

پیش‌بینی ریزش کارمندان با استفاده از الگوریتم‌های یادگیری گروهی مبتنی بر درخت تصمیم

سیده محبوبه مزارعی^{۱*}، جعفر پورامینی^۲

^۱ مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور مرکز بین‌الملل عسلویه، ایران

^۲ 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 به‌دست آمده‌است؛ بنابراین، پیش‌بینی ریزش کارکنان به مدیران منابع انسانی در استخدام نیروهای ماندگار و حفظ و نگهداری آنها کمک می‌کند. یکی از ابزارهای کارآمد در این خصوص استفاده از روش‌های مختلف داده‌کاوی است. تعداد کم نمونه‌ها و نامتوازن بودن داده‌های ریزش کارکنان و تنظیم ابرشاخص‌ها از جمله مشکلات استفاده از روش‌های داده‌کاوی برای پیش‌بینی ریزش کارکنان است. هدف این پژوهش، ارائه روش‌های مناسب کاهش ویژگی و پیش‌پردازش داده‌ها به همراه ارائه راهکار برای تنظیم مناسب ابرشاخص‌ها برای پیش‌بینی ریزش کارکنان با استفاده از روش‌های مختلف داده‌کاوی و الگوریتم‌های یادگیری ماشین تجمعی است. با توجه به نامتوازن بودن داده‌ها، از روش‌های کم نمونه‌گیری تصادفی و ترکیب آن با بیش نمونه‌گیری تصادفی برای متوازن‌سازی داده‌ها با نسبت‌های متفاوت استفاده شد. با توجه به مثبت بودن همه داده‌ها از روش کاهش ابعاد تجزیه ماتریس نامنفی (NMF) استفاده و با استفاده از روش‌های جستجوی ابرشاخص‌ها، مقادیر بهینه آن‌ها برای الگوریتم‌های پیشنهادی، تعیین شد. برای ارزیابی روش پیشنهادی از مجموعه داده‌های استاندارد با اندازه‌های مختلف استفاده شده‌است. نتایج روش پیشنهادی نیز با نتایج حاصل از سایر روش‌های مطرح در این حوزه مانند KNN، AdaBoost، DT و SVC مقایسه شده‌است. نتایج به‌دست آمده نشان می‌دهند که الگوی پیشنهادی این پژوهش نسبت به پژوهش‌هایی که در گذشته روی همین داده‌ها صورت گرفته، دارای دقت پیش‌بینی بهتری است. طبق بررسی‌های انجام شده 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 mazareei^{1*}, Jafar pooramini²

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

²Computer 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 reviews 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

* Corresponding author

* نویسنده عهده‌دار مکاتبات

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

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 تجربه آنها برای موفقیت سازمان مهم است. در

۱- مقدمه

تجزیه و تحلیل منابع انسانی نقش مهمی در هر جنبه‌ای از عملکرد منابع انسانی در سازمان‌ها، از جمله استخدام،

فصل ۱



این زمینه، ریزش کارکنان توجه بسیاری از کارشناسان منابع انسانی را به خود جلب کرده‌است، زیرا به‌طور گسترده به‌عنوان یک شاخص مهم رقابت‌پذیری در نظر گرفته می‌شود [۱]. ریزش کارکنان به فرآیند ازدست‌دادن کارمندان و جایگزینی آنها با استخدام‌های جدید اشاره دارد [۲]، که سازمان‌ها را در سراسر صنایع تحت تأثیر 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 سرمایه انسانی، استخدام، آموزش و توسعه کارکنان جدید متحمل می‌شود [۵]. بنابراین، درک علل ریزش کارکنان و توسعه 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 رقابت‌درازمدت بهره‌مند می‌شوند. هدف از انجام این پژوهش، استفاده از روش‌های داده‌کاوی و مقایسه 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 در نظر بگیرند. 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 هر یک وظایف مختلفی دارند: فهم تجاری، فهم داده، آماده‌سازی داده، الگوسازی، ارزیابی و پیاده‌سازی.

در این مقاله، از الگوریتم‌های یادگیری ماشین^۱ مختلف، برای پیش‌بینی ریزش کارکنان، استفاده شد. به‌منظور بهبود دقت پیش‌بینی، پیش‌پردازش‌های متفاوتی بر روی مجموعه داده‌های استاندارد، انجام شده‌است. نتایج به‌دست‌آمده نشان می‌دهد که روش پیشنهادی این پژوهش که استفاده از الگوریتم‌های بهینه‌شده یادگیری جمعی^۲ جنگل تصادفی (RF)^۳ و **ExtraTrees** است، به‌همراه استفاده از روش کاهش ابعاد «تجزیه ماتریس نامفی»^۴ (NMF)، متوازن‌سازی داده‌ها و تنظیم صحیح ابرشاخص‌ها^۵، باعث افزایش دقت پیش‌بینی می‌شود. جنگل تصادفی یک روش یادگیری جمعی مبتنی بر درخت است که در آن درختان زیادی با استفاده از زیرمجموعه‌های کوچکی از داده‌ها (با جایگزینی) که بوت‌استرپ^۶ نیز نامیده می‌شوند، ساخته می‌شوند. در نهایت، از رأی اکثریت برای تعیین پیش‌بینی الگو استفاده می‌شود. این مفهوم به‌عنوان **Bagging**^۷ 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^۸ مقاوم می‌کند [۲].

الگوریتم **ExtraTrees** نوعی روش یادگیری جمعی است که تعدادی درخت تصمیم تصادفی شده (معروف به درخت‌های اضافی) را در زیرنمونه‌های مختلف مجموعه داده برازش می‌دهد و از میانگین‌گیری برای بهبود دقت پیش‌بینی و مدیریت 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. با این حال، پس از

¹ Machine Learning

² Ensembl Learning

³ Random Forest

⁴ Non-negative matrix actorization

⁵ Hyper Parameters

⁶ Bootstrap

⁷ Bootstrap Aggregation

⁸ Over-fitting

انتخاب نقاط تقسیم، این دو الگوریتم بهترین را از بین همه زیرمجموعه ویژگی‌ها انتخاب می‌کنند.

این مقاله به شرح زیر تنظیم شده است: در بخش دو، پیشینه پژوهش مطالعه شده است. در بخش سه، روش پیشنهادی بررسی و ارزیابی شده و مهم‌ترین ابرشاخص‌ها تعیین شده‌اند. در بخش چهار نتایج پژوهش نشان داده شده است. بخش پنج به بحث و بررسی نتایج این پژوهش و مقایسه آنها با نتایج به دست آمده در مقالات قبل پرداخته و نتیجه‌گیری از پژوهش در بخش آخر آمده است.

۲- پیشینه پژوهش

در زمینه مدیریت منابع انسانی و روش‌های پیش‌بینی ریزش کارکنان، پژوهش‌های داخلی و خارجی فراوانی صورت گرفته و از روش‌های مختلفی برای این کار استفاده شده است. یکی از این روش‌ها استفاده از روش‌های مختلف داده‌کاوی و الگوریتم‌های یادگیری ماشین است که پیشتر توسط پژوهشگران مختلف مطالعه شده است. از جمله، نویسندگان مقاله [۷] قابلیت‌های پیش‌بینی هفت الگوریتم مختلف یادگیری ماشین، از جمله الگوریتم‌های به تازگی توسعه یافته XGB^1 را در مورد ریزش کارمندان مقایسه کردند. به طور 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 قاعده‌القایی، انجام دادند.

دولت‌آبادی و کی‌نیا [۹]، چهار الگوی پیش‌بینی ریزش مشتری و کارمند، از جمله درخت تصمیم $(DT)^2$ ، بیز ساده $(NB)^3$ ، ماشین بردار پشتیبان $(SVM)^4$ و شبکه عصبی $(NN)^5$ را بررسی کردند. الگوریتم SVM ، به بالاترین دقت پیش‌بینی $(99/83)$ دست یافت. پس از SVM ، روش NB با $92/37$ درصد بالاترین دقت پیش‌بینی را داشت.

محسن سرداری زارچی و همکارانش [۱۰] در سال ۱۳۹۷، در مقاله‌ای، با استفاده از مجموعه داده به دست آمده از نظرات کارکنان شرکت بهره‌برداری نفت و گاز کارون، مبنی بر رضایت‌مندی و تمایل به ترک سازمان، با استفاده از شبکه‌های عصبی مصنوعی به عنوان طبقه‌بند و الگوریتم

تکاملی ژنتیک چندمنظوره برای انتخاب ویژگی‌های مؤثر، یک سامانه خبره با دقت $0/88$ طراحی کردند.

مادارا پرت^۶ 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 DT ، RF ، رگرسیون لجستیک^۷ (LR)، بیز ساده گوسی (GNB)^۸، k -نزدیکترین همسایه (KNN)^۹ و SVM ، برای پیش‌بینی ریزش کارکنان استفاده کرده‌اند. بالاترین دقت پیش‌بینی در این پژوهش، مربوط به RF و برابر با $0/85$ است.

نویسندگان در مقاله [۲]، یک رویکرد تجزیه و تحلیل افراد را برای پیش‌بینی ریزش کارکنان پیشنهاد کرده‌اند، که با تمرکز بر کیفیت داده‌ها به جای کمیت، از یک داده حجیم به یک بافت داده 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 مرحله پیش‌پردازش، از روش رمزگذاری $One-hot$ ، برای تبدیل متغیرهای غیرعددی به متغیرهای عددی استفاده شده است. برای پیش‌بینی ریزش کارکنان از الگوهای یادگیری ماشین DT ، LR و SVM ، یادگیری تجمعی RF ، XGB ^{۱۰}، $VotingClassifier$ بر اساس رأی‌گیری اکثریت^{۱۱} و یک الگوی پشته‌ای مبتنی بر شبکه عصبی شامل شبکه‌های DNN^{12} ، CNN^{13} و $LSTM^{14}$ استفاده شده است. الگوها را با استفاده از مجموعه داده‌های شبیه‌سازی شده منابع انسانی با اندازه بزرگ (hr_data)، متوسط (IBM) و سپس یک مجموعه داده با اندازه کوچک واقعی با ۴۵۰ نمونه، آزمایش شدند و در نهایت، به ترتیب به دقت‌های $(0/96)$ ، $(0/99)$ و برای سه مجموعه داده 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.

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

¹ Extreme Gradient Boosting

² Decision Tree

³ Naïve Bayes

⁴ Support Vector Machines

⁵ Neural Network

نویسنده مقاله [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 سامانه 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^۱ (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 به‌وسیله DBGE ریزش کارکنان را بسیار دقیق‌تر پیش‌بینی می‌کنند.

نویسندگان در مقاله [۱]، از دو مجموعه داده متفاوت برای پیش‌بینی ریزش کارکنان استفاده کرده‌اند. نخستین مجموعه داده شامل اطلاعات منابع انسانی یک بانک منطقه‌ای در ایالات متحده آمریکا است که از سال ۲۰۱۳ تا ۲۰۱۶ جمع‌آوری شده است و شامل ۱۴۳۲۲ ورودی کارمند و ۲۴ ویژگی است. مجموعه داده دوم یک مجموعه داده شبیه‌سازی شده است که توسط شرکت IBM ایجاد شده و برای تسهیل تجزیه و تحلیل دقیق‌تر در این پژوهش گنجانده شده است. برای آماده‌سازی داده‌ها، همه 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 میانه آن ویژگی برای همه ورودی‌ها، جایگزین کرده‌اند. داده‌های با مقادیر غیرعددی را با روش رمزگذاری برچسب^۲، به مقادیر عددی تبدیل کرده‌اند. سپس، عملکرد ده الگوریتم مختلف یادگیری ماشین مانند KNN، DT، RF، GBT، XGB، LR، SVM، NN، LDA^۳ و NB را، بر روی ده زیرمجموعه کوچک، متوسط و بزرگ از مجموعه داده‌های اولیه که به‌طور تصادفی و بدون جای‌گذاری، انتخاب کرده‌اند، با استفاده از معیار ROC^۴ ارزیابی کردند. در این مقاله، دقت الگوریتم‌ها به‌طور میانگین، در مجموعه داده‌های کوچک برابر ۰/۸۳، در مجموعه داده‌های متوسط برابر با ۰/۸۱ و در مجموعه داده‌های بزرگ برابر با ۰/۸۶ به‌دست آمده است.

جین و همکاران [۱۲] ریزش کارکنان را با استفاده از الگوهای یادگیری ماشین SVM، DT و RF و مجموعه داده‌های سامانه اطلاعات منابع انسانی با بیش از چهارده هزار رکورد و ده ویژگی، پیش‌بینی کرد. بر اساس یافته‌های این پژوهش، عملکرد RF از سایر طبقه‌بندی‌کننده‌ها بهتر و دقیق‌تر بود. نویسندگان مقاله [۱۳] از مجموعه داده‌های نظرسنجی بزرگ و واقعی از سراسر اروپا برای شناسایی بهترین روش‌های یادگیری ماشین برای پیش‌بینی ریزش کارکنان استفاده کردند. KNN، DT، LR، RF، 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 مقاله [۴]، از سه آزمایش اصلی بر روی مجموعه داده شبیه‌سازی شده IBM Watson، برای پیش‌بینی ریزش کارکنان استفاده کرده‌اند. اولین آزمایش شامل آموزش مجموعه داده اصلی رده نامتوازن با الگوهای یادگیری ماشینی SVM با چند تابع هسته^۶، RF و KNN بود. در آزمایش دوم بر روی استفاده از رویکرد مصنوعی تطبیقی^۷ (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 مجموعه داده متوازن‌شده با روش ADASYN، برابر با ۰/۹۳ بود. همچنین، دقت الگوریتم RF بعد از انتخاب ۱۲ ویژگی از کل ۲۹ ویژگی، برابر با ۰/۹۱ به‌دست آمده است. در جدول (۱) مقایسه‌ای بین پژوهش‌های ارائه‌شده در این بخش نمایش داده شده است.

در این پژوهش، به مقالاتی که به‌تازگی منتشر شده و مبتنی بر الگوهای یادگیری ماشینی و عمیق هستند، و از مجموعه داده‌های منابع انسانی شبیه‌سازی شده Kaggle استفاده کرده‌اند [۴-۱]، استناد شده است. دلیل این انتخاب، وجود نتایج آزمایش‌های مربوط به دقت الگوهای پیش‌بینی برای این مجموعه داده‌های 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

ویژگی متفاوت، میانگین‌گیری و ویژگی‌های برتر انتخاب شد. در ادامه هر یک از موارد ذکر شده با جزئیات تشریح می‌شود.

۱-۳ مجموعه داده

باتوجه به این که پایگاه‌های داده 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 مجموعه داده شبیه‌سازی شده استاندارد که از پایگاه **Kaggle.com** استخراج شده‌اند، استفاده شده‌است. مجموعه داده منابع انسانی شرکت **IBM (HR IBM)**، شامل ۱۴۷۰ نمونه با ۳۴ ویژگی ورودی و مجموعه داده **hr_data**، یک مجموعه داده شبیه‌سازی شده است که شامل پانزده هزار نمونه با ده ویژگی ورودی است. متغیرهای مستقل آن‌ها در جدول (۱) نمایش داده شده‌اند. متغیر هدف، ترک خدمت کارکنان (**Left**) است.

۲-۳ آماده‌سازی و پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها به‌طور معمول، در مطالعات پیش‌بینی ریزش کارکنان انجام می‌شود، زیرا مجموعه داده‌ها اغلب حاوی ورودی‌های گم‌شده^۲، درجات مختلف نوفه و تفاوت‌های اساسی در مقیاس^۳ 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 مجموعه داده **hr_data**، مقدار ویژگی سیگاری بودن (**is_smoker**) برای بیشتر نمونه‌ها ثبت نشده‌است؛ بنابراین، این ویژگی، به دلیل فراوانی داده‌های گم‌شده و نیز تمامی نمونه‌هایی که دارای داده گم‌شده بودند، حذف شدند. با بررسی مجموعه داده **IBM**، مشخص شد که این مجموعه داده، داده گم‌شده‌ای ندارد. در مرحله بعد، به این دلیل که بعضی از الگوریتم‌ها، قادر نیستند به‌طور مستقیم داده‌های غیرعددی را پردازش کنند، در هر دو مجموعه داده، مقادیر و ویژگی‌های غیرعددی، مانند وضعیت تأهل (**MaritalStatus**) به مقادیر عددی تبدیل شدند. در مجموعه داده‌های مورد مطالعه در این پژوهش، ویژگی‌ها مقیاس‌های متفاوتی دارند. به‌طور مثال، مقدار ویژگی **MonthlyIncome** بین ۱۰۰۹ تا ۱۹۹۹۹ است، در حالی که مقدار ویژگی **Age** بین ۱۸ تا ۶۰ تغییر می‌کند؛

^۲ Missing Value

^۳ Scale

• 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 تجزیه ماتریس غیرمنفی، برای پیش‌گیری از پیچیدگی الگوها و افزایش عملکرد آن‌ها. از «مقداردهی اولیه غیرمنفی تجزیه مقدار مفرد دوگانه»^۱ (**NNDSVD**) استفاده شده‌است که مقادیر صفر با مقدار میانگین **X** پر شده‌اند.

استفاده از الگوریتم‌های تجمعی مبتنی بر درخت بهینه‌شده به‌وسیله 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.

باتوجه به اهمیت پیش‌پردازش در کارایی الگو، در مرحله 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 بازه [۰، ۱]، و کاهش ابعاد مجموعه داده با روش تجزیه ماتریس غیرمنفی پیشنهاد شده‌است.

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

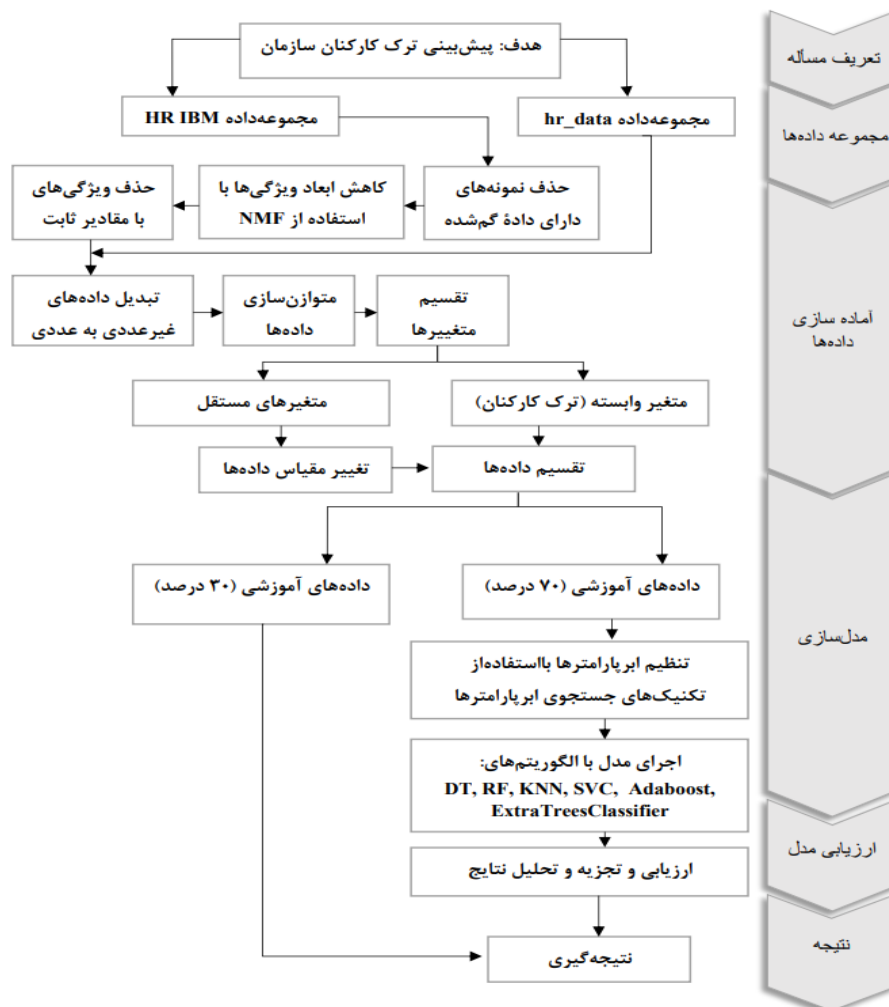
در نهایت، نتایج به‌دست‌آمده از ارزیابی الگو، با نتایج پیش‌بینی چهار الگوریتم یادگیری **DT**، **KNN**، **SVC** و **AdaBoost** مقایسه شدند.

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

^۱ Nonnegative Double Singular Value Decomposition

یادگیری ماشین به داده مقیاس‌بندی شده نیاز دارند.

بنابراین، نمی‌توان چنین داده‌هایی را به الگوریتم‌های یادگیری ماشین داد؛ چراکه بعضی از الگوریتم‌های



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

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

ریاضی هم‌ارز هستند. اگر $X \in \mathbb{R}^{m \times n}$ ، ماتریس ویژگی‌ها باشد (m تعداد مشاهدات و n تعداد ویژگی‌ها است)، NMF ماتریس‌های $W \in \mathbb{R}^{m \times r}$ و $H \in \mathbb{R}^{r \times n}$ را با شرط $r \ll \min\{m, n\}$ چنان می‌یابد که رابطه تقریبی $X \approx WH$ برقرار باشد. معادله تقریبی آن به صورت مسئله بهینه‌سازی که در رابطه (۱) نشان داده شده، تبدیل می‌شود.

پیدا کردن چنین تقریبی، نیازمند تابع هزینه‌ای است که کیفیت تقریب را به خوبی نشان دهد. یکی از این تابع‌ها می‌تواند اندازه فاصله دو ماتریس نامنفی از یکدیگر باشد. در این پژوهش، تابع هزینه بر مبنای روش فروبنیوس^۲ پایه‌ریزی شده است.

در این پژوهش‌ها، برای تنظیم دامنه ویژگی‌ها، داده‌ها در بازه $[0, 1]$ با روش $\min\max$ عادی‌سازی شدند. از آنجاکه تعداد ویژگی‌های مجموعه داده IBM زیاد است، نخست، تمامی ویژگی‌های با مقادیر ثابت که تأثیری در نتیجه پیش‌بینی الگوریتم‌ها ندارند، مانند `EmployeeCount` و `over18 standard hourse` از مجموعه داده IBM حذف شدند. سپس، باتوجه به این که همه داده‌ها دارای مقادیر مثبت هستند، از روش کاهش ابعاد «تجزیه ماتریس غیرمنفی»^۱ (NMF)، برای جلوگیری از پیچیدگی الگو استفاده شده و ابعاد فضای ویژگی‌ها به ۱۷ کاهش یافت. به طور کلی، تجزیه ماتریسی ابزاری برای تحلیل داده‌هاست. هر تجزیه تعبیرهای مختلفی را از ساختار ضمنی داده‌ها آشکار می‌سازد، که البته این تعبیرها از نظر

² Frobenius

¹ Non-negative matrix actorization

(جدول - ۱): ویژگی‌های مستقل در مجموعه داده IBM و hr_data
(Table- 1): Independent properties in the hr_data and IBM datasets

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

در NMF دو مبحث مقداردهی اولیه و شرط توقف، در دستیابی به عوامل ماتریسی بهین و هم‌گرایی نقش به‌سزایی دارند. مشکلی که اغلب الگوریتم‌های NMF با آن روبه‌رو هستند، عدم تضمین هم‌گرایی به نقطه کمینه سراسری است؛ زیرا تابع هزینه در NMF، هم‌زمان نسبت به دو عامل ماتریسی W و H نامحدب است؛ از این‌رو هم‌گرایی به نقطه ایستا هدفی است که در همه الگوریتم‌های NMF دنبال می‌شود. مقداردهی اولیه 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 نقطه ایستا را ضمانت کند، به‌طور حتم، به‌اندازه کافی از تعداد تکرارها خواهدکاست. در این پژوهش، از «مقداردهی اولیه غیرمنفی تجزیه مقدار مفرد دوگانه» استفاده شده‌است که مقادیر صفر با مقدار میانگین X پر شده‌اند. همچنین، از حل‌کننده نزول مختصات^۲ (CU)، برای بهینه‌سازی

(رابطه - ۱)

$$\min f(W, H) \cong \frac{1}{2} * \|X - WH\|_{Fro}^2 + \alpha_W * l1_{ratio} * n_{features} * \|vec(W)\|_1 + \alpha_H * l1_{ratio} * n_{samples} * \|vec(H)\|_1 + \frac{1}{2} * \alpha_W * (1 - l1_{ratio}) * n_{features} * \|W\|_{Fro}^2 + \frac{1}{2} * \alpha_H * (1 - l1_{ratio}) * n_{samples} * \|H\|_{Fro}^2, \\ W \geq 0, H \geq 0$$

در رابطه بالا، $\|W\|_{Fro}^2$ نرم فروبنیوس است که در رابطه ۲ و $\|vec(W)\|_1$ ضرب درایه‌ای هم L1 است که در رابطه (۳) نشان داده شده‌است.

$$\|W\|_{Fro}^2 = \sum_{i,j} w_{ij}^2 \quad (\text{رابطه - ۲})$$

$$\|vec(W)\|_1 = \sum_{i,j} abs(w_{ij}) \quad (\text{رابطه - ۳})$$

α_W ، α_H و $l1_{ratio}$ مقادیری ثابت هستند که برای چندبرابر کردن میزان هم‌سازی^۱ W و H به کار می‌رود که به‌منظور حذف هم‌سازی، مقادیر آنها برابر با صفر قرار داده شده‌است.

Regularization Term

^۲ Coordinate Descent solver

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

با بررسی‌های انجام‌شده، معلوم شد که درصد ترک کارکنان در مجموعه‌داده IBM، برابر با ۱۶/۱ درصد و در مجموعه‌داده 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^۱ از داده‌های اکثریت و سپس بیش نمونه‌گیری تصادفی^۲ از داده‌های اقلیت با نسبت ۰/۵۹۶۵ برای مجموعه‌داده IBM و ۰/۶۵۵۸ برای مجموعه‌داده hr_data استفاده شده‌است. در نهایت، مجموعه‌داده‌ها به دو قسمت آموزشی^۳، برای آموزش الگوها و آزمایشی^۴، برای ارزیابی عملکرد آن‌ها، با نسبت هفتاد به سی تقسیم شده‌اند.

۳-۳ - پردازش الگو و تعیین ابرشاخص‌ها

پس از انجام مراحل پیش‌پردازش و آماده‌سازی داده‌ها، در مرحله پردازش، دو الگوریتم پیشنهادی (ExtraTrees و RF) بهینه‌شده با تنظیم ابرشاخص‌ها و چهار الگوریتم مختلف یادگیری ماشین از جمله: الگوریتم DT، KNN، طبقه‌بند بردار پشتیبان^۵ و AdaBoostClassifier بر روی مجموعه آموزشی برازش داده و از مجموعه آزمایشی برای ارزیابی دقت الگوها استفاده شد. در این مرحله، تنظیم بهینه 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 مجموعه بهینه برای هر مسئله یادگیری ماشین، متفاوت خواهد بود؛ بنابراین، تنها راه برای یافتن بهترین تنظیمات، امتحان کردن تعدادی از آن‌ها در هر مجموعه از داده‌های جدید است.

در این پژوهش، جهت یافتن بهترین مقدار برای ابرشاخص‌ها از روش‌های GridSearchCV و RandomizedSearchCV و همچنین، آزمایش و خطا، استفاده شده‌است. با استفاده از روش RandomizedSearchCV، می‌توان شبکه‌ای^۶ 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، همه 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 به‌دست آمده‌اند، 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 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 نشان‌دهنده تعداد همسایه‌هایی است که برای جستارهای همسایگان استفاده می‌شوند و در این پژوهش برابر با یک منظور شد.

۳-۴ - انتخاب مؤثرترین ویژگی‌ها

باتوجه‌به این‌که تنها پیش‌بینی ریزش کارکنان بدون یافتن عوامل مؤثر بر آن کافی نخواهد بود، بنابراین، پس از 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 مؤثرتر را از مجموعه‌های داده برای انجام اعمال داده‌کاوی انتخاب می‌کند. روش‌های انتخاب ویژگی به‌طور کلی به سه دسته پالایه^۷، پوشانه^۸ و تعبیه‌شده^۹ 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 مؤثرترین 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^{۱۰} با چهار الگوریتم یادگیری RF، DT، ExtraTrees و AdaBoost، استفاده

⁷ Filter

⁸ Wrapper

⁹ Embedded

¹⁰ Recursive feature elimination

¹ Random Under Sampling

² Random Over Sampling

³ Train

⁴ Test

⁵ Support Vector Classification

⁶ Grid

۲-۴ نتایج آزمایش دوم

در این مرحله، برای بررسی اثر کاهش ابعاد تجزیه ماتریس غیرمنفی روی مجموعه داده IBM، نتایج عملکرد الگوریتم‌ها بعد از متوازن سازی و قبل از کاهش ابعاد، در جدول (۳) نشان داده شده است.

(جدول - ۳): عملکرد الگوریتم‌ها در مجموعه داده IBM قبل از

کاهش ابعاد و بعد از متوازن سازی

(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

۳-۴ نتایج آزمایش سوم

در آزمایش سوم عملکرد الگوریتم‌ها بعد از متوازن سازی در هر دو مجموعه داده و پس از کاهش ابعاد در مجموعه داده IBM، مورد بررسی قرار گرفته اند. نتایج این آزمایش در جدول ۴ نشان داده شده است.

(جدول ۴): عملکرد الگوریتم‌ها بعد از متوازن سازی

و کاهش ابعاد

(Table- 4): Performance of algorithms after balancing and dimensional reduction

معیار ارزیابی الگوریتم	مجموعه داده IBM		مجموعه داده 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

و مؤثرترین ویژگی‌ها در هر روش مشخص شد. سپس با میانگین گیری از نتایج همه روش‌ها، ویژگی‌هایی که در هشتاد درصد روش‌ها انتخاب شده بودند، به عنوان ویژگی‌های برتر منظور شدند. روش RFE از الگوریتم‌های یادگیری ماشین برای حذف کم اهمیت ترین ویژگی‌ها پس از آموزش بازگشتی، استفاده می کند تا در نهایت، تعداد ویژگی‌ها به تعداد تعیین شده (در این مقاله ۱۷ ویژگی) برسد.

۴- نتایج

این بخش، به بررسی نتایج حاصل از این پژوهش می پردازد. به منظور بررسی تأثیر متوازن سازی و همچنین کاهش ابعاد بر روی عملکرد الگوها، آزمایش‌ها در سه مرحله انجام گرفت:

۱- قبل از متوازن سازی داده‌ها

۲- بعد از متوازن سازی داده‌ها و قبل از کاهش ابعاد فضای ویژگی‌ها

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

۱-۴ نتایج آزمایش اول

الگوها با استفاده از ابزارها و کتابخانه‌های زبان پایتون مانند scikit-learn پیاده سازی شده اند.

دقت پیش بینی قبل از متوازن سازی و بر اساس معیارهای ارزیابی f1score و دقت^۱ محاسبه و نتایج مقایسه شدند. جدول (۲) عملکرد الگوریتم‌ها را قبل از متوازن سازی در مجموعه داده‌ها نشان می دهد.

(جدول - ۲): عملکرد الگوریتم‌ها قبل از متوازن سازی

(Table- 2): Performance of algorithms before balancing

معیار ارزیابی الگوریتم	مجموعه داده IBM		مجموعه داده 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

(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

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

۴-۴ نتایج انتخاب ویژگی

نتایج حاصل از یک روش انتخاب ویژگی صافی و چهار روش انتخاب ویژگی پوششی با الگوریتم‌های یادگیری RF, ExtraTrees, DT 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 داده شده‌اند. 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.

باتوجه به ارزیابی کمی عملکرد پیش‌بینی‌کننده‌ها قبل از متوازن‌سازی داده‌ها، بر اساس نتایج نشان داده شده در جدول (۲)، در مجموعه داده IBM، الگوریتم‌های یادگیری ماشین AdaBoost (برابر با ۴۴/۰۷) و

آنجا مشغول به کار بوده است» و «میزان دستمزد روزانه او»، هستند.

(جدول - ۶): عملکرد الگوریتم‌ها در مجموعه داده IBM بعد از 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 ExtraTreesClassifier	<u>93.74</u>	<u>93.42</u>
الگوی پیشنهادی ۲ RandomForest	92.57	92.06

۶- نتیجه گیری

باتوجه به گسترش روزافزون فناوری اطلاعات و تغییر مداوم محیط‌های کاری، این فرصت برای مدیران سازمان‌ها وجود دارد که بتوانند پایگاه داده‌های منابع انسانی خود را گسترش داده و با استفاده روش‌های مختلف داده کاوی تحلیل‌ها و پیش‌بینی‌های دقیق‌تری از منابع انسانی خود داشته باشند. در این پژوهش ضمن بیان رویکردهای پیشین در زمینه استفاده از داده کاوی به منظور پیش‌بینی ریزش کارکنان، سعی شد تا ضمن توسعه پژوهش‌های گذشته الگوهای جدید با پیش‌پردازش‌های متفاوت و ابرشاخص‌هایی با مقادیر مؤثرتر ارائه شود. اگرچه اندازه مجموعه داده (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 مبتنی بر درخت، بهترین گزینه برای طبقه‌بندی داده‌های منابع انسانی و پیش‌بینی ریزش کارکنان هستند. با مقایسه عملکرد پیش‌بینی الگوریتم‌ها در مجموعه داده 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 مجموعه داده 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 دست یافت. همچنین تنظیم بهینه ابرشاخص «تعداد درختان» در الگوریتم‌های تجمعی جنگل تصادفی و

DecisionTree (برابر با ۳۲/۵۹) و در مجموعه داده hr_data، الگوریتم‌های جنگل تصادفی (برابر با ۹۸/۳۸) و DecisionTree (برابر با ۹۵/۸۷)، عملکرد بهتری نسبت به سایر الگوها دارند. باتوجه به نامتوازن بودن داده‌ها، استفاده از معیار f1score برای ارزیابی عملکرد الگوریتم‌ها قبل از متوازن‌سازی داده‌ها، مناسب‌تر از معیار دقت است. همان‌طور که 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 بعد از متوازن‌سازی، عملکرد پیش‌بینی در تمامی الگوریتم‌ها، به‌طور قابل‌توجهی افزایش یافته است. همان‌طور که در جدول (۴) نشان داده شده است، در مجموعه داده hr_data، بهترین عملکرد مربوط به الگوریتم جنگل تصادفی (برابر با ۹۹/۵۲) و الگوریتم ExtraTrees (برابر با ۹۹/۱۸)، و در مجموعه داده IBM، بهترین عملکرد مربوط به الگوریتم ExtraTrees (برابر با ۹۵/۸۲) و پس از 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 مجموعه داده hr_data به مراتب بهتر از عملکرد آنها در مجموعه داده IBM است که این نتیجه را نیز می‌توان با اندازه مجموعه داده‌ها مرتبط دانست، چراکه الگوریتم‌های یادگیری تجمعی در برخورد با مجموعه داده‌هایی با اندازه بزرگ‌تر، عملکرد بهتری از خود نشان می‌دهند. با مقایسه جدول‌های (۳) و (۴) می‌توان نتیجه گرفت که استفاده از روش کاهش ابعاد NMF در مجموعه داده IBM، باعث کاهش پیچیدگی الگو و افزایش عملکرد الگوریتم‌ها شده است.

کارایی الگوی یادگیری ماشین، ارتباط مستقیمی با ویژگی‌های داده مورد استفاده برای آموزش آن دارد. اگر ویژگی‌های داده فراهم شده برای الگوی ML نامربوط باشد، روی کارایی الگو تأثیر منفی خواهد داشت. از طرف دیگر، اگر ویژگی‌های داده استفاده شده مناسب باشد، دقت الگوی یادگیری ماشین می‌تواند افزایش یابد. باتوجه به نتایج جدول (۵) مهم‌ترین عوامل مؤثر بر ریزش کارکنان سازمان عواملی همچون «سن»، «درآمد ماهیانه»، «میزان اضافه کاری»، «تعداد شرکت‌هایی که کارمند پیشتر در

[8] A. M. Esmaeeli 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.

[9] 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 عفاوی، م. سرداری زارچی، س. م. م. فاطمی بوشهری، «عوامل مؤثر بر تمایل به ترک سازمان با استفاده از الگوریتم‌های مبتنی بر شبکه 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. ۱۱-۱، 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.

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. باتوجه به شباهت ویژگی‌های مجموعه داده 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. باتوجه به نتایج این پژوهش در جدول (۷)، نتایج این پژوهش با نتایج پژوهش‌های قبلی مقایسه شده‌است.

7-References

۷- منابع

- [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. ۹۸-۹۳.
- [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»، *مجله سیستم‌ها و صنعت*، ۲۰۱۵.



سیده محبوبه مزارعی،
دانش‌نامه 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 موردعلاقه ایشان در زمینه هوش مصنوعی
است.

نشانی رایانامه ایشان عبارت است از:

s.m.mazarei60@gmail.com



جعفر پورامینی، دانش‌نامه
کارشناسی و کارشناسی ارشد
خود را در رشته مهندسی نرم‌افزار
در سال‌های ۱۳۷۸ و تحصیلات
خود را در مقطع دکتری در رشته
مهندسی فناوری اطلاعات در
سال ۱۳۹۷ به پایان رساند و

هم‌اکنون استادیار دانشگاه پیام‌نور استان قم است.
موضوعات موردعلاقه ایشان، زمینه داده‌کاوی، یادگیری
ماشین و یادگیری تقویتی است.

نشانی رایانامه ایشان عبارت است از:

J_pouramini@pnu.ac.ir