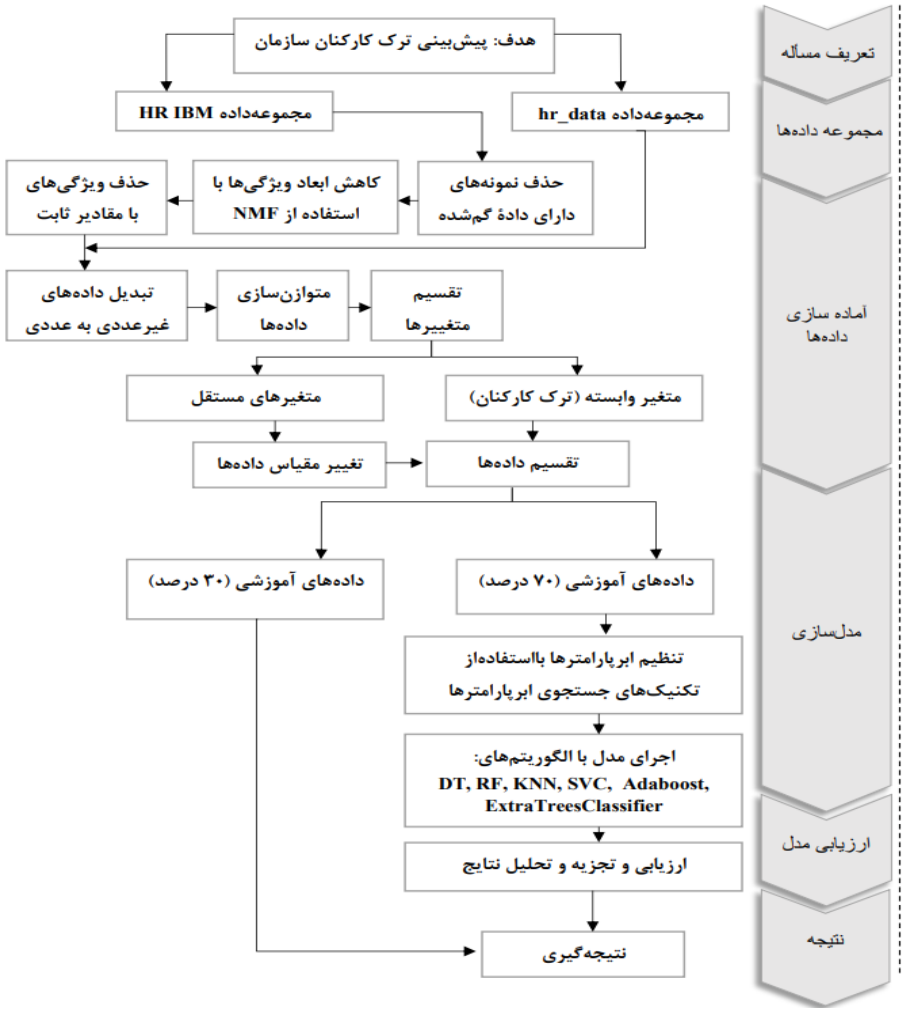
[1] Nonnegative Double Singular Value Decomposition

[2] Missing Value

[3] Scale

[DOI: 10.61186/jsdp.20.3.73]

بنابراین، نمی‌توان چنین داده‌هایی را به الگوریتم‌های یادگیری ماشین داد؛ چراکه بعضی از الگوریتم‌های یادگیری ماشین به دادهٔ مقیاس‌بندی‌شده نیاز دارند.



(شکل- ۱): نمایی کلی از فرایند پیش‌بینی با استفاده از استاندارد CRISP-DM

(Figure- 1) :An overview of the predicting process using CRI

در این پژوهش‌ها، برای تنظیم دامنهٔ ویژگی‌ها، داده‌ها در بازهٔ [۰،۱] با روش **minmax** عادی‌سازی شدند. ازآنجاکه تعداد ویژگی‌های مجموعه‌دادهٔ **IBM** زیاد است، نخست، تمامی ویژگی‌های با مقادیر ثابت که تأثیری در نتیجهٔ پیش‌بینی الگوریتم‌ها ندارند، مانند **EmployeeCount**، **over18** و **standard hourse** از مجموعه‌دادهٔ **IBM** حذف شدند. سپس، باتوجه‌به‌این‌که همهٔ داده‌ها دارای مقادیر مثبت هستند، از روش کاهش ابعاد «تجزیهٔ ماتریس غیرمنفی[1]» (**NMF**)، برای جلوگیری از پیچیدگی الگو استفاده شده و ابعاد فضای ویژگی‌ها به ۱۷ کاهش یافت. به‌طورکلی، تجزیهٔ ماتریسی ابزاری برای تحلیل داده‌هاست. هر تجزیه تعبیرهای مختلفی را از ساختار ضمنی داده‌ها آشکار می‌سازد، که البته این تعبیرها از نظر ریاضی هم‌ارز هستند. اگر $\mathbf{X} \in \mathbb{R}^{m \times n}$، ماتریس ویژگی‌ها باشد (**m** تعداد مشاهدات و **n** تعداد ویژگی‌ها است)، **NMF** ماتریس‌های $\mathbf{W} \in \mathbb{R}^{m \times r}$ و $\boldsymbol{H} \in \mathbb{R}^{r \times n}$ را با شرط $r \ll min\{m, n\}$ چنان می‌یابد که رابطهٔ تقریبی $\boldsymbol{X} \approx \boldsymbol{WH}$ برقرار باشد. معادلهٔ تقریبی آن به‌صورت مسئلهٔ بهینه‌سازی که در رابطهٔ (۱) نشان‌داده‌شده، تبدیل می‌شود.

پیداکردن چنین تقریبی، نیازمندِ تابع هزینه‌ای است که کیفیت تقریب را به‌خوبی نشان دهد. یکی از این تابع‌ها می‌تواند اندازه فاصلهٔ دو ماتریس نامنفی از یکدیگر باشد. در این پژوهش، تابع هزینه بر مبنای روش فروبنیوس[2] پایه‌ریزی شده‌است.

[1] Non-negative matrix actorization

[2] Frobenius

[DOI: 10.61186/jsdp.20.3.73]

(جدول- ۱): ویژگی‌های مستقل در مجموعه‌دادهٔ hr_data و IBM

(Table- 1): Independent properties in the hr_data and IBM datasets

مجموعه‌داده IBM			
نام ویژگی	شماره	نام ویژگی	شماره
MaritalStatus	18	Age	1
MonthlyIncome	19	BusinessTravel	2
MonthlyRate	20	DailyRate	3
NumCompaniesWorked	21	Department	4
DistanceFromHome	22	Over18	5
TotalWorkingYears	23	Education	6
TrainingTimesLastYear	24	EducationField	7
EmployeeNumber	25	EmployeeCount	8
PercentSalaryHike	26	OverTime	9
PerformanceRating	27	StandardHours	10
WorkLifeBalance	28	Gender	11
YearsAtCompany	29	HourlyRate	12
RelationshipSatisfaction	30	JobInvolvement	13
EnvironmentSatisfaction	31	JobLevel	14
StockOptionLevel	32	JobRole	15
YearsSinceLastPromotion	33	JobSatisfaction	16
YearsInCurrentRole	34	YearsWithCurrManager	17
مجموعه‌دادهٔ hr_data			
time_spend_company	6	work_accident	1
satisfaction_level	7	last_evaluation	2
promotion_last_5years	8	number_project	3
salary	9	average_montly_hours	4
department	10	is_smoker	5

(رابطه- ۱)

$$min\, f\,(W,H) \cong \frac{1}{2} * \| X - WH \|_{Fro}^{2} + alpha_W * l1_{ratio} * n_f eatures * \| vec(W) \|_1 + alpha_H * l1_{ratio} * n_s amples * \| vec(H) \|_1 + \frac{1}{2} * alpha_W * (\mathbf{1} - \boldsymbol{l1_{ratio}}) * \boldsymbol{n_f eatures} * \| \boldsymbol{W} \|_{\boldsymbol{Fro}}^{\mathbf{2}} + \frac{\mathbf{1}}{\mathbf{2}} * \boldsymbol{alpha_H} * (\mathbf{1} - \boldsymbol{l1_{ratio}}) * \boldsymbol{n_s amples} * \| \boldsymbol{H} \|_{\boldsymbol{Fro}}^{\mathbf{2}},\ W \geq 0, H \geq 0$$

در رابطهٔ بالا، $\| \boldsymbol{W} \|_{\boldsymbol{Fro}}^{\mathbf{2}}$ نرم فروبنیوس است که در رابطهٔ ۲ و $\| \boldsymbol{vec(W)} \|_{\mathbf{1}}$ ضرب درایه‌ای نرم **L1** است که در رابطهٔ (۳) نشان داده شده‌است.

(رابطه- ۲) $\| W \|_{Fro}^{2} = \sum_{ij} W_{ij}^{2}$

(رابطه- ۳) $\| vec(W) \|_1 = \sum_{ij} abs(W_{ij})$

Alpha، **Alpha_W** و **Alpha_H** مقادیری ثابت هستند که برای چندبرابرکردن میزان منظم‌سازی[1] **X**، **W** و **H** به کار می‌رود که به‌منظور حذف منظم‌سازی، مقادیر آنها برابر با صفر قرار داده شده‌است.

در **NMF** دو مبحث مقداردهی اولیه و شرط توقف، در دست‌یابی به عوامل ماتریسی بهین و هم‌گرایی نقش به‌سزایی دارند. مشکلی که اغلبِ الگوریتم‌های **NMF** با آن روبه‌رو هستند، عدم تضمین هم‌گرایی به نقطهٔ کمینهٔ سراسری است؛ زیرا تابع هزینه در **NMF**، هم‌زمان نسبت به دو عامل ماتریسی **W** و **H** نامحدب است؛ ازاین‌رو هم‌گرایی به نقطهٔ ایستا هدفی است که در همهٔ الگوریتم‌های **NMF** دنبال می‌شود. مقداردهی اولیهٔ ضعیف (مانند روش تصادفی) اغلب هم‌گرایی آرام و گاهی جواب‌های بی‌ربط و غلطی را نتیجه می‌دهد. اگر یک مقدار اولیه خوب بتواند به‌قدر کافی هم‌گرایی به نقطهٔ ایستا را ضمانت کند، به‌طور حتم، به‌اندازهٔ کافی از تعداد تکرارها خواهدکاست. در این پژوهش، از «مقداردهی اولیهٔ غیرمنفی تجزیهٔ مقدار مفرد دوگانه» استفاده شده‌است که مقادیر صفر با مقدار میانگین **X** پر شده‌اند. همچنین، از حل‌کنندهٔ نزول مختصات[2] (**CU**)، برای بهینه‌سازی

[1] Regularization Term

[2] Coordinate Descent solver

[DOI: 10.61186/jsdp.20.3.73]

استفاده شده‌است. تعداد تکرارها قبل از پایان زمان نیز برابر با ۵۰۰ تنظیم شده‌است.

با بررسی‌های انجام‌شده، معلوم شد که درصد ترک کارکنان در مجموعه‌دادهٔ **IBM**، برابر با ۱۶/۱ درصد و در مجموعه‌دادهٔ **hr_data**، برابر با ۲۳/۸ درصد است که نشان می‌دهد تعداد کارمندانی که سازمان را ترک کرده‌اند، نسبت به کارکنانی که مانده‌اند، خیلی کمتر است؛ بنابراین، داده‌ها نامتوازن هستند. برای متوازن‌سازی داده‌ها از روش کم نمونه‌گیری تصادفی[1] از داده‌های اکثریت و سپس بیش نمونه‌گیری تصادفی[2] از داده‌های اقلیت با نسبت ۰/۵۹۶۵ برای مجموعه‌دادهٔ **IBM** و ۰/۶۵۵۸ برای مجموعه‌دادهٔ **hr_data** استفاده شده‌است. در نهایت، مجموعه‌داده‌ها به دو قسمت آموزشی[3]، برای آموزش الگوها و آزمایشی[4]، برای ارزیابی عملکرد آن‌ها، با نسبت هفتاد به سی تقسیم شده‌اند.

۳-۳ - پردازش الگو و تعیین ابرشاخص‌ها

پس‌از انجام مراحل پیش‌پردازش و آماده‌سازی داده‌ها، در مرحلهٔ پردازش، دو الگوریتم پیشنهادی (**ExtraTrees** و **RF**) بهینه‌شده با تنظیم ابرشاخص‌ها و چهار الگوریتم مختلف یادگیری ماشین ازجمله: الگوریتم **KNN**، **DT**، طبقه‌بند بردار پشتیبان[5] و **AdaBoostClassifier**، بر روی مجموعهٔ آموزشی برازش داده و از مجموعهٔ آزمایشی برای ارزیابی دقت الگوها استفاده شد. در این مرحله، تنظیم بهینهٔ ابرشاخص‌های هر الگوریتم از اهمیت زیادی برخوردار است؛ زیرا ابرشاخص‌های تعیین‌شده برای هر الگو، در افزایش یا کاهش دقت الگوها بسیار تأثیرگذار هستند. ابرشاخص‌ها باید توسط دانشمند داده، قبل از آموزش مشخص شوند. مشکلی که در انتخاب ابرشاخص‌های مناسب وجود دارد، این است که مجموعهٔ بهینه برای هر مسئلهٔ یادگیری ماشین، متفاوت خواهدبود؛ بنابراین، تنها راه برای یافتن بهترین تنظیمات، امتحان‌کردن تعدادی از آن‌ها در هر مجموعه از داده‌های جدید است.

در این پژوهش، جهت یافتن بهترین مقدار برای ابرشاخص‌ها از روش‌های **GridSearchCV**، **RandomizedSearchCV** و همچنین، آزمایش و خطا، استفاده شده‌است. با استفاده از روش **RandomizedSearchCV**، می‌توان شبکه‌ای[6] از بازه‌هایی از مقادیر ابرشاخص‌ها را تعریف و به‌صورت تصادفی نمونه‌هایی از مقادیر این شبکه را انتخاب و ارزیابی کرد، سپس، بهترین مقدار ابرشاخص‌ها را در آن بازه مشخص کرد. در روش **GridSearchCV**، همهٔ ترکیبات شاخص‌های مشخص‌شده در شبکه، ارزیابی می‌شوند؛ بنابراین، به منابع محاسباتی بیشتری برای استفاده از این روش نیاز است. مهم‌ترین ابرشاخص‌هایی که در این جستجو به‌دست‌آمده‌اند، در فهرست زیر ذکر شده‌اند:

- در الگوریتم جنگل تصادفی مقدار بهینه برای ابرشاخص «تعداد درختان تصمیم در جنگل» (**n_estimators**)، برابر با ۱۵۰۰ به دست آمد. برای اندازه‌گیری کیفیت تقسیم در درختان، از معیار **Gini** استفاده شده‌است.
- در الگوریتم **ExtraTrees** نیز ابرشاخص **n_estimators** برابر با هزار تنظیم شد.
- به نظر می‌رسد مهم‌ترین ابرشاخص‌ها در الگوریتم **KNN** شاخص **n_neighbors** باشد که نشان‌دهندهٔ تعداد همسایه‌هایی است که برای جستارهای همسایگان استفاده می‌شوند و در این پژوهش برابر با یک منظور شد.

۴-۳- انتخاب مؤثرترین ویژگی‌ها

باتوجه‌به‌این‌که تنها پیش‌بینی ریزش کارکنان بدون یافتن عوامل مؤثر بر آن کافی نخواهدبود، بنابراین، پس‌از ساخت الگوها و ارزیابی عملکرد آن‌ها، با استفاده از روش‌های مختلف انتخاب ویژگی، مؤثرترین ویژگی‌ها در پیش‌بینی ریزش کارکنان جستجو شد. انتخاب ویژگی فرآیندی است که ویژگی‌های موثرتر را از مجموعه‌های داده برای انجام اعمال داده‌کاوی انتخاب می‌کند. روش‌های انتخاب ویژگی به‌طورکلی به سه دستهٔ پالایه[7]، پوشانه[8] و تعبیه‌شده[9] تقسیم می‌شوند. روش‌های پالایه از مشخصات آماری ویژگی‌ها استفاده و از الگوریتم یادگیری مستقل عمل می‌کنند. در مقابل، روش‌های پوشانه از الگوریتم‌های یادگیری ماشین برای انتخاب بهترین و موثرترین ویژگی‌ها استفاده می‌کنند. در این پژوهش از یک روش ترکیبی برای انتخاب بهترین ویژگی‌ها استفاده شده‌است. بدین ترتیب که نخست، از یک روش انتخاب ویژگی پالایه با عنوان **selectKBest** با آزمون آماری **chi2** و یک روش انتخاب ویژگی پوشانه با عنوان **RFE**[10]، با چهار الگوریتم یادگیری **RF**، **DT**، **ExtraTrees** و **AdaBoost**، استفاده

[1] Random Under Smapling
[2] Random Over Sampling
[3] Train
[4] Test
[5] Support Vector Classification
[6] Grid

[7] Filter
[8] Wrapper
[9] Embeded
[10] Recursive feature elimination

[DOI: 10.61186/jsdp.20.3.73]

و مؤثرترین ویژگی‌ها در هر روش مشخص شد. سپس با میانگین‌گیری از نتایج همهٔ روش‌ها، ویژگی‌هایی که در هشتاد درصد روش‌ها انتخاب شده‌بودند، به‌عنوان ویژگی‌های برتر منظور شدند. روش **RFE** از الگوریتم‌های یادگیری ماشین برای حذف کم‌اهمیت‌ترین ویژگی‌ها پس‌از آموزش بازگشتی، استفاده می‌کند تا در نهایت، تعداد ویژگی‌ها به تعداد تعیین‌شده (در این مقاله ۱۷ ویژگی) برسد.

۴- نتایج

این بخش، به بررسی نتایج حاصل از این پژوهش می‌پردازد. به‌منظور بررسی تأثیر متوازن‌سازی و همچنین کاهش ابعاد بر روی عملکرد الگوها، آزمایش‌ها در سه مرحله انجام گرفت:

۱- قبل‌از متوازن‌سازی داده‌ها

۲- بعداز متوازن‌سازی داده‌ها و قبل‌از کاهش ابعاد فضای ویژگی‌ها

سپس، نتایج حاصل از انتخاب ویژگی‌های برتر، در جدول (۵) و نتایج ارزیابی عملکرد الگو با ویژگی‌های انتخاب‌شده در جدول (۶) نشان داده شده‌است.

۴-۱ نتایج آزمایش اول

الگوها با استفاده از ابزارها و کتابخانه‌های زبان پایتون مانند **scikit-learn** پیاده‌سازی شده‌اند.

دقت پیش‌بینی قبل از متوازن‌سازی و بر اساس معیارهای ارزیابی **f1score** و دقت[1] محاسبه و نتایج مقایسه شدند. جدول (۲) عملکرد الگوریتم‌ها را قبل از متوازن‌سازی در مجموعه‌داده‌ها نشان می‌دهد.

(جدول- ۲): عملکرد الگوریتم‌ها قبل‌از متوازن‌سازی
(Table- 2): Performance of algorithms before balancing

معیار ارزیابی الگوریتم	مجموعه‌دادهٔ IBM		مجموعه‌دادهٔ hr_data	
	f1score	دقت	f1score	دقت
AdaBoost	44.07	85.03	92.09	96.25
KNN	23.7	76.64	89.53	94.8
DecisionTree	32.59	79.37	95.87	98.02
SVC	10.67	84.81	16.81	78.38
الگوی پیشنهادی ۱ EetraTrees	**20**	**85.49**	**97.6**	**98.87**
الگوی پیشنهادی ۲ RandomForest	**25.58**	**85.49**	**98.38**	**99.24**

[1] Accuracy

۴-۲ نتایج آزمایش دوم

در این مرحله، برای بررسی اثر کاهش ابعاد تجزیه ماتریس غیرمنفی روی مجموعه‌دادهٔ **IBM**، نتایج عملکرد الگوریتم‌ها بعداز متوازن‌سازی و قبل از کاهش ابعاد، در جدول (۳) نشان داده شده‌است.

(جدول- ۳): عملکرد الگوریتم‌ها در مجموعه‌دادهٔ IBM قبل‌از کاهش ابعاد و بعداز متوازن‌سازی
(Table- 3): Performance of Algorithms in IBM Dataset Before Dimension Reduction and After Balancing

معیار ارزیابی الگوریتم	f1score	دقت
AdaBoost	84.48	83.67
KNN	85.77	84.81
DecisionTree	87.10	86.17
SVC	83.48	82.77
الگوی پیشنهادی ۱ EetraTrees	**91.61**	**91.61**
الگوی پیشنهادی ۲ RandomForest	**90.91**	**90.70**

۴-۳ نتایج آزمایش سوم

در آزمایش سوم عملکرد الگوریتم‌ها بعداز متوازن‌سازی در هر دو مجموعه‌داده و پس‌از کاهش ابعاد در مجموعه‌داده **IBM**، مورد بررسی قرار گرفته‌اند. نتایج این آزمایش در جدول ۴ نشان‌داده‌شده است.

(جدول ۴): عملکرد الگوریتم‌ها بعداز متوازن‌سازی و کاهش ابعاد
(Table- 4): Performance of algorithms after balancing and dimensional reduction

معیار ارزیابی الگوریتم	مجموعه‌دادهٔ IBM		مجموعه‌دادهٔ hr_data	
	f1score	دقت	f1score	دقت
AdaBoost	83.47	82.31	93.92	93.81
KNN	88.98	87.3	95.87	95.68
DecisionTree	88.41	86.62	98.21	98.16
SVC	77.43	76.87	67.9	67.54
الگوی پیشنهادی ۱ EetraTrees	**95.82**	**95.69**	**99.18**	**99.17**
الگوی پیشنهادی ۲ RandomForest	**94.67**	**94.33**	**99.52**	**99.52**

[DOI: 10.61186/jsdp.20.3.73]

(جدول- ۵): نمایش ویژگی برتر انتخاب‌شده با روش‌های SelectKbest و RFE
(Table- 5): display of the top 5 features selected by SelectKbest and RFE methods

روش انتخاب ویژگی	ExtraTrees RFE	AdaBoost RFE	RandomForest RFE	DecisionTree RFE	SelectKbest	نتیجه (انتخاب ویژگی‌های با فراوانی انتخاب بالای هشتاد درصد)
Age	Y	Y	Y	Y	Y	Y
BusinessTravel	Y	N	N	Y	N	N
DailyRate	Y	Y	Y	Y	Y	Y
Department	N	N	N	N	N	N
DistanceFromHome	N	Y	Y	Y	Y	Y
Education	N	N	N	N	N	N
EducationField	Y	N	N	Y	N	N
EmployeeCount	N	N	N	N	N	N
EnvironmentSatisfaction	Y	Y	N	Y	Y	Y
Gender	Y	N	N	N	N	N
HourlyRate	Y	Y	Y	N	Y	Y
JobInvolvement	Y	N	Y	Y	N	N
JobLevel	N	Y	N	Y	N	N
JobRole	Y	N	N	N	N	N
JobSatisfaction	Y	N	Y	Y	Y	Y
MaritalStatus	Y	N	N	N	N	N
MonthlyIncome	Y	Y	Y	Y	Y	Y
MonthlyRate	N	Y	Y	N	Y	Y
NumCompaniesWorked	Y	Y	Y	Y	Y	Y
OverTime	Y	Y	Y	Y	Y	Y
PercentSalaryHike	Y	Y	N	N	Y	Y
PerformanceRating	N	N	N	N	N	N
RelationshipSatisfaction	N	N	N	N	N	N
StockOptionLevel	N	Y	N	Y	Y	Y
TotalWorkingYears	N	Y	Y	Y	Y	Y
TrainingTimesLastYear	Y	Y	Y	N	N	N
WorkLifeBalance	N	N	Y	N	Y	Y
YearsAtCompany	N	Y	Y	Y	Y	Y
YearsInCurrentRole	N	N	Y	Y	Y	Y
YearsSinceLastPromotion	Y	Y	N	N	N	N
YearsWithCurrManager	Y	Y	N	Y	Y	Y

۴-۴ نتایج انتخاب ویژگی

نتایج حاصل از یک روش انتخاب ویژگی صافی و چهار روش انتخاب ویژگی پوششی با الگوریتم‌های یادگیری **RF**، **DT**، **ExtraTrees** و **AdaBoost**، و ترکیب آن‌ها در جدول (۵) نمایش داده شده‌است. در ستون نتیجه، ویژگی‌هایی که در هشتاد درصد موارد یعنی در چهار مورد از روش‌های بالا انتخاب شده‌اند، به‌عنوان ویژگی‌های برتر انتخاب شده‌اند. در این جدول ویژگی‌هایی که انتخاب شده‌اند، با حرف **Y** و ویژگی‌هایی که انتخاب نشده‌اند، با حرف **N** نمایش داده‌شده‌اند. همچنین، نتایج مربوط به عملکرد الگوها، با ویژگی‌های انتخاب‌شده در «جدول ۶» نشان داده شده‌است.

۵- بحث و بررسی

این بخش، نتایج پژوهش را بررسی کرده و به نکات جدید این پژوهش می‌پردازد.

باتوجه‌به ارزیابی کمی عملکرد پیش‌بینی‌کننده‌ها قبل‌از متوازن‌سازی داده‌ها، بر اساس نتایج نشان‌داده‌شده در جدول (۲)، در مجموعه‌دادهٔ **IBM**، الگوریتم‌های یادگیری ماشین **AdaBoost** (برابر با ۴۴/۰۷) و

[DOI: 10.61186/jsdp.20.3.73]

DecisionTree (برابر با ۳۲/۵۹) و در مجموعه‌دادهٔ **hr_data**، الگوریتم‌های جنگل تصادفی (برابر با ۹۸/۳۸) و **DecisionTree** (برابر با ۹۵/۸۷)، عملکرد بهتری نسبت به سایر الگوها دارند. باتوجه‌به نامتوازن بودن داده‌ها، استفاده از معیار **f1score** برای ارزیابی عملکرد الگوریتم‌ها قبل‌از متوازن‌سازی داده‌ها، مناسب‌تر از معیار دقت است. همان‌طورکه در جدول (۲) مشاهده می‌شود، عملکرد الگوریتم‌ها با معیار **f1score** بسیار پایین‌تر از عملکرد آنها با معیار دقت است؛ به‌طور مثال، عملکرد الگوریتم **Adaboost** با معیار دقت، برابر با ۸۵/۰۳ و با معیار **f1score**، برابر با ۴۴/۰۷ است. به‌وضوح مشاهده می‌شود که بعداز متوازن‌سازی، عملکرد پیش‌بینی در تمامی الگوریتم‌ها، به طور قابل‌توجهی افزایش یافته‌است. همان‌طورکه در جدول (۴) نشان داده شده‌است، در مجموعه‌دادهٔ **hr_data**، بهترین عملکرد مربوط به الگوریتم جنگل تصادفی (برابر با ۹۹/۵۲) و الگوریتم **ExtraTrees** (برابر با ۹۹/۱۸)، و در مجموعه‌دادهٔ **IBM**، بهترین عملکرد مربوط به الگوریتم **ExtraTrees** (برابر با ۹۵/۸۲) و پس‌از آن الگوریتم جنگل تصادفی (برابر با ۹۴/۶۷) است. این نتیجه با هدف الگوریتم‌های یادگیری تجمعی که استفاده از یادگیرندگان ضعیف و ترکیب آنها برای رسیدن به نتایج بهتر است، به‌طور کامل، سازگار است. مشاهده می‌شود که عملکرد الگوریتم‌ها در مجموعه‌دادهٔ **hr_data** به‌مراتب بهتر از عملکرد آنها در مجموعه‌دادهٔ **IBM** است که این نتیجه را نیز می‌توان با اندازه مجموعه‌داده‌ها مرتبط دانست، چراکه الگوریتم‌های یادگیری تجمعی در برخورد با مجموعه‌داده‌هایی با اندازهٔ بزرگ‌تر، عملکرد بهتری از خود نشان می‌دهند. با مقایسه جدول‌های (۳) و (۴) می‌توان نتیجه گرفت که استفاده از روش کاهش ابعاد **NMF** در مجموعه‌داده **IBM**، باعث کاهش پیچیدگی الگو و افزایش عملکرد الگوریتم‌ها شده‌است.

کارایی الگوی یادگیری ماشین، ارتباط مستقیمی با ویژگی‌های دادهٔ مورداستفاده برای آموزش آن دارد. اگر ویژگی‌های دادهٔ فراهم‌شده برای الگوی **ML** نامربوط باشد، روی کارایی الگو تأثیر منفی خواهدداشت. از طرف دیگر، اگر ویژگی‌های دادهٔ استفاده‌شده مناسب باشد، دقت الگوی یادگیری ماشین می‌تواند افزایش یابد. باتوجه‌به نتایج جدول (۵) مهم‌ترین عوامل مؤثر بر ریزش کارکنان سازمان عواملی همچون «سن»، «درآمد ماهیانه»، «میزان اضافه‌کاری»، «تعداد شرکت‌هایی که کارمند پیشتر در آنجا مشغول به کار بوده است» و «میزان دستمزد روزانهٔ او»، هستند.

(جدول- ۶): عملکرد الگوریتم‌ها در مجموعه‌دادهٔ IBM بعداز انجام مهندسی ویژگی

(Table- 6): Algorithm performance in IBM dataset after feature engineering

معیار ارزیابی الگوریتم	f1score	دقت
AdaBoost	78.38	78.23
KNN	88.07	86.85
DecisionTree	86.52	84.81
SVC	81.56	80.73
الگوی پیشنهادی ۱ EetraTreesClassifier	**93.74**	**93.42**
الگوی پیشنهادی ۲ RandomForest	**92.57**	**92.06**

۶- نتیجه‌گیری

باتوجه‌به گسترش روزافزون فناوری اطلاعات و تغییر مداوم محیط‌های کاری، این فرصت برای مدیران سازمان‌ها وجود دارد که بتوانند پایگاه‌داده‌های منابع انسانی خود را گسترش داده و با استفاده روش‌های مختلف داده‌کاوی تحلیل‌ها و پیش‌بینی‌های دقیق‌تری از منابع انسانی خود داشته‌باشند. در این پژوهش ضمن بیان رویکردهای پیشین در زمینهٔ استفاده از داده‌کاوی به‌منظور پیش‌بینی ریزش کارکنان، سعی شد تا ضمن توسعه پژوهش‌های گذشته الگوهای جدید با پیش‌پردازش‌های متفاوت و ابرشاخص‌هایی با مقادیر مؤثرتر ارائه شود. اگرچه اندازهٔ مجموعه‌داده (از نظر تعداد ویژگی‌ها و تعداد نمونه‌ها)، در انتخاب نوع الگوریتم‌ها و عملکرد آنها بسیار مؤثر است، اما نتایج این پژوهش تأیید می‌کند که استفاده از الگوریتم‌های تجمعی، به‌ویژه الگوریتم‌های مبتنی‌بر درخت، بهترین گزینه برای طبقه‌بندی داده‌های منابع انسانی و پیش‌بینی ریزش کارکنان هستند. با مقایسهٔ عملکرد پیش‌بینی الگوریتم‌ها در مجموعه‌دادهٔ **IBM** (برابر با ۹۵/۸۲)، نسبت به پژوهش‌های قبل [۱-۴] و اختلاف زیادی که وجود دارد، همچنین، افزایش ۱/۵ درصدی دقت پیش‌بینی الگوریتم‌ها در مجموعه‌دادهٔ **hr_data** (برابر با ۹۹/۵۲) نسبت به پژوهش [۲] می‌توان نتیجه گرفت که نخست، متوازن‌سازی داده‌ها در داده‌های نامتوازن، در بهبود نتایج پیش‌بینی‌ها بسیار مؤثر است، دوم، با استفاده از روش‌های کاهش ابعاد مناسب، می‌توان به نتایج بهتری دست‌یافت. همچنین تنظیم بهینهٔ ابرشاخص «تعداد درختان» در الگوریتم‌های تجمعی جنگل تصادفی و

[DOI: 10.61186/jsdp.20.3.73]

ExtraTrees، بر افزایش یا کاهش دقت آنها مؤثر است. با تحلیل ویژگی‌های مؤثر بر ریزش کارکنان در این پژوهش، می‌توان نتیجه گرفت که مسائل مالی بیشترین تأثیر را بر ریزش کارکنان داشته‌اند. باتوجه‌به شباهت ویژگی‌های مجموعه‌دادهٔ **IBM** با ویژگی‌های کارکنان سازمان‌های ایران، می‌توان از نتایج این پژوهش برای پیش‌بینی ریزش کارکنان سازمان‌های داخلی و تحلیل ویژگی‌های مؤثر بر آن استفاده کرد؛ بنابراین، مدیران سازمان‌ها می‌توانند با افزایش حقوق، مزایا و اضافه‌کاری به کارکنان خود، از تحمیل هزینه‌های ناشی از ریزش کارکنان بر سازمان پیشگیری کنند. باتوجه‌به نتایج این پژوهش در جدول (۷)، نتایج این پژوهش با نتایج پژوهش‌های قبلی مقایسه شده‌است.

۷- منابع 7-References

[1] Y. Zhao, M. K. Hryniewicki, F. Cheng, B. Fu, and X. Zhu, "Employee turnover prediction with machine learning: A reliable approach," in *Proceedings of SAI intelligent systems conference*, 2018: Springer, pp. 737-758.

[2] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, "From big data to deep data to support people analytics for employee attrition prediction," *IEEE Access*, vol. 9, pp. 60447-60458, 2021.

[3] M. Pratt, M. Boudhane, and S. Cakula, "Employee Attrition Estimation Using Random Forest Algorithm," *Baltic Journal of Modern Computing*, vol. 9, no. 1, pp. 49-66, 2021.

[4] S. S. Alduayj and K. Rajpoot, "Predicting employee attrition using machine learning," in *2018 international conference on innovations in information technology (iit)*, 2018: IEEE, pp. ۹۳-۹۸.

[5] A. Huda and N. Ardi, "Predictive Analytic on Human Resource Department Data Based on Uncertain Numeric Features Classification," *Int. J. Interact. Mob. Technol.*, vol. 15, no. 8, pp. 172-181, 2021.

[6] M. Al Akasheh, E. F. Malik, O. Hujran, and N. Zaki, "A Decade of Research on Data Mining Techniques for Predicting Employee Turnover: A Systematic Literature Review," *Available at SSRN 4401862*.

[7] P. Ajit, "Prediction of employee turnover in organizations using machine learning algorithms," *algorithms*, vol. 4, no. 5, p. C5, 2016.

[۸] ا. م. اسماعیلی سیکارودی، ر. قوسی، ع. اسماعیلی سیکارودی، «یک رویکرد داده‌کاوی برای پیش‌بینی ریزش کارکنان. مطالعه موردی: تولید قطعات خودروی اراک»، مجلهٔ سیستم‌ها و صنعت، ۲۰۱۵.

[8] A. M. Esmaieeli Sikaroudi, R. Ghousi, and A. Sikaroudi, "*A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing*", Journal of industrial and systems engineering, vol. 8, no. 4, pp. 106-121, 2015.

[۹] س. حسن‌خانی دولت‌آبادی، ف. کی‌نیا، «طراحی مدل پیش‌بینی ریزش مشتری و کارمند بر اساس روش داده‌کاوی و پیش‌بینی‌کننده عصبی»، IEEE Xplore، ۲۰۱۷.

[9] S. H. Dolatabadi and F. Keynia, "*Designing of customer and employee churn prediction model based on data mining method and neural predictor,"* in *2017 2nd International Conference on Computer and Communication Systems (ICCCS)*, 2017: IEEE, pp. 74-77.

[۱۰] س. خضیری عفراوی، م. سرداری زارچی، س. م. م. فاطمی بوشهری، «عوامل مؤثر بر تمایل به ترک سازمان با استفاده از الگوریتم‌های مبتنی‌بر شبکهٔ عصبی و ژنتیک چندهدفه»، مدیریت منابع انسانی در صنعت نفت، ۱۳۹۷.

[10] S. Khaziri Afravi, M. Sardari Zarchi, S. M. M. Fatemi Bushehri, "*Factors affecting the tendency to leave the organization using algorithms based on multi-objective neural network and genetics*", Human Resource Management in the Oil Industry, 1397.

[11] X. Cai *et al.*, "DBGE: employee turnover prediction based on dynamic bipartite graph embedding," *IEEE Access*, vol. 8, pp. 10390-10402, 2020.

[12] P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees' attrition: a machine learning approach," *SN Applied Sciences*, vol. 2, pp. ۱-۱۱, ۲۰۲۰.

[13] M. Lazzari, J. M. Alvarez, and S. Ruggieri, "Predicting and explaining employee turnover intention," *International Journal of Data Science and Analytics*, vol. 14, no. 3, pp. 279-292, 2022.

[14] X. Gao, J. Wen, and C. Zhang, "An improved random forest algorithm for predicting employee turnover," *Mathematical Problems in Engineering*, vol. 2019, 2019.

[15] M. Teng, H. Zhu, C. Liu, and H. Xiong, "Exploiting network fusion for organizational turnover prediction," *ACM Transactions on Management Information Systems (TMIS)*, vol. 12, no. 2, pp. 1-18, 2021.

[16] N. Jain, A. Tomar, and P. K. Jana, "A novel scheme for employee churn problem using multi-attribute decision making approach and machine learning," *Journal of Intelligent Information Systems*, vol. 56, pp. 279-302, 2021.

[DOI: 10.61186/jsdp.20.3.73]

سیده محبوبه مزارعی، دانش‌نامهٔ کارشناسی خود را در رشته ریاضی کاربردی در رایانه، از دانشگاه صنعتی امیرکبیر در سال ۱۳۸۲ دریافت کرده‌است. همچنین تحصیلات کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات گرایش معماری سازمانی در دانشگاه پیام‌نور بوشهر، مرکز بین‌الملل عسلویه در سال ۱۴۰۱ به پایان رسانیده‌است. موضوع پایان‌نامه ایشان، پیش‌بینی ریزش کارکنان سازمان با استفاده از داده‌کاوی، بوده‌است. موضوعات موردعلاقهٔ ایشان در زمینهٔ هوش مصنوعی است.

نشانی رایانامهٔ ایشان عبارت است از:

s.m.mazarei60@gmail.com

جعفر پورامینی، دانش‌نامهٔ کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی نرم‌افزار در سال‌های ۱۳۷۸ و تحصیلات خود را در مقطع دکتری در رشتهٔ مهندسی فناوری اطلاعات در سال ۱۳۹۷ به پایان رساند و هم‌اکنون استادیار دانشگاه پیام‌نور استان قم است. موضوعات موردعلاقهٔ ایشان، زمینهٔ داده‌کاوی، یادگیری ماشین و یادگیری تقویتی است.

نشانی رایانامهٔ ایشان عبارت است از:

J_pouramini@pnu.ac.ir

[DOI: 10.61186/jsdp.20.3.73]