یک طبقه بندی بیز ساده برای مدارک علمی

چکیده

طبقه‌بندی دستی افراد به دسته‌های مختلف براساس مدرک تحصیلی‌شان، کار خسته‌کننده‌ای است و این ممکن است که تغییری نسبت‌به سناریوی بحث شده، ایجاد کند. این مقاله یک روش طبقه‌بندی با استفاده از معیار الگوریتم طبقه‌بندی بیز ساده برای طبقه‌بندی افراد به دسته‌های مختلف براساس نمایش برخی ویژگی‌های مدارک علمی‌شان پیشنهاد می‌کند. نتایج آزمایشی قابل ارزیابی نشان می‌دهد که روش طبقه‌بندی پیشنهادی می‌تواند امیدوارکننده باشد و در جای دیگر استفاده شود. روش پیشنهادی، به‌صورت آزمایشی با90٪ دقت مقادیر kappa، اثبات کارایی بالای آن را تأیید می‌کند. این روش طبقه‌بندی می‌تواند کار دستی این جهان را کاهش دهد و به‌راحتی در دسته‌بندی کمک کند.

کلمات کلیدی : طبقه‌بندی، داده‌کاوی، مدارک علمی، kappa، بیز ساده

1. مقدمه

متعدد پیش آمده که یک فرد براساس تجزیه‌وتحلیل مدارک علمی که در زندگی کسب کرده است، در نظر گرفته شود. در چنین مواردی، طبقه‌بندی افراد باتوجه به مدرک تحصیلی آموزشی‌شان می‌تواند به تصمیمی فنی آزادانه، و عاری از هرگونه تعصب بسیار کمک کند و از این‌رو می‌تواند قابل اجرا باشد.

این مقاله روشی برای دسته‌بندی مدارک علمی با استفاده از معیار الگوریتم طبقه‌بندی بیز ساده پیشنهاد می‌کند. این روش می‌تواند در بسیاری از برنامه‌های کاربردی مانند تفکیک براساس فهرست کوتاه ارتباط آموزشی، برای استخدام افراد براساس میزان تحصیلات و غیره، استفاده شود. سازمان این مقاله در زیر آورده شده است: بخش2 شامل بررسی ادبیات می‌شود. بخش3 الگوریتم بیزین ساده و روش طبقه‌بندی پیشنهادی را توضیح می‌دهد. بخش4 تجزیه‌وتحلیل نتایج آزمایشی براساس جدول‌بندی ذکر شده است و بخش5 مقاله را نتیجه‌گیری می‌کند.

الگوریتم بیزین ساده یک الگوریتم طبقه‌بندی کلاسیک است که بهره‌وری خود را در برنامه‌های مختلف و چند مقالهٔ نمایش بهره‌وری از طبقه‌بندی‌کننده که در اینجا ورد بحث است، ثابت کرده است.

مقالهٔ Mauricio A.Valle و همکاران روش پیش‌بینی ویژگی‌های تعیین‌کننده درمورد یک الگوریتم طبقه‌بندی بیز ساده شامل یک روش تست براساس اعتبارسنجی متقابل را بحث می‌کند. تأیید آزمایشی صفات اجتماعی و جمعیت شناختی است که به پیش‌بینی عملکرد آینده عامل فروش در یک مرکز تلفن کمک نمی‌کند.

Dunja Mladenic و همکاران. پژوهش با انتخاب ویژگی‌های کمک برای طبقه‌بندی با استفاده از مشخصات خاص و توانایی یادگیری طبقه‌بندی بیش از یک دادهٔ متنی که توزیع نابرابری است. زمانی‌که دامنه و ویژگی‌های الگوریتم طبقه‌بندی شده، در نظر گرفته شود، عملکرد طبقه‌بندی افزایش می‌یابد. Dong Tao و همکاران. مقالهٔ بهبود الگوریتم بیزی ساده را با ترکیب روش کلاسیک با یک روش انتخاب ویژگی براساس شاخص Gini پیشنهاد می‌کند. این روش ترکیبی، عملکرد طبقه‌بندی متن را بهبود می‌بخشد.

Kabir Md Faisal و همکاران. پژوهش ترکیب روش خوشه‌بندی k-means با الگوریتم طبقه‌بندی ساده برای افزایش دقت. این روش خوشه‌بندی گروه‌های نمونه‌های آموزشی را به دسته‌بندی‌های مشابه، پس از آن همهٔ گروه‌ها تحت طبقه‌بندی بیز ساده آموزش داده می‌شوند. این روش برای تأیید بهبود دقت است.

Santra A.K. و همکاران. تحقیقات ثابت می‌کند که درمورد استفاده از کاربرد وب، درحالی‌که از یک طبقه‌بندی بیز ساده به‌جای درخت تصمیم استفاده می‌کند، که زمان صرف شده برای طبقه‌بندی و حافظه کاهش پیدا می‌کند. مقاله نشان می‌دهد که ماهیت استقلال شرطی ویژگی‌ها در الگوریتم اصلی بیز ساده در بعضی موارد ضعیف به‌نظر می‌رسد و روش وزن‌دهی محلی که از الگوریتم کلاسیک از نظر دقت بهتر است، پیشنهاد می‌کند. Pradeepta K. Sarangiet و همکاران. مقاله استخراج ویژگی با استفاده از تجزیه LU به‌دنبال استفاده از طبقه‌بندی بیز ساده برای تشخیص الگو توصیف می‌کند. این کاربرد جهانی طبقه‌بند را نشان می‌دهد.

Birant D وYildrim . مقالهٔ تحقیقاتی تأیید آزمایشی اثر توزیع‌های مختلف برروی ویژگی های مورد بحث. مشاهده شده است که کاربرد توزیع براساس طبیعت ویژگی‌ها به‌جای استفاده از یک توزیع در سراسر تمام ویژگی‌ها دقت را افزایش می‌دهد. AbeerBadr El Din AhmedوIbrahim SayedElarabarticle بحث درمورد استفاده از الگوریتم‌های طبقه‌بندی برای پیش‌بینی نمرهٔ نهایی دانش آموزان است. مقاله Ron Kohavi ترکیب طبقه‌بند بیز را با درخت تصمیم که به‌عنوان NBTree نامیده است، برای افزایش دقت طبقه‌بند پیشنهاد می‌کند.

همچنین دریافتند که استقلال شرطی کلاس در مورد مجموعه داده کوچک منفعل است اما درصورت مجموعه داده‌های بزرگ، این فرض منجر به اشکال در طبقه‌بندی و کاهش دقت می‌شود.

Shasha Wang و همکاران. مقاله نسخهٔ ارتقا یافتهٔ طبقه‌بندی NBTree ترکیبی را پیشنهاد و آن را به‌عنوان NBTree چندگانه به‌نام MNBTree، که در آن یک طبقه‌بند بیزین سادهٔ چندجمله‌ای برای گره‌های برگ درخت تصمیم کاربرد دارد. علاوه‌براین، برای افزایش عملکرد، بداههٔ دیگری با گنجاندن طبقه‌بندی چند کلاسه ساخته شده است و سیستم به‌عنوان نسخهٔ چندکلاسه MNBTree نامیده می‌شود.

با توجه به مقالات پژوهشی فوق‌الذکر، جوانب مثبت الگوریتم طبقه‌بندی بیزین به‌طور کامل مطالعه شده و دریافتند که این الگوریتم بهترین خواهد شد با توجه به ماهیت داده‌های مورد استفاده برای آزمایش متشکل از هردو دادهٔ عددی و متنی که به‌طور مستقل کمک به طبقه‌بندی باشد.

2. سیستم پیشنهادی

2.1 مرور کلی بیز ساده

الگوریتم طبقه‌بندی بیز بسیار ساده است که در آن فرض بر این است که ویژگی‌های طبقه‌بندی مستقل هستند و بین آنها هیچ همبستگی وجود ندارد. بسیاری از محققان دریافته‌اند که این فرض استقلال در تمام موارد که دیگر روش‌های جایگزین ارائه شده، برای افزایش عملکرد کار نمی‌کند. یکی از روش‌های جایگزین را Liang xiao Jiang ارائه داده است.

تکنیک اصلی بیز ساده بر احتمال شرطی و حداکثر وقوع احتمال است. الگوریتم بیز ساده براساس توضیحات ارائه شده به شرح زیر است:

- فرض کنید G مجموعهٔ آموزش با N متغیر است که در آن K به‌عنوان بردار ویژگی بُعد M در $M=\{M1,M2,\ldots Mk\}$ نشان می‌دهد.
- برای 'P' کلاس‌های $C_1, C_2,\ldots C_p$ وجود دارد. با توجه به این طبقه‌بندی بیز ساده، T یک متغیر متعلق به کلاس C_x است که فقط زمانی آن را به یک احتمال شرطی بالاتر از هرکلاس دیگر C_y می‌برد، که در آن $x \neq y$ باشد.

$$P(C_x \mid T) > P(C_y \mid T) \quad \text{و} \quad P(C_x \mid T) = (P(T \mid C_x * P(C_x)) / P(T)$$

- از کلاس مستقل مشروط فرض شده است،

$$P(M_k \mid C_x) = P(M_1) * P(M_2) * P(M_3) \ldots *P(M_k)\ P(M \mid C_x) = \prod_{i=1}^{k}$$

- کلاس C_x به‌عنوان طبقهٔ خروجی پیش‌بینی شده $P(M \mid C_x) * P(C_x) > P(M \mid C_y) * P(C_y)$، که در آن $1\ p \geq \leq x,y$ و $x \neq y$ است.

2.2 شرح دیتاست

سناریویی که اینجا مطرح شده با جزئیات مدارک علمی افراد که آنها در دورهٔ آموزشی‌شان کسب کرده‌اند سروکار دارد. براساس این جزئیات، افراد به کلاس‌های مختلف طبقه‌بندی می‌شوند. در اینجا سه کلاس در نظر گرفته شده و آنها عبارتند از:

- کم.
- متوسط.
- زیاد.

ویژگی‌های گرفته شده در دست بررسی در زیر آورده شده:

- عدد کل سال‌هایی که برای تحصیل صرف شده.
- درجهٔ تحصیلی کسب شده.
- موضوع اصلی / رشته تخصصی مربوط به سناریو تحت بررسی.

عدد کل سال‌هایی که برای آموزش سپری شده به‌عنوان یک ویژگی تعیین عمق دانش به‌دست آمده، به‌طور مستقیم به زمان سرمایه‌گذاری در یادگیری متناسب با آموزش در نظر گرفته شده است.

درجه علمی به‌دست آمده توسط یک فرد، نقش حیاتی را در تصمیم‌گیری مهارت شخص بازی می‌کند.

زمینه‌های تخصصی افراد به‌عنوان ویژگی‌های در نظر گرفته در تعیین دلیل انعطاف‌پذیری این سیستم شده که در انواع حالات استفاده می‌شود و این ویژگیِ کاربرد وابسته است.

شامل مجموعهٔ داده در حدود 25000 ورودی که در آن 20000 ورودی برای آموزش و باقی مانده 5000 ورودی برای تست بهره‌وری طبقه‌بندی و تعیین چگونه خوب یادگرفته، به طبقه‌بندی داده‌ها استفاده می‌شود. نمایش مجموعه داده به‌صورت زیر است:

- نمایش کل داده‌ها عبارتست از $D = \{D_1, D_2, \ldots D_{25000}\}$
- نمایش کل داده‌های آموزشی عبارتست از $\{D_1, D_2, \ldots D_{20000}\}$
- نمایش داده‌های تست عبارتست از $\{D_{20001}, D_{20002}, \ldots D_{25000}\}$

یک فرد با مدرک علمی مشابه ممکن است تحت تفکیک حالات مختلف شده باشد، چون طبقه‌بندی کاربرد وابسته است.

2.3 آموزش و تست مجموعه داده نمونه

این مجموعهٔ آموزشی نمونه مطابق با انتخاب یک کاندید برای آموزش دپارتمان شیمی در زیر ارایه شده است. این جدول به‌وضوح ویژگی‌های مختلف مورد استفاده برای طبقه‌بندی و کاربرد وابسته را نشان می‌دهد. داده‌های تست مشابه داده‌های آموزش بدون ستون کلاس خواهد بود که می‌تواند با کمک اجرای الگوریتم زیر پیش‌بینی کند که در زیر ارائه شده است.

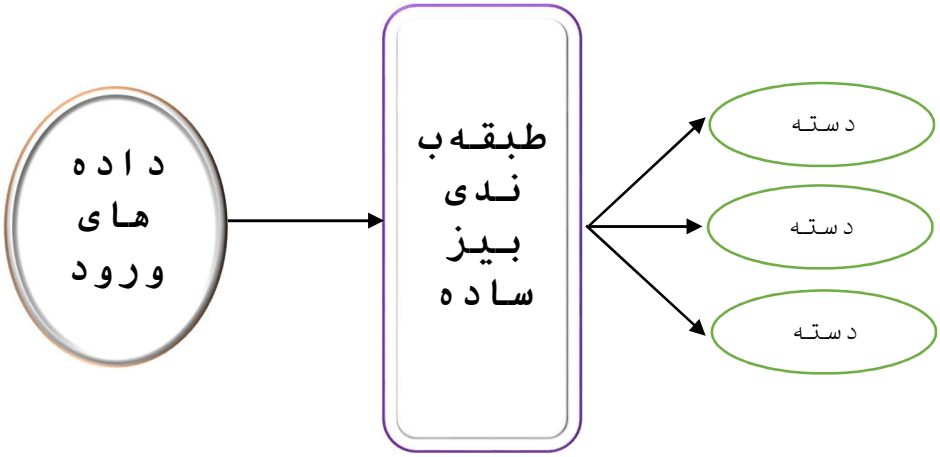
جدول2.

کلاس	تخصص	تعداد سال‌ها	مدرک تحصیلی
متوسط	شیمی	15	کارشناسی علوم‌پایه
کم	تجارت	17	کارشناسی ارشد بازرگانی
زیاد	شیمی	17	کارشناسی ارشد علوم‌پایه
زیاد	شیمی	25	دکترا
کم	مدیریت	18	کارشناسی ارشد مدیریت تجاری
کم	فیزیک	15	کارشناسی علوم‌پایه
کم	زیست‌شناسی	22	دکترا

توجه: اگر یک فرد یا مدرک علمی بالا در درجه‌بندی PG، برچسب کلاس کم دارد به‌دلیل ویژگی مساحت تخصص کاربرد وابسته است.

2.4 روش طبقه‌بندی

داده‌های آموزشی برای دسته‌بندی داده‌ها به سه دستهٔ تحت بررسی که در بخش بالا براساس ویژگی‌ها تعیین شده‌اند، استفاده می‌شوند. معماری برای چنین روش طبقه‌بندی در شکل 1 آورده شده است.



شکل 1. معماری طبقه‌بندی ارائه شده است.

2.5 الگوریتم

روش: الگوریتم بیز ساده برای مدارک علمی

- شروع
- مقداردهی اولیه

  nc — تعداد دسته‌ها

  na تعداد ویژگی‌ها

  N — تعداد نمونه‌ها

- برای هر کلاس C_i انجام می‌دهد

  محاسبهٔ احتمال قبلی $P(C_i) = \sum C_i / \sum N,\ i \in \{1,nc\}$

- برای هر کلاس C_i انجام می‌دهد

  برای هر کلاس A_j انجام می‌دهد

  محاسبهٔ احتمال شرطی $P(A_j|C_i) = \sum C_i$

با $Aj / \sum C_i , i \in \{1,..,nc\}$ ،

و $j \in \{1,..,na\}$

- برای هر کلاس Ci انجام می‌دهد

محاسبه احتمال شرطی چند تایی $K \text{ i.e } P(K|C_i) = P(A_1|C_i) * P(A_2|C_i) * \ldots * P(A_{na}|Ci)$

- برای هر کلاس Ci انجام می‌دهد

محاسبه احتمال قبلی چند تایی $K \text{ i.e } P(C_i) * P(K|C_i)$

- پیش‌بینی

اگر $((P(C_p) * P(K|C_p)) > (P(C_q) * P(K|C_q)))$

پیش‌بینی C_p ←

در غیر این‌صورت

پیش‌بینی C_q ←

که در آن $p,q \in \{1,..,nc\}$ و $p \neq q$ باشد

- پایان

3. نتایج آزمایش و تجزیه‌وتحلیل

خروجی طبقه‌بندی تعداد نمونه‌های تست گرفته شده نشان می‌دهد که طبقه‌بندی در سه کلاس مختلف جدول است در جدول1. براساس این جدول معیارهای عملکرد محاسبه شده است و درجدول2 و جدول3 فهرست شده‌اند. جدول2 دقت و مقادیر طبقه‌بندیkappa نشان می‌دهد. جدول3 فهرست حساسیت، ویژگی و درجهٔ توزیع سه سطح مختلف بحث شده را فهرست کرده است.

جدول1. تعداد نمونه‌هایی که در سطح‌های مختلف دسته‌بندی شده‌اند.

پیش‌بینی/	**دسته‌های**

انتظار	کم	متوسط	زیاد
کم	1233	126	0
متوسط	100	1716	126
زیاد	0	133	1566

جدول2. دقت و معیارهای Kappa طبقه‌بندی

پارامترها	مقدار
دقت	**0.903**
Kappa	**0.8528**

جدول3. میزان حساسیت و گزینش‌پذیری برای سه سطح

پارامترها	دسته کم	دسته متوسط	دسته زیاد
حساسیت	0.9250	0.8689	0.9255
ویژگی	0.9656	0.9253	0.9598
توزیع	0.2666	0.3950	0.3384

3.1 معیارهای عملکرد

به‌طور سنتی، دقت، اصلی‌ترین معیار برای تعیین طبقه‌بندی‌کننده بود. اما وقتی که مجموعه داده به‌خاطر تمایلش به‌سمت کلاس غالب، کاملا انحراف‌دار است، دقت، غیرقابل اعتماد است بنابراین مقادیر گوناگون دیگر نشان‌می‌دهند که می‌توان به‌وضوح توانایی حقیقی یک طبقه‌بندی‌کننده را توصیف کرد.

میزان کارایی برای تعیین عملکرد دسته‌بندی‌کننده‌هایی که در زیر، توضیح داده‌شده، استفاده می‌شود.

3.1.1 حساسیت:

حساسیت به‌معنای توانایی طبقه‌بندی‌کننده در شناسایی دسته‌ی مثبت از دسته‌ی مثبت و منفی و به‌طور صحیح از دستهٔ منفی است. حساسیت، نسبت حقیقی (صحیح) ارزش‌های مثبت را نشان می‌دهد.

(منفی پیش‌بینی‌شده + مثبت پیش‌بینی‌شده) / مثبت پیش‌بینی‌شده = حساسیت

3.1.2 ویژگی:

ویژگی، توانایی طبقه‌بند را به‌طور مناسب در مستثنی نمودن ارزش دسته‌های مثبت از دسته‌های منفی و دسته‌های منفی از مثبت‌ها نشان می‌دهد. ویژگی، نسبت ارزش‌های حقیقی منفی را تشریح می‌کند.

(مثبت به اشتباه پیش‌بینی‌شده + منفی پیش‌بینی‌شده) / منفی پیش‌بینی‌شده = ویژگی

3.1.3 دقت:

دقت، ملاکی‌ست برای تعیین مقداری از ارزش‌های دسته‌ی مثبت که به‌درستی از مثبت‌ها دسته‌بندی شده‌اند و مقداری از دسته‌ی منفی که دقیقا به‌عنوان منفی علامت‌گذاری شده‌اند. دقت، نشان‌دهنده ارزش‌هایی‌ست که به‌درستی دسته‌بندی شده‌اند.

(کل منفی‌ها+ کل مثبت‌ها) / (منفی پیش‌بینی‌شده + مثبت پیش‌بینی‌شده) = دقت

3.1.4 Kappa:

ارزش Kappa نشان‌دهندهٔ توافق ارزیابی بینابین و شرایط مناسب توسط مقدار1 Kappa است. به‌همان اندازه که ارزش کاهش می‌یابد، شرایط سطح نیز کاهش می‌یابد.Kappa ، ارزش آماری نرمالی، که دقت سیستم تعریف‌شده توسط کاربر، و آنکه از یک سیستم فرضی تصادفی مقایسه می‌کند را نشان‌می‌دهد. این ثابت شده که بیشترین ملاک مؤثر در تعیین بازده طبقه‌بندی‌کننده، مقایسهٔ سیستم‌شان با یک سیستم ایده‌آل است. این، همچنین تعیین می‌کند که طبقه‌بندی‌کننده چگونگی دسته‌بندی را آموخته یا مقادیر آموزشی را به‌خاطر سپرده، به‌عبارت دیگر، طبقه‌بندی‌کننده رابطهٔ بین ویژگی‌ها با روش دقیق را درک می‌کند و نه اینکه فقط کلاس مقادیر تکراری را، تکرار کند.

$$\text{Kappa} = (\text{دقت} - \text{تصادفی}) / (1-\text{تصادفی})$$

تصادفی= ((منفی صحیح + مثبت ناصحیح) * (منفی صحیح + منفی ناصحیح) + (منفی ناصحیح + مثبت صحیح) * (مثبت ناصحیح + مثبت صحیح)) تقسیم‌بر (کل مثبت‌ها + کل منفی‌ها) * (کل مثبت‌ها + کل منفی‌ها)

3.1.5 توزیع

توزیع نسبت دستهٔ مثبت به کل جمعیت آن است و به‌صورت

(کل منفی‌ها + کل مثبت‌ها) / (منفی به اشتباه پیش‌بینی‌شده + مثبت پیش‌بینی‌شده) = توزیع

بدست می‌آید.

3.1.6 نمادهای استفاده شده

جدول4 شامل نمادهای مقادیر مورد استفاده برای محاسبهٔ عملکرد ذکر شده بالا است. این مقادیر مبنای تعیین محاسبهٔ عملکرد ذکر شدهٔ بالا است. این مقادیر معمولا از ماتریس درهم‌ریختگی بدست آمده‌اند. یک ماتریس درهم‌ریختگی، یک ماتریس دو در دو دارای ارزش‌های اصلی کلاس مثبت و منفی است که توسط طبقه‌بند، بعد از آموزش، پیش‌بینی شده است.

جدول4. نمادگذاری‌هایی که در محاسبهٔ ملاک‌های عملکرد استفاده شده است:

مقداری که به‌درستی تشخیص داده شده(مثبت حقیقی)
مقداری که به اشتباه صحیح تشخیص داده شده(مثبت کاذب)
مقداری که به‌درستی ناصحیح تشخیص داده شده(منفی حقیقی)
مقداری که به اشتباه ناصحیح تشخیص داده شده(منفی کاذب)
مقادیر مشمول دستهٔ X به‌عنوان کلاس X شناخته شده و دستهٔ Y به‌عنوان کلاس Y
مقادیر مستثنی دستهٔ Y به‌عنوان کلاس X شناخته نشده و دستهٔ X به‌عنوان کلاس Y شناخته نشده
کل مثبت‌ها همهٔ مقادیری که به‌عنوان دستهٔ مثبت شناخته شده است
کل منفی‌ها همهٔ مقادیری که به‌عنوان دستهٔ منفی شناخته شده است

3.2 تحلیل:

کارایی نتایجی که الگوریتم طبقه بندی بیز ساده ارائه می‌کند، وقتی‌که مشخصات(نشانه‌ها)، غیرعددی هستند به‌خوبی اثبات می‌شود. چرا که اصالت طبیعی دسته‌بندی‌کننده، محاسبه براساس بیشترین درست‌نمایی حادثه

هستند، نتایج دسته‌بندی رضایت‌بخش‌اند و برخی دسته‌بندی‌کننده‌ها می‌توانند در کاربردهای گوناگون جایی که یک شخص می‌تواند براساس درجه‌ی علمی اکتسابی‌اش طبقه‌بندی کند، استفاده شود.

دقت بالا و ارزش بالای Kappa نمایان می‌شود(اشاره دارد به) که دسته‌بندی توسط دانشی که درطول فرآیند آموزش، کسب‌شده، ساخته شده و از این‌رو ثابت می‌شود که یک سیستم امیدوارکننده است. ارزش بالای Kappa همچنین معرفی می‌کند آنچه را که سیستم به‌منظور طبقه‌بندی نسبت‌به حفظ مقادیر آموزش دیده است.

ارزش بالای حساسیت و اختصاصی‌بودن، مشخص می‌کند که دسته‌بندی‌کننده، قادر به شناسایی یک کلاس(سطح)مثبت به‌صورتی که از مثبت‌ها و مانع‌شدن(مستثنی‌کردن) کلاس‌های منفی از مثبت و برعکس.

الگوریتم ساده‌ی بیز که کارایی‌اش در وضعیت‌های گوناگون زندگی ثابت شده‌است، باردیگر کارایی‌اش در این سناریوی متعهدشده، اثبات شده‌است.

4. نتیجه‌گیری:

یک روش طبقه‌بندی براساس طبقه‌بندی بیز ساده برای طبقه‌بندی افراد مختلف در سطوح مختلف براساس ویژگی‌های مختلف مطابق با وضعیت تحصیلی‌شان، در این مقاله پیشنهاد می‌شود. کار آینده ممکن است با افزایش تعداد ویژگی‌های مربوط به تحصیل برای تعیین صلاحیت دانش اشخاص و دسته‌بندی آن‌ها برطبق آن، انجام‌شود. علاوه‌بر این، ممکن است الگوریتم‌های مختلف طبقه‌بندی دیگر، برای بررسی تغییرات در معیارهای عملکرد و تحلیل توانایی دسته‌بندی طبقه‌بندهای مختلف، مورد استفاده قرارگیرند.