پیش‌بینی بی‌پاسخی در یک آمارگیری کارگاهی با استفاده از ترکیب روش‌های یادگیری ماشین

علیرضا رضایی [۱]، مجتبی گنجعلی [۲]، احسان بهرامی [۳]

تاریخ دریافت: ۹۹/۵/۲۰
تاریخ پذیرش: ۹۹/۱۱/۱

چکیده:

بی‌پاسخی در آمارگیری‌ها منبعی برای بروز خطا در نتایج آمارگیری است و سازمان‌های ملی آماری همواره به دنبال راهکارهایی برای کنترل و کاهش آن هستند. پیش‌بینی واحدهای نمونه‌گیری بی‌پاسخ در آمارگیری قبل از اجرای آمارگیری از جمله راهکارهایی است که می‌تواند کمک زیادی به کاهش و درمان بی‌پاسخی آمارگیری داشته باشد. با توسعه‌های اخیر فناوری و تسهیل در محاسبات پیچیده امکان به‌کارگیری روش‌های یادگیری ماشین، مانند درخت‌های رگرسیون و رده‌بندی یا ماشین بردار پشتیبان در بسیاری از مسائل از جمله پیش‌بینی بی‌پاسخی واحدهای نمونه‌گیری در آمارگیری‌ها فراهم‌شده است. ما در این مقاله ضمن مرور کلی روش‌های فوق، واحدهای نمونه‌گیری بی‌پاسخ را در یک آمارگیری کارگاهی با استفاده از آن‌ها پیش‌بینی خواهیم کرد و نشان خواهیم داد ترکیب روش‌های فوق دارای دقت بیشتری در پیش‌بینی درست بی‌پاسخی نسبت به هرکدام از روش‌ها است.

واژه‌های کلیدی: بی‌پاسخی، درخت رگرسیون و طبقه‌بندی، رگرسیون لوژستیک، ماشین بردار پشتیبان.

۱ مقدمه

آمارگیری‌های کارگاهی ابزار بسیار مهمی در تولید آمارهای رسمی کشورها هستند. آمارگیری کارگاهی بر اساس تعریف دائره‌المعارف روش‌های تحقیق آمارگیری (۲۰۰۸) [۹] آمارگیری است که ساختار، رفتار و خروجی‌های کارگاه را به‌جای افراد موردبررسی قرار می‌دهد. این آمارگیری‌ها همواره در معرض بروز بی‌پاسخی هستند و سازمان‌های آماری در مراحل مختلف طراحی، اجرا و استخراج آن‌ها به دنبال راهکارهایی برای کاهش و درمان بی‌پاسخی هستند. پیش‌بینی واحدهای نمونه‌گیری بی‌پاسخ قبل از اجرای آمارگیری راهکاری است که می‌تواند کمک زیادی در کاهش بی‌پاسخی آمارگیری داشته باشد. در یک آمارگیری اگر پرسشنامه آمارگیری برای یک واحد نمونه‌گیری تکمیل نشود بی‌پاسخی از نوع واحد و درصورتی‌که تعدادی از سؤالات پرسشنامه تکمیل نشود از نوع قلم خواهد بود. روش‌های رگرسیون لوژستیک یا پروبیت، روش‌های رایجی در شناسایی و تحلیل عوامل بروز بی‌پاسخی هستند. در این روش‌ها اغلب متغیر بی‌پاسخی به‌عنوان یک متغیر دو رده‌ای به‌صورت ۱ برای نشان دادن بی‌پاسخی و ۰ برای نشان دادن پاسخ در نظر گرفته می‌شود. در این حالت موارد عدم دسترسی به پاسخ‌گوی مطلع یا عدم همکاری پاسخ‌گو که منجر به بی‌پاسخی شده است یکسان در نظر گرفته می‌شود. تحقیقاتی در زمینه مدل‌بندی بی‌پاسخی در آمارگیری‌ها انجام‌شده است. دورانت و استیل [۶] برای مجموعه‌ای از آمارگیری‌های خانواری انگلستان با استفاده از مدل رگرسیون لوژستیک چند متغیره، عوامل مؤثر بر بروز بی‌پاسخی را بر عدم همکاری پاسخ‌گو و عدم دسترسی به پاسخ‌گوی مطلع بررسی کردند. سیلر [۱۶] عوامل مؤثر بر بروز بی‌پاسخی را با فرض دو رده‌ای بودن متغیر پاسخ برای آمارگیری گرایش کسب‌وکار در آلمان با استفاده از مدل بندی پویا تحلیل نمود. او از روش رگرسیون لوژستیک برای مدل‌بندی بی‌پاسخی استفاده کرد و به دلیل پانلی بودن داده‌ها و مستقل نبودن مشاهدات از یکدیگر و وجود ناهمگنی مشاهده ناپذیر از روش معادلات برآوردگر تعمیم‌یافته [۴] استفاده کرد. استفاده از روش‌های یادگیری ماشین [۵] مانند درخت‌های رگرسیون و رده‌بندی [۶] و ماشین بردار

[۱] گروه آمار دانشگاه شهید بهشتی، تهران، ایران، alireza.re85@gmail.com
[۲] گروه آمار دانشگاه شهید بهشتی، تهران، ایران.
[۳] گروه آمار دانشگاه شهید بهشتی، تهران، ایران.

[4] Generalized Estimating Equations (GEE)
[5] Machine learning
[6] Classification And Regression Trees (CART)
[7] Support Vector Machine (SVM)

پشتیبان [۷] در مدل بندی بی‌پاسخی با توجه به توسعه فناوری و تسهیل در انجام محاسبات پیچیده در دهه اخیر رو به افزایش بوده است. یادگیری ماشین بخشی از علوم رایانه است که هدف آن آموزش رایانه برای یادگیری و عمل بدون برنامه‌نویسی صریح است [۱۲]. به بیان دیگری نیز می‌توان گفت یادگیری ماشین به روش‌های محاسباتی اطلاق می‌شود که از تجربه برای بهبود عملکرد یا انجام پیش‌بینی‌های دقیق استفاده می‌کند. تجربه دلالت به اطلاعاتی دارد که قبل از انجام تحلیل وجود دارد و اغلب به‌صورت داده‌های الکترونیکی قابل‌دسترس هستند. یادگیری ماشین ارتباط زیادی با آمار و تحلیل داده‌ها دارد زیرا الگوریتم‌های آموزش‌گیرنده در آن از داده‌ها استفاده می‌کنند [۱۳]. یادگیری ماشین نقش کلیدی در مسائل کاربردی به‌خصوص در موارد رده‌بندی و پیش‌بینی دارد و استفاده از آن در سال‌های اخیر در آمار رسمی نیز توسعه‌یافته است. فیپس و تت [۱۵] با استفاده از مدل درخت رگرسیون، عوامل مؤثر بر بروز بی‌پاسخی را در آمارگیری آمارهای فرصت شغلی در آمریکا موردبررسی قرار دادند. ارپ و همکاران [۷] با استفاده از مدل درخت رگرسیون، بی‌پاسخی را در یک آمارگیری کشاورزی مورد تحلیل قرار دادند و همچنین در سال ۲۰۱۸ با استفاده از درخت رگرسیون به تحلیل بی‌پاسخی در آمارگیری‌های کارگاهی طولی پرداختند. کرچنر و سیگنورینو [۱۱] روش ماشین بردار پشتیبان را در پیش‌بینی بی‌پاسخی در یک آمارگیری خانواری به کار بردند و با مقایسه آن با روش مدل رگرسیون لوژستیک نشان دادند دقت ماشین بردار پشتیبان از رگرسیون لوژستیک بیشتر است. رضایی قهرودی و همکاران [۲] کاربرد روش‌های یادگیری آماری [۸] را در آمار رسمی موردبررسی قرار دادند. رضایی و همکاران [۱] با استفاده از روش درخت‌های رگرسیون و طبقه‌بندی به شناسایی عوامل مؤثر بر بروز بی‌پاسخی در یک طرح آمارگیری کارگاهی پرداختند و نشان دادند سازمان اجرای آمارگیری، بی‌پاسخی کارگاه در دوره قبلی آمارگیری و تعداد کارکن کارگاه تأثیر زیادی در بی‌پاسخی کارگاه‌ها دارد. ما در این مقاله ضمن معرفی آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیش‌تر مرکز آمار ایران با مرور روش‌های مدل رگرسیون لوژستیک، درخت‌های رگرسیون و رده‌بندی و ماشین بردار پشتیبان به پیش‌بینی بی‌پاسخی کارگاه‌ها در طرح مذکور با استفاده از این روش‌ها خواهیم پرداخت و نشان خواهیم داد دقت ماشین بردار پشتیبان در پیش‌بینی کارگاه‌های بی‌پاسخ از دو روش دیگر بیشتر است و علاوه بر این با پیشنهاد یک روش ترکیبی مبتنی بر سه روش فوق، دقت ماشین بردار پشتیبان را افزایش‌خواهیم داد و نتیجه خواهیم گرفت روش ترکیبی دارای مطلوبیت بیش‌تری در پیش‌بینی درست بی‌پاسخی کارگاه‌ها نسبت به سایر روش‌ها خواهد داشت.

۲ آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیش‌تر

آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیشتر مرکز آمار ایران، یکی از آمارگیری‌های مهم در نظام آماری ایران است که سالانه با هدف تأمین نیازهای حساب‌های ملی و منطقه‌ای کشور در حوزه صنعت انجام می‌شود. این آمارگیری از سال ۱۳۵۱ تاکنون همه‌ساله به‌جز سال‌های ۱۳۵۶ و ۱۳۵۷ اجراشده است [۳]. کارگاه‌ها در این آمارگیری بر اساس تعداد کارکن به دوطبقه کارگاه‌های صنعتی ۱۰ تا ۴۹ نفر کارکن و ۵۰ نفر کارکن و بیش‌تر تقسیم می‌شوند. تمام کارگاه‌های ۵۰ نفر کارکن و بیش‌تر آمارگیری می‌شوند و نمونه‌ای از کارگاه‌های ۱۰ تا ۴۹ نفر کارکن بر اساس یک طرح نمونه‌گیری احتمالی طبقه‌بندی‌شده انتخاب می‌شود و کارگاه‌های انتخاب‌شده در نمونه مورد آمارگیری قرار می‌گیرند. متغیرهای استان محل فعالیت کارگاه، فعالیت اصلی کارگاه برحسب طبقه‌بندی استاندارد بین‌المللی فعالیت‌های اقتصادی *ISIC* ویرایش چهارم و تعداد کارکن کارگاه برای طبقه‌بندی مورداستفاده قرار می‌گیرند. متغیر اصلی موردبررسی در این آمارگیری، ارزش‌افزوده کارگاه‌های صنعتی است که از تفاضل ارزش ستانده و ارزش مصرف واسطه به دست می‌آید. هر یک از متغیرهای ارزش ستانده و ارزش مصرف واسطه نیز دارای اجزایی هستند که در پرسشنامه به‌تفصیل از پاسخ‌گو پرسیده می‌شوند. پرسشنامه توسط مأمور آمارگیر با مصاحبه رودررو با مسئول کارگاه تکمیل می‌شود یا در اختیار کارگاه قرار داده می‌شود و کارگاه پس از تکمیل پرسشنامه، آن را در اختیار مأمور آمارگیری قرار می‌دهد. مدت‌زمان اجرای آمارگیری حدود ۴ ماه است و تعداد نمونه آن تا سال ۱۳۹۴ به‌طور متوسط ۱۶۰۰۰ کارگاه و در سال‌های ۱۳۹۵ به بعد به حدود ۲۲۰۰۰ نمونه افزایش‌یافته است. با توجه به تمام شماری کارگاه‌های صنعتی ۵۰ نفر کارکن و بیشتر و کسر نمونه‌گیری حداقل ۵۰ درصدی در بخش کارگاه‌های ۱۰ تا ۴۹ نفر کارکن در هر دوره آمارگیری، درصد هم‌پوشانی نمونه‌ها در سال‌های مختلف آمارگیری زیاد است و معمولاً تعداد زیادی از کارگاه‌ها همواره در نمونه انتخاب می‌شوند [۱]. نرخ پاسخ‌گویی این طرح در سال ۱۳۹۷ برابر ۷/۷۷ درصد است [۳]. این نرخ در آمارگیری کارگاه‌های صنعتی کانادا در سال ۲۰۱۹ برابر ۷/۸۵ درصد [۱۸] و در آمارگیری سال ۲۰۱۶ آمریکا برابر ۶۳ درصد [۱۹] بوده است.

[8] Statistical learning

۳ تعریف مسئله و راهکارهای در نظر گرفته‌شده برای حل آن

فرض کنید نمونه موردبررسی به‌صورت

$$U = (M_1, x_1), (M_2, x_2), \ldots, (M_n, x_n)$$

باشد که در آن M_i نشانگر بی‌پاسخی کارگاه به سؤالات پرسشنامه (۱: بی‌پاسخ برای همه سؤالات و ۰: پاسخ به همه یا بخشی از سؤالات) و $x_i' = (x_{i1}, x_{i2}, , x_{ip})$ بردار متغیرهای کمکی باشد که M_i به مقادیر آن‌ها وابسته باشد. مقادیر این متغیرهای کمکی قبل از اجرای آمارگیری برای تمام نمونه‌ها معلوم و در جدول ۱ عناوین آن‌ها آورده شده است. مسئله اصلی، تعیین مدل مناسب برای پیش‌بینی M_i با توجه به مقادیر x_i است. ما در این مقاله برای بررسی و حل مسئله از مدل رگرسیون لوژستیک، درخت‌های رگرسیون و رده‌بندی ($CART$) و ماشین بردار پشتیبان (SVM) و ترکیب این سه روش استفاده خواهیم کرد.

۱.۳ رگرسیون لوژستیک

در مدل رگرسیون لوژستیک فرض می‌شود که $(M_i|x_i) \sim B(1, p_i)$ و M_iها از یکدیگر مستقل هستند و نحوه ارتباط p_i با x_i به‌صورت $log\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = x_i'\beta$ است. با تشکیل تابع درستنمایی مشاهدات به‌صورت:

$$L(\beta|x) = \prod_{i \in N_1} \frac{exp(x_i'\beta)}{1+exp(x_i'\beta)} \prod_{i \in N_0} \frac{1}{1+exp(x_i'\beta)} \qquad (1)$$

که در آن N_0 و N_1 واحدهای باپاسخ و بی‌پاسخ را نشان می‌دهد، β و در نتیجه p_i از روش برآورد ماکسیمم درستنمایی به دست می‌آید و نهایتاً برآورد M_i با در نظر گرفتن مقدار ۱ برای p_i بزرگ‌تر یا مساوی ۰/۵ و ۰ برای مقادیر p_i کوچک‌تر از ۰/۵ حاصل می‌شود. در این روش دو فرض مطرح است. اول توزیع شرطی M_i به‌شرط مقادیر متغیرهای کمکی و استقلال بین آن‌ها و دوم رابطه خطی $log\frac{p_i}{1-p_i}$ با مقادیر متغیرهای کمکی. استفاده از چنین مدل‌هایی برای بررسی عوامل مؤثر بر بی‌پاسخی دارای معایبی است. رگرسیون لوژستیک ممکن است با مسائلی در نیکویی برازش مواجه باشد. علاوه بر این در کاربرد به‌خصوص در طراحی و اجرای آمارگیری‌ها رده‌بندی کارگاه‌ها به رده‌های قابل تفسیر بر اساس مخاطره بی‌پاسخی (مثلاً مخاطره زیاد، مخاطره متوسط و مخاطره پایین) اهمیت زیادی دارد که این نتیجه معمولاً یا این مدل‌ها به دست نمی‌آید. کارگاه‌هایی که در رده بی‌پاسخ یا باپاسخ بر اساس مدل تعیین‌شده قرار می‌گیرند اغلب دارای مقادیر متفاوتی نسبت به متغیرهای کمکی در نظر گرفته‌شده هستند به‌خصوص برای متغیرهای کمکی پیوسته‌ای که رابطه آن‌ها با نرخ پاسخ یک رابطه یکنوا نباشد و این مسئله در حالاتی که اثرات متقابل معناداری بین متغیرهای کمکی وجود دارد به‌مراتب پیچیده‌تر نیز خواهد شد. به همین دلیل تعیین عوامل مؤثر بر بی‌پاسخی بر اساس مدل رگرسیون لوژستیک به‌گونه‌ای که قابل تفسیر باشند کار مشکلی است [۱۵].

۲.۳ درخت‌های رگرسیون و رده‌بندی

درخت‌های رگرسیون و رده‌بندی($CART$)، روش آماری برای مسائل رده‌بندی یا پیش‌بینی است که مبتنی بر الگوریتم‌های ساخت درخت هستند [۱۴]. برایمن و همکاران [۴] روش‌شناسی را برای ساخت درخت‌های تصمیم با عنوان درخت‌های رگرسیون و رده‌بندی ارائه دادند. درخت رده‌بندی در حالاتی مورداستفاده قرار می‌گیرد که متغیر مورد نظر گسسته باشد و هدف آن تعیین رده‌ای است که متغیر هدف در آن قرار می‌گیرد. درصورتی‌که متغیر هدف پیوسته باشد از درخت رگرسیون با هدف پیش‌بینی مقدار آن استفاده می‌شود. $CART$ یک روش ناپارامتری است و در استفاده از آن نیازی به در نظر گرفتن توزیع‌های آماری خاص برای M_iها و استقلال بین آن‌ها نیست. وجود داده‌های پرت در متغیرهای کمکی تأثیر معناداری بر نتایج ندارد و همچنین تفسیر خروجی آن نسبت به مدل رگرسیون لوژستیک از مطلوبیت بیشتری برخوردار است. سه عنصر مهم در $CART$ وجود دارند. اول قواعد دسته‌بندی داده‌ها بر اساس مقادیر متغیرهای کمکی، دوم قواعد توقف و تشکیل شاخه‌های پایانی [9] و سوم پیش‌بینی متغیر مورد نظر در شاخه‌های پایانی. روش کار به این صورت است که ابتدا یک متغیر کمکی (مثلاً j-امین متغیر) به‌دلخواه انتخاب و درصورتی‌که پیوسته باشد با توجه به یک نقطه انشعاب (s) [10] از دامنه مقادیر آن به دو ناحیه $R_1(j,s)$ و $R_2(j,s)$ تقسیم می‌شود که در آن $R_1(j,s) = \{x_j | x_j \leq s\}$ و $R_2(j,s) = \{x_j | x_j > s\}$ و اگر متغیر کمکی گسسته با q مقدار ممکن باشد آنگاه به تعداد $2^{(q-1)} - 1$ افراز ممکن از q مقدار متغیر کمکی وجود خواهد داشت به‌طوری‌که هر افراز شامل دو گروه باشد. در هر ناحیه $p_{k,jsb}$ به‌صورت

$$\hat{p}_{k,jsb} = \frac{1}{N_b} \sum_{x_i \in R_b(j,s)} I(M_i = k)$$

برآورد می‌شود که در آن N_b تعداد مشاهداتی است که در ناحیه $R_{b(j,s)}, b = 1,2$ قرار می‌گیرند. سپس s و j طوری تعیین می‌شوند که مقدار

[9] Terminal nodes
[10] Split point
[11] Gini index
[12] Cross entropy

$\sum_{k=0}^{1}\hat{p}_{k,jsb}(1-\hat{p}_{k,jsb})$ مینیمم شود. عبارت -1 $\sum_{b=1}^{2}N_b\sum_{k=0}^{1}\hat{p}_{k,jsb}(1-\hat{p}_{k,jsb})$ شاخص جینی [11] نام دارد و می‌توان از معیارهای دیگر مانند آنتروپی $-\sum_{k=0}^{1}\hat{p}_{k,jsb}log\,\hat{p}_{k,jsb}$ تعریف می‌شود نیز استفاده مقطع [12] که به‌صورت $R_1(j,s)$ و $R_2(j,s)$ کرد. در ادامه برای هرکدام از ناحیه‌های مشخص‌شده مراحل فوق تکرار می‌شود تا نهایتاً بر اساس یک قاعده توقف، $|T|$ ناحیه $R_1,R_2,...,R_{|T|}$ یا شاخه پایانی به دست آید. برای تعیین قاعده توقف، تابع مخاطره درخت را به‌صورت

$$R(T)=\sum_{m=1}^{|T|}N_m\sum_{k=0}^{1}\hat{p}_{mk}(1-\hat{p}_{mk})$$

نشان می‌دهیم. یک قاعده توقف می‌تواند به این صورت باشد که مقدار کاهش در $R(T)$ به دلیل تجزیه شاخه‌های پایانی در هر مرحله نسبت به مرحله قبل از یک مقدار آستانه بیشتر باشد. این قاعده توقف مناسبی نیست زیرا مقدار کاهش در $R(T)$ نسبت به افزایش تعداد مراحل لزوماً یک تابع نزولی نیست و ممکن است مقدار کاهش در مراحل بعد از مرحله توقف بیشتر از مقدار آستانه باشد. برای رفع این مشکل می‌توان به این صورت عمل کرد که ابتدا یک درخت بزرگ T_0 با تعداد شاخه‌های زیاد را به‌گونه‌ای تشکیل می‌دهیم که تعداد مشاهدات در هر شاخه نهایی از یک مقدار تعیین‌شده (مثلاً ۵) بیش‌تر باشد، سپس آن را بر اساس ادغام شاخه‌های پایانی یا غیر پایانی هرس می‌کنیم. اگر T نشانگر هر درختی باشد که از هرس کردن T_0 به‌دست‌آمده باشد مسئله اصلی تعیین T به‌گونه‌ای است که مناسب‌ترین نیکویی برازش به داده‌ها را داشته باشد و دچار بیش برازشی نباشد. ازآنجایی‌که نیکویی برازش و بیش‌برازشی با تعداد شاخه‌های نهایی درخت رابطه دارد به دنبال یک نقطه تعادل بین مقدار نیکویی برازش و تعداد شاخه‌های نهایی خواهیم بود. این مقدار را با α (پارامتر پیچیدگی [13]) نشان می‌دهیم و معیار هزینه پیچیدگی [14] زیر را بر اساس $R(T)$ به‌صورت زیر تعریف می‌کنیم.

$$C_\alpha(T)=R(T)+\alpha|T| \qquad (۲)$$

در رابطه (۲) هر چه α بزرگ‌تر باشد اندازه درخت کوچک‌تر و نیکویی برازش آن کمتر خواهد بود و برعکس هرچه α کوچک‌تر باشد اندازه درخت بزرگ‌تر و بیش‌برازشی آن نیز بیشتر خواهد بود. مسئله در این حالت یافتن $T\in T_0$ طوری است که برای هر α ، $C_\alpha(T)$ مینیمم شود. برای هر مقدار دلخواه α می‌توان نشان داد که یک درخت یکتای T_α وجود دارد که $C_\alpha(T)$ را مینیمم کند. برای یافتن T_α ، از ادغام شاخه‌های نهایی T_0 شروع خواهیم کرد به این صورت که اگر شاخه‌ای در T_0 که ادغام می‌شود را با T_t و درخت حاصل از ادغام T_t در T_0 را با T_1 نشان دهیم آنگاه $T_1=T_0-T_t$ و T_1 را طوری تشکیل می‌دهیم که $C_\alpha(T_1)-C_\alpha(T_0)$ مینیمم شود. مینیمم کردن $C_\alpha(T_1)-C_\alpha(T_0)$ معادل با مینیمم کردن $R(t)-R(T_t)+\alpha(1-|T_t|)$ و در نتیجه $\frac{R(t)-R(T_t)}{|T_t|-1}$ خواهد بود. به همین ترتیب در مرحله s از هرس کردن به دنبال T_s هستیم که $C_\alpha(T_s)-C_\alpha(T_{s-1})$ مینیمم شود. قرار می‌دهیم $\alpha^{(s)}=\frac{R(t)-R(T_t^{s-1})}{|T_t^{s-1}|-1}$ و این فرآیند را تا جایی ادامه می‌دهیم که همه شاخه‌ها ادغام شوند. خروجی این فرآیند شامل دنباله متناهی از زیر درخت‌ها به‌صورت $T_0\supseteq T_1\supseteq\ldots\supseteq T_m$ و $0=\alpha^0\le\alpha^1\le\ldots\le\alpha^{m-1}\le\alpha^m$ خواهد بود که در آن T_m درختی است که شاخه ندارد. می‌توان نشان داد که T_α در این دنباله وجود دارد. احتمال بی‌پاسخی و پاسخ‌گویی در هر ناحیه به‌صورت $k(m)=M_i$ و $\hat{p}_{mk}=\frac{1}{N_m}\sum_{x_i\in R_m}I(M_i=k),k=0,1$ به‌صورت $arg\,max_k\,\hat{p}_{mk},m=1,\ldots,|T|$ برآورد می‌شود. اگر ضرر ناشی از خطای رده‌بندی در بین رده‌ها متفاوت باشد می‌توان یک ماتریس ضرر به‌صورت $L=\begin{bmatrix}L_{00} & L_{01}\\ L_{10} & L_{11}\end{bmatrix}$ تعریف کرد که در آن L_{01} مقدار ضرر ناشی از پیش‌بینی نادرست بی‌پاسخی و L_{10} مقدار ضرر ناشی از پیش‌بینی نادرست پاسخ‌گویی است و در $C_\alpha(T)$ تأثیر L_{01} و L_{10} را به‌صورت

$$C_\alpha(T)=\sum_{m=1}^{|T|}N_m(L_{01}+L_{10})\hat{p}_{m0}\hat{p}_{m1}+\alpha|T| \qquad (۳)$$

در نظر گرفت. برای جزئیات بیشتر به [۱۰] مراجعه نمایید.

۳.۳ ماشین بردار پشتیبان

ماشین بردار پشتیبان (SVM) یک الگوریتم یادگیری ماشین راهنماییده [15] است که داده‌ها را با آن برای طبقه‌بندی یا رگرسیون تحلیل می‌کنند. SVM یک روش غیر احتمالی است و با ایجاد یک یا چند ابر صفحه در فضای R^p حاصل از متغیرهای کمکی موجود در داده‌های آموزش، یعنی $x_i=(x_{i1},x_{i2},\ldots,x_{ip})$ ، فضای R^p را به دو زیر فضا یا دوطبقه تقسیم می‌کند به‌طوری‌که نزدیک‌ترین x_i در هر طبقه به ابر صفحه ایجادشده دارای بیشترین فاصله با آن باشد [۱۰] . اگر فرض کنید x_i ها بردارهایی در R^p هستند و M_i برابر با صفر را معادل ۱- در نظر بگیریم آنگاه $\hat{M}_i$ از رابطه زیر به دست می‌آید :

$$\hat{M}_i=\begin{Bmatrix}1 & w.x_i+b\ge 0\\ -1\equiv 0 & w.x_i+b<0)\end{Bmatrix},i=1,2,\ldots,n$$

[13] Complexity Parameter (cp)
[14] Cost complexity
[15] Supervised machine learning algorithm
[16] Support vectors
[17] Margin

x_i هایی که برای آن‌ها رابطه $w.x_i+b=\pm 1$ برقرار است، بردارهای پشتیبان [16] نام دارند، فاصله بین دو ابر صفحه‌ای که از این بردارهای پشتیبان می‌گذرد، کناره [17] نام دارد و ابر صفحه ایجادشده در SVM در بین این دو ابر صفحه‌ای در فاصله مساوی با هرکدام از آن‌ها قرار دارد. در این حالت کناره دارای بیشترین مقدار ممکن خود را دارد و برابر با $\frac{2}{||w||}$ است. برای ایجاد چنین ابر صفحه‌ای b ، w و ς_i طوری تعیین می‌شوند که عبارت

$$\frac{1}{2}||w||^2+C\sum_{i=1}^{n}\varsigma_i \qquad (۴)$$

با توجه به شروط

$$M_i(w.x_i+b)\geq 1-\varsigma_i \quad , \quad c_i\geq 0 \quad i=1,\ldots,n \qquad (۵)$$

مینیمم شود. $\varsigma_i=max\{0,1-M_i(w.x_i+b)\}$ متغیر به‌هم‌ریختگی است [18] و برای حالاتی به کار می‌رود که مشاهدات به دلیل تعدادی مشاهده مختل [19] به‌صورت کاملاً خطی جدا پذیر نیستند و ابر صفحه جداکننده، همه مشاهدات را به طبقات صحیح منتسب نمی‌کند (حالت کناره نرم [20]). C عددی نامنفی است و مقدار اهمیت ς_i ها را نشان می‌دهد. هر چه مقدار C بزرگ‌تر باشد تعداد مشاهداتی که به طبقات اشتباه منتسب می‌شوند کمتر خواهد بود و برعکس با C برابر با صفر ابر صفحه‌ای به دست خواهد آمد که هیچ مشاهده‌ای به‌درستی طبقه‌بندی نمی‌شود. با کمک تابع لاگرانژ، مینیم کردن رابطه (۴) با شروط (۵) به‌صورت

$$L(w,b,\alpha)=\frac{1}{2}||w||^2+C\sum_{i=1}^{n}\varsigma_i-\sum_{i=1}^{n}\alpha_i[M_i(w.x_i+b)-1] \qquad (۶)$$

تبدیل خواهد شد. برای سهولت در محاسبات مرتبط با رابطه (۶)، از مسئله دوگانه ولف استفاده می‌شود و در نتیجه α_i ها طوری تعیین می‌شوند که عبارت

$$\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_jM_iM_jK(x_i,x_j) \qquad (۷)$$

با توجه به شروط

$$\sum_{i=1}^{n}\alpha_iM_i=0 \quad , \quad 0\leq\alpha_i<C \quad i=1,\ldots,n \qquad (۸)$$

مینیمم شود و در ادامه w و b به ترتیب از روابط $w=\sum_{i=1}^{n}\alpha_iM_i.x_i$ و $b=\frac{1}{S}\sum_{i=1}^{S}M_i-w.x_i^*$ که در آن x_i^* ، بردارهای پشتیبان و S تعداد آن‌ها است، به دست خواهند آمد. $K(x_i,x_j)$ تابع هسته‌ای است و از آن به‌عنوان ترفندی برای حالاتی به‌جز کناره نرم که مشاهدات به‌صورت غیرخطی جدا پذیر هستند استفاده می‌شود. تابع‌های هسته‌ای متعددی وجود دارند و در بین آن‌ها تابع مبانی شعاعی [21] به‌صورت $K(x_i,x_j)=exp(-\gamma||x_i-x_j||)^2$ کاربرد بیشتری دارد. برای جزئیات بیشتر در ارتباط با مطالب فوق به [۱۷] ، [۱۰] و [۵] مراجعه نمایید.

۴ تحلیل وضعیت پاسخ‌گویی کارگاه‌ها بر اساس رگرسیون لوژستیک، $CART$ و SVM

پیش‌بینی وضعیت پاسخ‌گویی کارگاه‌ها قبل از اجرای آمارگیری بسیار سودمند است زیرا بر اساس آن، کارگاه‌هایی که احتمال پاسخ‌گویی آن‌ها پایین است شناسایی خواهند شد و سازمان آماری می‌تواند برای جلب پاسخ‌گویی آن‌ها برنامه‌ها و اقدامات مؤثرتری را برای جلب همکاری کارگاه‌ها پیش‌بینی و اجرا کند و بروز بی‌پاسخی را در اجرای آمارگیری کاهش دهد. برای تحلیل عوامل مؤثر بر بی‌پاسخی از نتایج طرح آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیشتر سال‌های ۱۳۹۷ و ۱۳۹۶ استفاده‌شده است. متغیرهای کمکی موردبررسی در جدول ۱ آورده شده‌اند. داده‌های آمارگیری به دو گروه داده‌های آموزش و آزمایش با نرخ به ترتیب ۷۵ درصد و ۲۵ درصد تقسیم شد. برای تشکیل مدل رگرسیون لوژستیک از تمامی اثرات اصلی و اثرات متقابل دوتایی در مدل استفاده شد. در مدل $CART$ برای رسیدن به حداکثر صحت متعادل شده دنباله‌ای از پارامترهای پیچیدگی مختلف (cp) در بازه $[0,0/1]$ بافاصله ۰/۰۰۰۱ مورداستفاده قرار گرفت. بیش‌ترین صحت متعادل شده با $cp=$ ۰/۰۰۱۵ به دست می‌آید و مقدار آن برابر با ۰/۶۲۴۲ خواهد شد. در روش SVM با در نظر گرفتن مقادیر C و γ به ترتیب برابر با ۱۲۸ و ۰/۲ بیشترین صحت متعادل شده به دست می‌آید. در جدول ۲ نحوه طبقه‌بندی وضعیت پاسخ‌گویی و در جدول ۳ شاخص‌های ارزیابی مدل برازش داده‌شده سه روش بر اساس ماتریس به‌هم‌ریختگی [22] و مقایسه آن‌ها با مقادیر واقعی آورده شده است. بر اساس نتایج جدول ۳ ملاحظه می‌شود درستی روش طبقه‌بندی بدون مدل برابر با $\frac{662}{865}=$ ۰/۷۶۵۳ است و با هر سه مدل، این عدد کمی بهبود می‌یابد و بیشترین مقدار آن در روش درخت رده‌بندی با عدد $\frac{63+621}{865}=$ ۰/۷۹۰۸ است. نرخ پیش‌بینی درست باپاسخی در روش درخت رده‌بندی از دو روش دیگر بیشتر و برابر با $\frac{621}{667}=$ ۰/۹۳۸۱ است. هرچند شاخص‌های مطرح‌شده معیارهایی برای ارزیابی نیکویی برازش روش‌ها هستند

[18] Slack variable
[19] Noisy data
[20] Soft margin
[21] Radial based function
[22] Confusion matrix

اما ازآنجایی‌که در مسئله ما شناسایی درست کارگاه‌های بی‌پاسخ از اهمیت بیشتری برخوردار است به همین دلیل شاخص نرخ پیش‌بینی درست بی‌پاسخی و شاخص درستی متعادل شده نسبت به سایر شاخص‌ها از اهمیت بیشتری برخوردار هستند و می‌توانند معیارهای مناسب‌تری برای ارزیابی روش‌ها باشد. بر این اساس، ماشین بردار پشتیبان نسبت به دو روش دیگر از ارجحیت بیشتری برخوردار است و در شناسایی کارگاه‌های بی‌پاسخ نسبت به دو روش دیگر بهتر عمل کرده است به‌طوری‌که نرخ پیش‌بینی درست بی‌پاسخی و درستی متعادل شده آن دارای بیشترین مقدار نسبت به دو روش دیگر و به ترتیب برابر با ۰٫۳۵۹۶ = $\frac{۷۳}{۲۰۳}$ و ۰٫۶۲۷۷ = $\frac{۷۳+۵۹۳}{۸۶۵}$ است.

جدول ۱: متغیرهای کمکی مورداستفاده در مدل‌ها

نام متغیر	مقادیر
EMPL	تعداد کارکنان کارگاه
ORG	سازمان‌های اجرای آمارگیری شامل ۳۲ سازمان کارگاه‌ها بر اساس استان محل استقرار آن‌ها به این سازمان‌ها منتسب می‌شوند. (کارگاه‌های صنعتی ثبت‌شده در سازمان بورس اوراق بهادار دارای سازمان اجرای آمارگیری جداگانه هستند.)
IND	گروه فعالیت اقتصادی کارگاه حاصل تجمیع کدهای دو رقمی *ISIC۴* شامل ۷ طبقه به شرح زیر: (۱) کدهای ۱۰، ۱۱ و ۱۲ ؛ (۲) کدهای ۱۳، ۱۴ و ۱۵ ؛ (۳) کدهای ۱۶، ۱۷، ۱۸ و ۳۱ ؛ (۴) کدهای ۱۹، ۲۰، ۲۱ و ۲۲ ؛ (۵) کدهای ۲۳ و ۲۴ ؛ (۶) کدهای ۲۵، ۲۸ و ۳۳ ؛ (۷) کدهای ۲۶، ۲۷ ، ۳۳ ؛ (۸) کدهای ۲۹، ۳۰
OWN_MGMT	نوع مالکیت و نحوه مدیریت کارگاه شامل ۳ طبقه به شرح زیر: (۱) مالکیت خصوصی، مدیریت خصوصی ؛ (۲) مالکیت خصوصی، مدیریت عمومی؛ (۳) مالکیت عمومی، مدیریت عمومی
R_PREV_SURV	وضعیت پاسخ‌گویی کارگاه در دوره قبلی آمارگیری شامل ۳ طبقه به شرح زیر: (۱) با پاسخ در آمارگیری قبلی ؛ (۲) بی‌پاسخ در آمارگیری قبلی به دلیل عدم همکاری یا عدم دسترسی پاسخ‌گو ؛ (۳) بی‌پاسخ در آمارگیری قبلی به دلیل عدم انتخاب در نمونه
CO	وجود دفتر مرکزی برای کارگاه شامل ۳ طبقه به شرح زیر: (۱) بدون دفتر مرکزی ؛ (۲) دارای دفتر مرکزی در محدوده شهری فعالیت کارگاه ؛ (۳) دارای دفتر مرکزی خارج از محدوده شهری فعالیت کارگاه
IZ	محل استقرار کارگاه شامل ۳ طبقه به شرح زیر: (۱) مستقر در شهرک یا ناحیه صنعتی؛ (۲) مستقر در مناطق آزاد ؛ (۳) مستقر در خارج از شهرک یا ناحیه صنعتی و مناطق آزاد

جدول ۲: نحوه طبقه‌بندی وضعیت پاسخ‌گویی سه روش و مقایسه آن‌ها با مقادیر واقعی

	پیش‌بینی رگرسیون لوژستیک			پیش‌بینی درخت رده‌بندی			پیش‌بینی ماشین بردار پشتیبان		
واقعی	بی‌پاسخ	با پاسخ	مجموع	بی‌پاسخ	با پاسخ	مجموع	بی‌پاسخ	با پاسخ	مجموع
بی‌پاسخ	۵۵	۱۴۸	۲۰۳	۶۳	۱۴۰	۲۰۳	۷۳	۱۳۰	۲۰۳
با پاسخ	۴۸	۶۱۴	۶۶۲	۴۱	۶۲۱	۶۶۲	۶۹	۵۹۳	۶۶۲
مجموع	۱۰۳	۷۶۲	۸۶۵	۱۰۴	۷۶۱	۸۶۵	۱۴۲	۷۲۳	۸۶۵

جدول ۳: شاخص‌های ارزیابی روش‌های در نظر گرفته‌شده

در این جدول درستی عبارت است از نسبت تعداد کارگاه‌ها با پیش‌بینی درست وضعیت پاسخ‌گویی به کل کارگاه‌ها و نرخ عدم اطلاع نشان‌دهنده درستی در حالتی است که بدون استفاده از هر مدلی، وضعیت پاسخ‌گویی همه کارگاه‌ها، باپاسخ در نظر گرفته شود و برابر نسبت کارگاه‌های باپاسخ به تعداد کل کارگاه‌ها است. نرخ پیش‌بینی درست باپاسخی (بی‌پاسخی) عبارت است از نسبت تعداد کارگاه‌های با پیش‌بینی صحیح با پاسخ(بی‌پاسخ) به تعداد کل کارگاه‌های باپاسخ (بی‌پاسخ) و درستی متعادل شده نیز برابر میانگین نرخ پیش‌بینی درست باپاسخی و بی‌پاسخی است

شاخص	رگرسیون لوژستیک	درخت رده‌بندی	ماشین بردار پشتیبان
درستی (*Accuracy*)	۰/۷۷۳۴	۰/۷۹۰۸	۰/۷۶۹۹
نرخ عدم اطلاع (*No Information Rate*)	۰/۷۶۵۳	۰/۷۶۵۳	۰/۷۶۵۳
نرخ پیش‌بینی درست باپاسخی (*Sensitivity*)	۰/۹۲۷۵	۰/۹۳۸۱	۰/۸۹۵۸
نرخ پیش‌بینی درست بی‌پاسخی (*Specificity*)	۰/۲۷۰۹	۰/۳۱۰۳	۰/۳۵۹۶
درستی متعادل شده (*Balanced Accuracy*)	۰/۵۹۹۲	۰/۶۲۴۲	۰/۶۲۷۷

۵ ترکیب روش‌های فوق برای بهبود نرخ پیش‌بینی درست بی‌پاسخی

در بخش قبل ملاحظه شد روش ماشین بردار پشتیبان دارای دقت بیشتری در پیش‌بینی درست بی‌پاسخی نسبت به دو روش دیگر است. با مدل ماشین بردار پشتیبان از هر ۱۰۰ کارگاه بی‌پاسخ حدود ۳۵ کارگاه به‌درستی به‌عنوان کارگاه بی‌پاسخ شناسایی می‌شوند و به نظر می‌رسد این مقدار در کاربرد نمی‌تواند مطلوبیت قابل‌توجهی داشته باشد. می‌توان با تعریف تابع مخاطره جدید و در نظر گرفتن مقادیر ضرر بزرگ‌تر برای پیش‌بینی نادرست بی‌پاسخی این نسبت را افزایش داد اما از طرف دیگر مقدار پیش‌بینی نادرست باپاسخی و در نتیجه درستی متعادل شده کاهش می‌یابد. ما در این مقاله برای بهبود پیش‌بینی درست بی‌پاسخی از مدل ترکیبی به‌صورت زیر استفاده کردیم.

$$\hat{M}_i = \begin{cases} ۱ & M_i^{SVM} = ۱ \, or \, M_i^{CART} = ۱ \, or \, M_i^{LOGREG} = ۱ \\ ۰ & o.w \end{cases}$$

که در آن M_i^{SVM} ، M_i^{CART} و M_i^{LOGREG} برآورد وضعیت پاسخ‌گویی کارگاه بر اساس مدل‌های ماشین بردار پشتیبان، درخت رده‌بندی و رگرسیون لوژستیک است. جداول ۴ و ۵ به ترتیب ماتریس به‌هم‌ریختگی و شاخص‌های ارزیابی این روش را در مقابل روش ماشین بردار پشتیبان نشان می‌دهند. همان‌طوری که در جداول ۴ و ۵ ملاحظه می‌شود نرخ پیش‌بینی درست بی‌پاسخی و درستی متعادل شده روش ترکیبی از روش ماشین بردار پشتیبان بیشتر است. در جدول ۶ نتایج اعتبار سنجی متقابل $fold - ۵$ روش ترکیبی آورده شده است. برای اعتبار سنجی از میانگین و ضریب تغییرات شاخص‌های جدول ۵ حاصل از ۵ بار تکرار استفاده‌شده است.

۶ نتیجه‌گیری

هدف این مقاله انتخاب روش مناسب برای پیش‌بینی وضعیت پاسخ‌گویی کارگاه‌ها با استفاده از مدل‌های رگرسیون لوژستیک، *CART* و *SVM* بود. بر اساس نتایج حاصل در پیش‌بینی وضعیت بی‌پاسخی کارگاه، روش *SVM* دارای کارایی بهتری است به‌طوری‌که مقدار نرخ پیش‌بینی درست بی‌پاسخی و درستی متعادل شده آن از سایر روش‌ها بیشتر بود. این روش یک روش ناپارامتری است و در استفاده از آن نیازی به فرض‌های توزیع آماری و استقلال بین مشاهدات نیست. بر اساس نتایج جدول ۳ دقت روش رگرسیون لوژستیک در پیش‌بینی درست بی‌پاسخی از دو روش دیگر کمتر است و هرچند ممکن است با در نظر گرفتن اثرات متقابل مرتبه بالاتر در آن به نتایجی بهتر از *CART* و *SVM* دست‌یافت اما فرض‌های خطی بودن لوجیت احتمال بی‌پاسخی، استقلال مشاهدات و تفسیر خروجی مدل مسئله مهمی است که مطلوبیت استفاده از مدل رگرسیون لوژستیک را در مقایسه با *CART* و *SVM* کاهش می‌دهد. *CART* نیز مانند *SVM* روش ناپارامتری است و انجام محاسبات آن دارای زمان کمتری نسبت به *SVM* است. ازآنجایی‌که پیش‌بینی درست بی‌پاسخی در این مسئله بسیار مهم است می‌توان با ترکیب روش‌ها نتایج بهتری را برای پیش‌بینی درست بی‌پاسخی به دست آورد. در این مسئله ترکیب روش‌ها منجر به افزایش حدود ۱۰ درصدی در نرخ پیش‌بینی درست بی‌پاسخی خواهد شد و درستی متعادل شده نیز حدود ۳ درصد بهبود می‌یابد. بر اساس مدل حاصل از *SVM* یا روش ترکیبی پیشنهادشده می‌توان قبل از آمارگیری، کارگاه‌هایی که دارای احتمال زیادی در بی‌پاسخی هستند را شناسایی و تمهیدات لازم را برای جلب همکاری آن‌ها با برنامه‌ریزی و اقدامات مؤثر انجام داد و از این طریق به افزایش کیفیت نتایج آمارگیری کمک کرد.

جدول ۴: نحوه طبقه‌بندی وضعیت پاسخ‌گویی روش ماشین بردار پشتیبان و ترکیبی و مقایسه با مقادیر واقعی

	پیش‌بینی ماشین بردار پشتیبان			روش ترکیبی		
واقعی	بی‌پاسخ	با پاسخ	مجموع	بی‌پاسخ	با پاسخ	مجموع
بی‌پاسخ	۷۳	۱۳۰	۲۰۳	۹۰	۱۱۳	۲۰۳
با پاسخ	۶۹	۵۹۳	۶۶۲	۹۴	۵۶۸	۶۶۲
مجموع	۱۴۲	۷۲۳	۸۶۵	۱۸۴	۶۸۱	۸۶۵

جدول ۵: شاخص‌های ارزیابی روش‌های ماشین بردار پشتیبان و ترکیبی

شاخص	ماشین بردار پشتیبان	ترکیبی
درستی (*Accuracy*)	۰/۷۶۹۹	۰/۷۶۰۷
نرخ عدم اطلاع (*No Information Rate*)	۰/۷۶۵۳	۰/۷۶۵۳
نرخ پیش‌بینی درست باپاسخی (*Sensitivity*)	۰/۸۹۵۸	۰/۸۵۸۰
نرخ پیش‌بینی درست بی‌پاسخی (*Specificity*)	۰/۳۵۹۶	۰/۴۴۳۳
درستی متعادل شده (*Balanced Accuracy*)	۰/۶۲۷۷	۰/۶۵۰۷

جدول ۶: نتایج حاصل از اعتبار سنجی متقابل $fold-5$ روش ترکیبی

شاخص	میانگین	ضریب تغییرات
درستی (*Accuracy*)	۰/۷۶۹۱۴۱۹	۰/۰۰۸۹۸۸۳۴۴
نرخ عدم اطلاع (*No Information Rate*)	۰/۸۳۱۱۹۲۵	۰/۰۱۴۶۶۱۲۴۲
نرخ پیش‌بینی درست باپاسخی (*Sensitivity*)	۰/۸۷۶۴۰۷۵	۰/۰۰۹۱۴۹۴۲۹
نرخ پیش‌بینی درست بی‌پاسخی (*Specificity*)	۰/۴۱۹۹۹۶۶	۰/۰۶۰۰۰۱۷۰۱
درستی متعادل شده (*Balanced Accuracy*)	۰/۶۴۸۲۰۲۰	۰/۰۱۳۷۹۴۸۰۶

مراجع

[۱] رضایی، علیرضا؛ گنجعلی مجتبی و بهرامی، احسان (۱۳۹۹). شناسایی عوامل مؤثر بر بروز بی‌پاسخی در آمارگیری‌های کارگاهی با استفاده از درخت رده‌بندی. یزد: مجموعه مقالات پانزدهمین کنفرانس آمار ایران.

[۲] رضایی قهرودی، زهرا؛ رنجی، حسن و رضایی، علیرضا (۱۳۹۸). یادگیری آماری و کاربردهای آن در آمار رسمی. تهران: پژوهشکده آمار ایران.

[۳] مرکز آمار ایران (۱۳۹۸). نتایج طرح آمارگیری از کارگاه‌های صنعتی ۱۰ نفر کارکن و بیشتر سال ۱۳۹۶. تهران: مرکز آمار ایران.

[4] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J., (1984). *Classification and regression trees. Monterey*, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

[5] Cortez, C. and Vapnik, V., (1995). Support Vector Network, *Machine learning*. **20 (3)**, 273–297.

[6] Durrant, Gabriele, B. and Steele, F., (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society: series A (statistics in society)*, **172**

(2), 361-381.

[7] Earp, M., Mitchell, M., McCarthy, J. and Kreuter, F., (2014). Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey. *Journal of Official Statistics*, **30(4)**, 701–719.

[8] Earp, M., Toth, D., Phipps, P. and Oslund, C., (2018). Assessing Nonresponse in a Longitudinal Establishment Survey Using Regression Trees. *Journal of Official Statistics*,**34(2)**, 463–481.

[9] Encyclopedia of Survey Research Methods (2008). published online 2011, edited by P.J. Lavrakas.

[10] Hastie, T., Tibshirani, R. and Friedman, J., (2009). *The Elements of Statistical Learning*. 2th edition, Springer.

[11] Kirchner, A. and Signorino, C S., (2018). Using Support Vector Machines for Survey Research.*Survey Practice*, **11(1)**.

[12] Mitchell, T., M., (1997). *Machine Learning*. McGraw-Hill, p. 2.

[13] Mohri, M., Rostamizadeh, A. and Talwalkar, A., (2018). *Foundations of Machine Learning*, 2th edition. The MIT Press.

[14] Pham, B. and Prakash, I., (2018). Machine Learning Methods of Kernel Logistic Regression and Classification and Regression Trees for Landslide Susceptibility Assessment at Part of Himalayan Area, India. *Indian Journal of Science and Technology*, **11(12)**, 1-10.

[15] Phipps, P. and Toth, D., (2012). Analyzing Establishment Nonresponse Using an Interpretable Regression Tree Model with Linked Administrative Data. *The Annals of Applied Statistics*, **6(2)**, 772-794.

[16] Seiler, C., (2010). Dynamic Modelling of Nonresponse in Business Surveys. *Ifo Working Paper No. 93*.

[17] Shuzhan. (2018). Understanding the mathematics behind Support Vector Machines. from https://shuzhanfan.github.io/2018/05/understanding-mathematics-behind-support-vector-machines/

[18] Statistics Canada, (2020). Annual Survey of Manufacturing and Logging Industries (ASML), Detailed information for 2019. from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey amp;SDDS=2103

[19] U.S. Census Bureau, (2018). Annual Survey of Manufactures Methodology, from https://www.census.gov/programs-surveys/asm/technical-documentation/methodology.html