

# افزایش بازدهی نیروی کار با استفاده از تکنیک های داده کاوی

مرضیه جوادی<sup>۱</sup>؛ حمیدرضا جوادی<sup>۲</sup>

## چکیده

با توجه به اهمیت و ضرورت مدیریت کارمندان و اطلاع از رضایتمندی یا عدم رضایتمندی آنها از شرایط کاری که تاثیر مستقیم بر کارایی آنها و بسط و توسعه اهداف کاری دارد، نظارت مداوم و دریافت بازخورد آنها در جهت شناسایی و تشخیص به موقع روند افت انگیزه در افراد می تواند، یکی از عوامل مهم در توسعه کسب و کار باشد. مدیران و سرپرستان در صورت جمع آوری اطلاعات فردی و پرسشنامه های رضایت شغلی در بازه های زمانی معین، و مدیریت داده های موجود، و با شناسایی موارد خاص، می توانند در مدیریت نیروهای کاری خود به بهترین کارایی دست پیدا کنند.

در این مقاله با استفاده از رویکردهای داده کاوی و با استفاده از ابزار داده کاوی متن باز WEKA، مجموعه داده کارمندان به منظور انتخاب روشی با بالاترین دقت مورد بررسی و ارزیابی قرار می گیرد.

## کلمات کلیدی

داده کاوی، مدیریت کارمندان، پیش بینی بازدهی کارمندان.

## improving staff performance using data mining techniques

marzieh javadi; hamid reza javadi

### ABSTRACT

With regard to the importance and necessity of management of employees and their satisfaction of working conditions that have the direct impact on their efficiency and expansion and development work goals, constant feedback received them may be one of the important factors in development of business to identify and recognize dropping the motivations.

managers in the case for collecting information and individual questionnaire of job satisfaction in constant time, and management of existing data, and by identifying specific cases, they can manage their work forces to the best efficiency.

In this essay, with using data mining methods and WEKA, staff's dataset evaluate to choosing a way with highest precision.

### KEYWORDS

Data mining, staff management, predicting staff performance

## ۱. مقدمه

واضح است که وجود منابع انسانی واجد شرایط و با انگیزه برای ارائه خدمات مناسب ضروری است [۴]. رضایت شغلی، که به عنوان میزان گرایش کارمندان به شغل آن ها تعریف می شود، از فاکتورهای اصلی انگیزه در نیروی کار است. با توجه به اینکه کارمندان وقتی راضی نگه داشته می شوند، روزهای کمتری را برای تولید بیشتر صرف می نمایند و به شرکت وفادار می مانند، بالا نگه داشتن روحیه کارمندان می تواند سود بالایی برای هر شرکت به همراه داشته باشد. در این حال نتایج عملکرد ضعیف کارکنان بر بازدهی کل مجموعه انکار پذیر نیست. اکثر مشکلات را می توان به انتظارات نامشخص، کمبود منابع و تجهیزات، آموزش نامناسب، حقوق پایین، شرایط کار و زندگی دشوار و یا فقدان انگیزه نسبت داد.

marziehJavadi@gmail.com

<sup>۱</sup> دانشجوی کارشناسی ارشد مهندسی نرم افزار، دانشگاه آزاد اسلامی، واحد زنجان، گروه آموزشی مهندسی کامپیوتر، زنجان، ایران

hmdjavadi@gmail.com

<sup>۲</sup> دانشجوی مهندسی صنایع، دانشگاه آزاد اسلامی، واحد قزوین، قزوین، ایران

یکی از عناصر کلیدی برای موفقیت در حفظ کارکنان و بهبود در عملکرد آن‌ها، استفاده از روش‌هایی است که با تکیه بر شواهد و اطلاعات، و برای آن شرایط و آن گروه از کارکنان تبیین شده باشند [۴].

عوامل بسیاری در بهبود یا حفظ رضایت کارکنان وجود دارد، که کارفرمایان در جهت افزایش بازدهی می‌توانند، آن‌ها را به اجرا در آورند. یکی از عوامل موثر در جهت حفظ پرسنل، که هم ریشه در عوامل فردی و هم ریشه در عوامل مرتبط با شغل دارند، بررسی عوامل موثر در رضایت شغلی آن‌ها است [۴]. برای اندازه‌گیری رضایت کارکنان، بسیاری از شرکت‌ها از پرسشنامه‌های رضایت شغلی و یا جلسات چهره به چهره با کارکنان برای دستیابی به اطلاعات استفاده می‌نمایند. هر کدام از این تکنیک‌ها دارای جوانب مثبت و منفی است که باید با دقت انتخاب و به کار گرفته شود. از آنجائیکه نظرسنجی‌ها اغلب ناشناس هستند به کارکنان اجازه می‌دهد تا با آزادی بیشتر و بدون ترس از عکس‌العمل به سوالات پاسخ دهند. ترکیب مناسبی از نظرسنجی و جلسات چهره به چهره می‌تواند ابزار بسیار خوبی برای شناسایی مشکلات خاص منجر به کاهش روحیه گردد. جمع‌آوری بازخوردهای این نظر سنجی‌ها، می‌تواند یک بینش واقعی را نسبت به نگرش فرد به شغل و محل کارش، فراهم آورد. این اطلاعات می‌تواند در مدیریت کارکنان با هدف توسعه شغلی، موثر باشد.

بسیاری از کارشناسان یکی از بهترین راه‌های حفظ رضایت کارکنان را، ایجاد احساس شرکت در تیم یا خانواده می‌دانند. برخی از کارکنان نیز تنها دریافت پاداش بیشتر یا ارائه امکانات رفاهی بر رضایت آن‌ها تاثیرگذار است. اما اصلی‌ترین فاکتور در رضایت شغلی می‌تواند احترام به کارکنان و کار آن‌ها باشد.

در بخش ۲ از این نوشتار، مجموعه داده‌ای که برای آموزش و آزمایش در این مقاله مورد استفاده قرار گرفته، توصیف گردیده است. در بخش ۳، یک مدل تشخیصی روند افت انگیزه در میان کارمندان با استفاده از تکنیک‌های داده‌کاوی، معرفی گردیده و در ادامه تحلیل و ارزیابی مدل معرفی شده با استفاده از تکنیک‌های داده‌کاوی ارائه می‌گردد. در بخش ۴، روش پیاده‌سازی و در بخش ۵، نتیجه آزمایشات و مقایسه آن‌ها آمده است.

## ۲. توصیف مجموعه داده

مجموعه داده مورد استفاده در این نوشتار، از بیش از ششصد پرسشنامه رضایت شغلی، یک شرکت ارائه‌کننده خدمات، که با توجه به حفظ محرمانگی اطلاعات، نام آن در این تحقیق محفوظ نگه داشته شده، در فاصله سال‌های ۸۴ تا ۹۰ استخراج گردیده است. ویژگی‌های این مجموعه داده در جداول (۱) و (۲) آمده است.

جدول (۱): توصیف مشخصه‌های مجموعه داده مورد استفاده

شماره	نام مشخصه	مقادیر	توضیحات
۱	AGE	numeric	سن
۲	SEX	male, female	جنسیت
۳	education	A,B,C,D,E	سطح تحصیلات، که برای زیر دیپلم از حرف A، دیپلم از حرف B، کاردانی از حرف C، کارشناسی از حرف D، کارشناسی ارشد از حرف E استفاده شده است.
۴	Marital-status	single,married	وضعیت تاهل
۵	Job-experience	numeric	سابقه کار
۶	Adverse-factor	no, yes	وجود عوامل نامطلوب
۷	Job-entertaining	۱،۲،۳،۴،۵	میزان سرگرم‌کنندگی شغل
۸	Job-boring	۱،۲،۳،۴،۵	میزان خسته‌کننده بودن شغل
۹	Unpleasant-job	no, yes	نامطبوع تصور نمودن شغل
۱۰	Comfort-during-work	۱۰،۲۰،۳۰،۴۰،۵۰، ۶۰،۷۰،۸۰،۹۰،۱۰۰	احساس آرامش در حین کار
۱۱	Being-satisfied	۱،۲،۳،۴،۵	راضی بودن از کار به طور نسبی
۱۲	No-interest	۱،۲،۳،۴،۵	میزان بی‌علاقگی به کار
۱۳	Happier-than-colleagues	no, yes	خوشحال تر بودن از سایر همکاران
۱۴	Choose-it-again	no, yes	انتخاب این شغل در آینده
۱۵	Task-done	۱۰-۲۰،۳۰-۴۰،۵۰-۶۰، ۷۰-۸۰،۹۰-۱۰۰	درصد وظایف تکمیل شده از آخرین نظرسنجی تا کنون

جدول (۲) : ویژگی های مجموعه داده مورد استفاده

۶۷۴	تعداد نمونه ها
۱۵	تعداد مشخصه ها
بلی	مقادیر مفقود شده
۷	تعداد صفات عددی
۸	تعداد صفات اسمی

### ۳. فرآیند داده کاوی برای پیش بینی کارایی نیروی کار

داده کاوی عموماً برای استخراج دانش از مجموعه داده ها و پیش بینی وضعیت جدید در حالتیکه پردازش داده ها به صورت دستی دشوار است، به کار می رود. روش های اصلی داده کاوی به دو دسته توصیفی و پیش بینی کننده تقسیم می شوند. ابتدا بر روی مجموعه داده کارمندان، عملیات پیش پردازش را انجام می دهیم. سپس با استفاده از داده های پیش پردازش شده و روش های داده کاوی، عملیات آموزش را ادامه داده و به ساخت مدل با استفاده از این الگوریتم های زیر می پردازیم.

#### ۱.۱.۳ روش های کلاسه بندی

کلاسه بندی یکی از مهم ترین و پرکاربردترین روش های داده کاوی می باشد که در یک فرآیند دو مرحله ای انجام می گیرد. در مرحله اول (مرحله یادگیری) ، الگوریتم دسته بندی با آنالیز کردن مجموعه داده ها، مدل را می سازد و به کشف روابط میان داده ها می پردازد و در مرحله دوم از مدل برای دسته بندی داده ها استفاده می کند [۱]. برخی از روش های کلاسه بندی در ادامه توصیف شده است.

#### ۱.۱.۳ Rules

از این مجموعه الگوریتم های کلاسه بندی، ZeroR یک الگوریتم کلاسه بندی ساده است که می تواند بعنوان کران پایین کارایی بکار رود. این الگوریتم کلاس اصلی داده آزمایشی (در صورتی که غیر عددی باشد) یا مقدار میانگین (در صورتی که عددی باشد) را محاسبه می کند.

#### ۱.۲.۲ کلاسه بندی بیزین!

کلاسه بندهای بیزین، کلاسه بندهای آماری هستند و احتمال تعلق داشتن داده به کلاس خاصی را پیش بینی می کنند. آن ها از قضیه بیز استفاده می کنند [۲]. دقت و سرعت آن ها برای پایگاه داده های بزرگ، بالا می باشد. بیزین ساده<sup>۲</sup>، احتمال تمام کلاس ها را به شرط داده جدید محاسبه و سپس کلاس با بالاترین احتمال را برای آن داده انتخاب می نماید. اعمال این الگوریتم بر داده های مورد استفاده، دقت ۵۹,۰۵۰۴٪ و اعمال الگوریتم شبکه های بیزین بر داده های مورد استفاده، دقت ۷۳,۲۹۳۸٪ را به همراه داشته است.

#### ۱.۳.۳ کلاسه بندهای درختی

در یک مدل کلاسه بندی درختی<sup>۳</sup> که بوسیله یک ساختار شبیه به درخت نمایش داده می شود، هر گره داخلی نماینده یک ویژگی و هر شاخه نماینده یکی از نتایج ممکن و هر برگ نشان دهنده یک کلاسه بندی است. با توجه به الگوریتم انتخاب شده، مدل درخت تصمیم متفاوت است. الگوریتم DecisionStump درخت تصمیم یک سطحی و الگوریتم LMT مدل درختی منطقی می سازد. الگوریتم random tree تعدادی از مشخصه های تصادفی را برای ساخت درخت در نظر می گیرد. الگوریتم REPTree یادگیرنده درختی سریعی است که از هرس کاهش دهنده خطا استفاده می کند. مقایسه دقت این الگوریتم ها برای مجموعه داده مورد استفاده، در جدول (۳) آمده است.

جدول (۳): دقت کلاسه بند های درختی

نام کلاسه بند درختی	دقت کلاسه بند
DecisionStump	٪ ۵۱,۳۳۵۳
J۴۸	٪ ۸۳,۲۳۴۴
LMT	٪ ۸۴,۷۱۸۱
RandomForest	٪ ۸۴,۷۱۸۱
RandomTree	٪ ۸۳,۶۷۹۵
REPTree	٪ ۷۷,۴۴۸۱

### ۴,۱,۳. الگوریتم های MetaLearning

metalearner ها، کلاسه بندهای ساده را گرفته و به یادگیرنده های قدرتمند تبدیل می نمایند.

#### ۱,۴,۱,۳. Bagging

یک کلاسه بندی ترکیبی که طی مراحل نمونه هایی را با جایگزینی از داده های آموزشی انتخاب کرده و کلاسه بندی روی آنها انجام می گردد. در بین نتایج، کلاسی بعنوان نتیجه انتخاب می شود که بیشتر از سایر کلاس ها، به عنوان کلاس توصیه شده است [۳]. این روش برای افزایش دقت بکار برده می شود. اعمال این الگوریتم، دقت ۸۱,۸۹۹۱٪ را به همراه داشته است.

#### ۲,۴,۱,۳. Boosting

در این روش، الگوریتم کلاسه بندی به هر نمونه از داده ها وزنی را اختصاص می دهد. سپس الگوریتم چندین مرحله تکرار می شود و در هر تکرار، وزن هر نمونه اصلاح می شود. در نهایت امکان انتخاب نمونه هایی با وزن بیشتر، بالاتر است. AdaBoost مشهورترین نوع Boosting می باشد. دقت الگوریتم AdaBoost، ۵۱,۳۳۵۳٪ می باشد.

#### ۳,۴,۱,۳. استفاده از MetaLearner ها برای ایجاد کلاسه بند توسط کاربر

نتایج اعمال دو کلاسه بند metalearner که بر روی کلاسه بندهای درختی آزمایش شده اند، در جداول (۴) و (۵) آمده است.

جدول (۴): اعمال کلاسه بند AdaBoostM۱ بر کلاسه بندهای درختی

نام کلاسه بند درختی	دقت کلاسه بند
DecisionStump	۵۱,۳۳۵۳٪
J۴۸	۸۵,۴۵۹۹٪
RandomForest	۸۱,۸۹۹۱٪
RandomTree	۸۴,۴۲۱۴٪
REPTree	۸۱,۰۰۸۹٪

جدول (۵): اعمال کلاسه بند Bagging بر کلاسه بندهای درختی

نام کلاسه بند درختی	دقت کلاسه بند
DecisionStump	۵۱,۱۸۶۹٪
J۴۸	۸۵,۷۵۶۷٪
RandomForest	۸۵,۰۱۴۸٪
RandomTree	۸۴,۷۱۸۱٪
REPTree	۸۱,۸۹۹۱٪

### ۲,۲. پیش پردازش، آموزش و آزمایش

یکی از تکنیک هایی که برای ارزیابی کارایی الگوریتم ها بکار می رود، cross validation می باشد. در cross fold validation، k داده اولیه را بصورت تصادفی به k زیرمجموعه یا fold با اندازه های تقریباً مساوی تقسیم می کند. آموزش و تست، k مرتبه انجام می گیرد. در تکرار i ام، پارامتر i به عنوان مجموعه تست و سایر پارامترها به عنوان آموزشی در مدل می باشند. هر نمونه از زمان ثابتی برای آموزش و برای تست استفاده می کند. دقت کلاسه بندی، به صورت تعداد کل کلاسه بندی های درست در k تکرار تقسیم بر مجموع تعداد چندتایی ها در داده های اولیه بدست می آید [۲].

برای آموزش و آزمایش بر روی این مجموعه داده از روش 10 – fold validation استفاده شده است. در این روش، مجموعه داده به صورت تصادفی به ۱۰ قسمت تقسیم می شود و در ۱۰ آزمایش مختلف، هر بار یکی از قسمت ها به عنوان مجموعه داده آزمایش و ۹ قسمت دیگر به عنوان مجموعه داده آموزشی به کار می رود.

### ۳,۲. تحلیل و ارزیابی

در اینجا اگر داده ها را در دو دسته کارمندی که کمتر از ۲۰٪ وظایف خود را تکمیل نموده اند، و سایرین دسته بندی نمائیم، ارتباط میان کلاس های پیش بینی شده و کلاس های واقعی در جدول (۶) که به confusion matrix معروف است، آمده است.

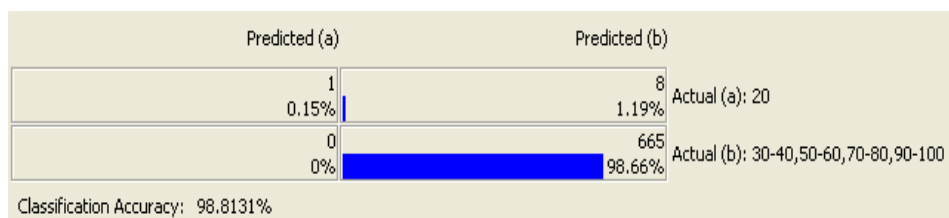
جدول (۶): confusion matrix

		کلاس های پیش بینی شده	
		a	b
کلاس های واقعی	a	(TP) True positive	(FN) False negative
	b	(FP) False positive	(TN) True negative

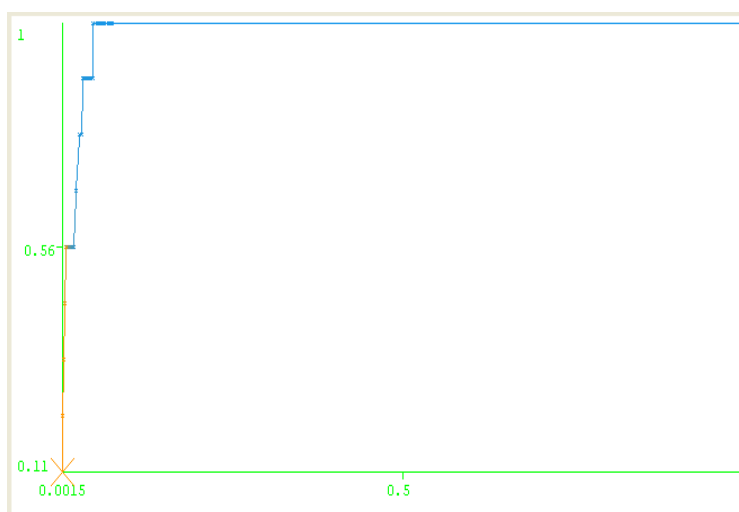
نرخ موفقیت کل، که از تقسیم تعداد کلاسه بندی های درست بر تعداد کل کلاسه بندی ها، به دست می آید. به صورت زیر است:

از آنجائیکه در اینجا هدف تشخیص افت شدید کارایی است، پس معیار حساسیت، عبارت است از تعداد نمونه هایی که به درستی عملکرد آن ها ضعیف تشخیص داده شده، نسبت به تعداد کل نمونه هایی که واقعا دارای عملکرد ضعیف هستند. معیار ویژگی، عبارت است از تعداد نمونه هایی که به درستی دارای عملکرد عادی تشخیص داده می شوند، نسبت به تعداد کل نمونه هایی که واقعا دارای عملکرد عادی هستند. مقدار پیش بینی مثبت، از تقسیم تعداد نمونه هایی که به درستی دارای عملکرد ضعیف تشخیص داده می شوند بر تعداد کل نمونه هایی که دارای عملکرد ضعیف تشخیص داده می شوند. مقدار پیش بینی منفی، نیز از تقسیم تعداد نمونه هایی که به درستی دارای عملکرد عادی تشخیص داده می شوند بر تعداد کل نمونه هایی که دارای عملکرد عادی تشخیص داده می شوند.

تحلیل نتایج برای برخی از الگوریتم هایی که بالاترین دقت را به همراه داشته اند، در ادامه آمده است. confusion matrix مربوط به اعمال فرا یادگیرنده bagging بر J۴۸ در تصویر (۱) و منحنی آستانه آن در تصویر (۲) آمده است. همچنین confusion matrix مربوط به Random Forest نیز در تصویر (۳) آمده است. ماتریس confusion مربوط به اعمال فرا یادگیرنده AdaBoostM۱ بر J۴۸، با ماتریس مربوط به اعمال فرا یادگیرنده bagging بر J۴۸، یکسان است.



تصویر (۱): ماتریس confusion اعمال فرا یادگیرنده bagging بر J۴۸



تصویر (۲): منحنی آستانه اعمال فرا یادگیرنده bagging بر J۴۸

Predicted (a)		Predicted (b)		
4	0.59%	5	0.74%	Actual (a): 20
1	0.15%	664	98.52%	Actual (b): 30-40,50-60,70-80,90-100

Classification Accuracy: 99.1098%

تصویر (۳) : ماتریس confusion اعمال الگوریتم RandomForest

نتایج حاصل از مقایسه معیارهای دقت، حساسیت، ویژگی، پیش بینی مثبت و پیش بینی منفی در جدول (۷) آمده است.

جدول (۷) : معیارهای ارزیابی مربوط به برخی از روش های کلاسه بندی

پیش بینی منفی	پیش بینی مثبت	ویژگی	حساسیت	دقت	روش دسته بندی
%۹۸,۸۱	%۱۰۰	%۱۰۰	%۱۱,۱۱	%۹۸,۸۱۳۱	Bagging with J۴۸
%۹۹,۲۵	%۸۰	%۹۸,۸۴	%۴۴,۴۴	%۹۹,۱۰۹۸	RandomForest

#### ۴. پیاده سازی

نرم افزار WEKA [۵] الگوریتم های یادگیری را برای اعمال کردن بر روی مجموعه داده ها در اختیار کاربران قرار می دهد. همچنین شامل ابزار های مختلفی برای تبدیل قالب های داده می باشد. نرم افزار WEKA نسخه ۷.۳ ، برای پیاده سازی روش های فوق مورد استفاده قرار گرفته است و برای هر الگوریتم از مقادیر پیش فرض نرم افزار استفاده گردیده است.

#### ۵. نتیجه

در این مقاله، مجموعه داده های کارمندان، با هدف پیش بینی روند افت کارایی، مورد بررسی و ارزیابی قرار گرفته است. با ارزیابی های انجام شده، به کمک الگوریتم k fold cross validation با مقدار  $k = 10$  ، برای داده ها، اعمال تکنیک های RandomForest ، فرا یادگیرنده Bagging بر روی J۴۸ ، بالاترین دقت را به همراه داشته است.

#### ۶. مراجع

[1] مظهری، نیلوفر؛ ایمانی، مهدی؛ جودکی، مجید؛ قلیچ پور، احمد؛ "مروری بر کلاسه بندی و الگوریتم های آن"، سومین کنفرانس داده کاوی ایران، ۱۳۸۸.

[2] Han, J, Kamber, M "Data Mining Concepts and Techniques", Second Edition, Morgan Kaufman, San Francisco, ۲۰۰۶.

[3] Witten, I. H, Frank, E, Hall, M.A "Data Mining :Practical Machine Learning Tools and Techniques", Third Edition, Morgan Kaufmann, Burlington, ۲۰۱۱.

[4] Dieleman .M; Harnmeijer .J.W; "Improving health worker performance in search of promising practices ", Evidence and Information for Policy, Department of Human Resources for Health, Geneva, September ۲۰۰۶

[5] <http://www.cs.waikato.ac.nz/me/weka>

<sup>۱</sup> Bayesian Classification  
<sup>۲</sup> Naive Bayesian  
<sup>۳</sup> Tree classification model