



نقش آمار در پژوهش‌های علمی با تاکید بر سرشماری‌های ثبتی مبنا

زهرا رضائی قهرودی
دانشگاه تهران
۲۸ مهرماه ۱۳۹۹

z.rezaeigh@ut.ac.ir

نقش آمار در مدرن سازی آمار رسمی

Modernization of official statistics

<https://www.unece.org/stats/mos.html>

High-level Group for the Modernization of Statistical Production and Services.

- Oversee development of frameworks,
- sharing information, tools and methods, which support the modernization of statistical organizations.

<https://statswiki.unece.org/display/hlgbas>

UNECE: United Nations Economic Commission for Europe

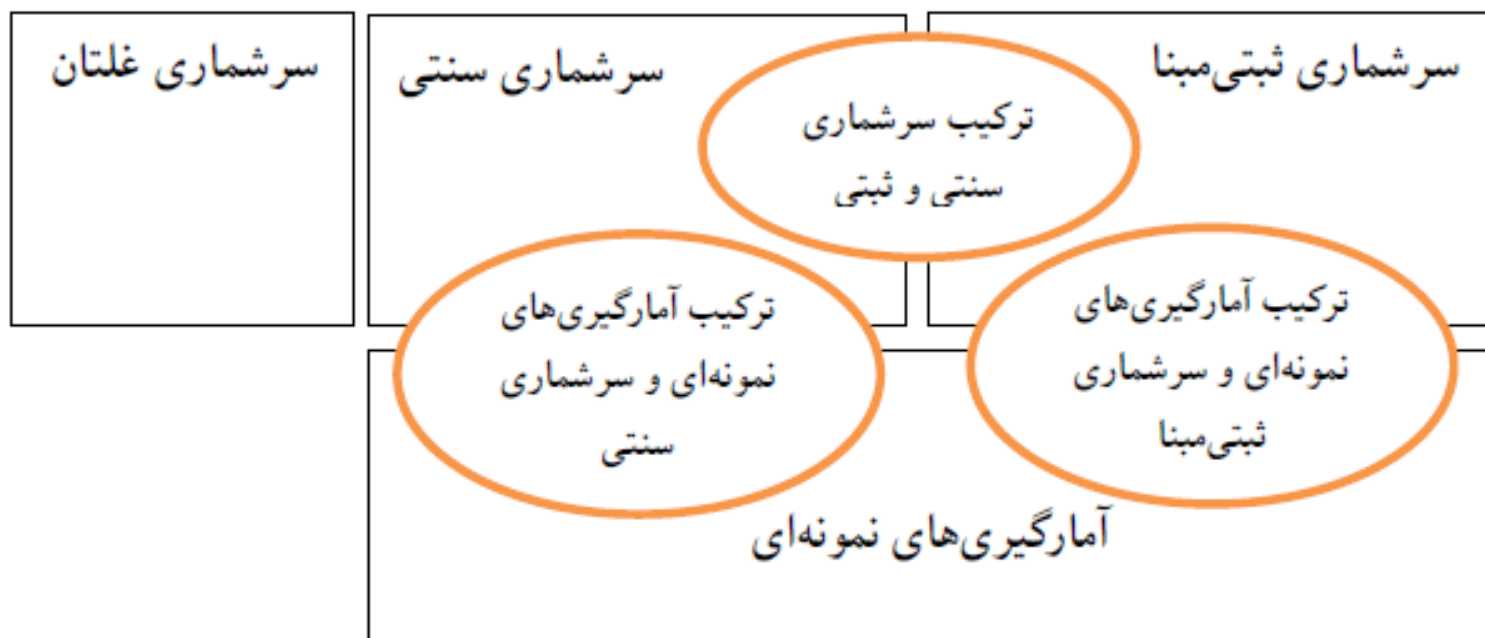
Machine Learning for Official Statistics

HLG-MOS page:

[High-Level Group for the Modernisation of Official Statistics](https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home)

<https://statswiki.unece.org/display/ML/Machine+Learning+for+Official+Statistics+Home>

روش‌های اجرای سرشماری‌ها



روش‌های به دست‌آوردن اطلاعات سرشماری برای هدف‌های آماری

وضعیت اجرای سرشماری در کشورهای عضو کمیسیون اقتصادی سازمان ملل برای اروپا UNECE



UNECE REGION



استرالیا
بلژیک
هلند
فنلاند
سوئد
نروژ
دانمارک
ایسلند

استونی
لتونی
لهستان
آلمان
اسپانیا
ترکیه

نقش آمار در اجرای سرشماری‌ها

What important information will we find out about dwellings?

- location
- type
- ownership
- time of construction (period)
- number of rooms
- usable floor area
- amenities (cold water supply, washing facilities, toilet, type of heating)

Why is a register-based census good for Estonia?

- Compared to conventional censuses, a register-based census saves the state an average of 25 x 1,300,000 minutes, or 541,667 hours, or 67,708 working days, or 280 full-time staff years.
- A register-based census enables receiving results faster than a conventional census, as it reduces the number of human errors made by enumerated persons, which would have to be corrected during the data validation period in a conventional census. If the same national databases are regularly used for censuses, it is possible to use standardised software for the validation and analysis of census data.
- A register-based census has lower costs, because it eliminates the need for census interviews which is the most labour- and cost-intensive stage of conventional censuses.
- The quality of a register-based census depends on the quality of data. High data quality should be an important objective for each national database.
- By holding a register-based census, Estonia once again confirms its status as one of the world's leading countries in the field of information technology.

Statistics Estonia
Tatari 51, 10134 Tallinn
tel +372 625 9300
stat@stat.ee

Databases and census characteristics

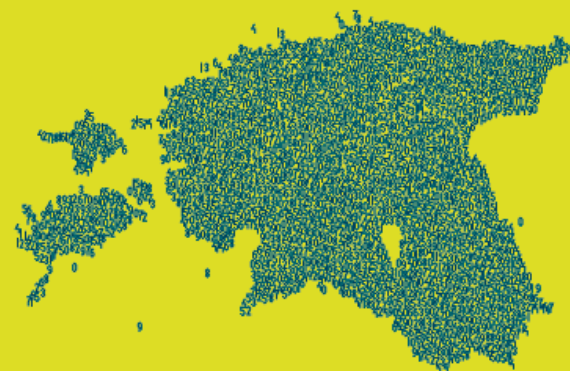
Characteristic	RR	ETR	EHIS	MKR	TÖR	EHR	KR	ARIREG	RKOARR	STAR	KIRST	KVKR	PKR	EMPIS	KIR	KOPIS	ADS	REL 2011
Permanent place of residence																		
Previous permanent place of residence																		
Sex																		
Age																		
Legal marital status																		
Country/place of birth																		
Citizenship																		
Relationships b/w household members																		
Location of job																		
Current activity status																		
Occupation																		
Economic sector																		
Status in employment																		
Level of education																		
Year of entering country																		
Tenure status of dwelling																		
Technical data on dwelling																		

RR – Population Register
ETR – Register of Residence and Work Permits
EHIS – Estonian Education Information System
MKR – Register of Residence and Work Permits
TÖR – Register of Employment
EHR – State Register of Construction Works
KR – Land Register
ARIREG – Commercial Register
RKOARR – State Register of State and Local Government Agencies
STAR – Register of Social Services and Benefits

KIRST – Health Insurance Database
KVKR – National Defence Obligation Register
PKR – State Pension Insurance Register
EMPIS – register of persons registered as unemployed and jobseekers, and of provision of labour market services
KIR – Register of Prisoners and Probationers
KOPIS – Register of Mandatory Funded Pension
ADS – Address Data System
REL 2011 – Population and Housing Census 2011



Register-based population and housing census



in Estonia

in 2020

نقش آمار در اجرای سرشماری‌ها

- تعیین تکلیف در خصوص اینکه چه اقلامی قرار است از طریق ثبت گردآوری شود
- چه اقلامی از طریق سرشماری آمیخته‌مد یا سرشماری‌های سنتی
- برای اقلامی که قرار است از طریق ثبت گردآوری شود ← اتصال رکوردی

اتصال رکوردی: اتصال داده‌ها در دو یا چند پایگاه‌داده با هدف یکپارچه‌سازی اطلاعات و ویژگی‌های مربوط به یک واحد مشخص

هدف اتصال رکوردی:

- ✓ حذف رکوردهای تکراری در پایگاه داده‌های مختلف
- ✓ یکی کردن اطلاعات رکوردهای مشابه در یک پایگاه داده نهایی

عملکرد اتصال رکوردی به این صورت است که واحدهایی که دارای رکوردهای مشترک هستند را پیدا کرده و ویژگی‌های متفاوتی که در پایگاه‌های داده متفاوت در مورد رکورد مشخص ثبت شده است را در کنار هم قرار داده و در نهایت یک پایگاه‌داده ترکیبی از رکوردهای مشترک و رکوردهای یکتا تشکیل دهیم.

سرشماری: سرشماری ثبتی مبنا

کارکرد مؤثر نظام ثبت‌های اداری از طریق:

- امکان برقراری پیوند بین رکوردهای مختلف از طریق کلید شناسایی یکتا

شناسه‌های یکتا: Record Linkage

- شناسه‌ی یکتای مکان
- شناسه‌ی یکتای فرد

اتصال هر فرد به آدرس یکتای مکان
پیونددهنده‌ی ثبت جمعیت و ثبت واحدهای مسکونی

اتصال قطعیته:

ناجور

جور

این نوع اتصال (یا عدم اتصال) زمانی اتفاق می‌افتد که کدشناسایی دقیق یک واحد در دو پایگاه داده مختلف پیدا شود و عمل اتصال اتفاق می‌افتد.

اشکالات:

- خطای در درج کدشناسایی هنگام پر کردن پرسشنامه، و یا وارد کردن در پایگاه‌داده، وجود بیش از یک کد شناسایی برای واحدهای مختلف (به عنوان مثال شماره شناسنامه و کد ملی
- گمشدگی در کدشناسایی

عدم وجود کد یکتای:

اتصال احتمالاتی

ناجور

جور

بر اساس متغیرهای کلیدی مشترک
(مانند نام، تاریخ تولد، آدرس و ...) بین
پایگاه داده‌های مختلف

تاریخچه اتصال رکوردی

- ✓ اتصال رکوردی اولین بار در سال ۱۹۵۰ و در حوزه پزشکی توسط هالبرت دان ظاهر شد
- ✓ در سال ۱۹۵۹ نیوکامب اتصال رکوردی احتمالاتی
- ✓ در سال ۱۹۶۹ فلگی و سانتر نظریه اتصال رکوردی
- ✓ در سال ۱۹۸۹ جارو روش‌های دیگر اتصال رکوردی احتمالاتی .
- ✓ در سال ۱۹۹۵ وینکلر
- ✓ در سال ۱۹۹۹ پگلیکا و سری روش اتصال رکوردی فاصله محور
- ✓ ...

مراحل اتصال رکوردی

✓ هدف از انجام اتصال رکوردی و شناسایی جامعه هدف

✓ آماده سازی برای انجام اتصال

- انتخاب متغیرهای کلیدی

- پاکسازی داده و بهبود کیفیت داده (اکثر داده‌ها دارای ناپاکی هستند و داده‌ها به فرمت‌های مختلف وارد شده‌اند: یکسان‌سازی فرمت تاریخ تولد، تبدیل تمامی حروف به عدد و ...)

- استانداردسازی متغیرهای کلیدی (ویرایش Editing، تجزیه Parsing، قالب‌بندی Formatting و مطابقت و هماهنگی concordance)

بخشی که شامل نام و نام خانوادگی است به دو بخش نام و نام خانوادگی تجزیه کرد.

کدنویسی: ° برای مرد و ۱ برای زن

✓ اتصال

- تصمیم در مورد روش‌های اتصال

- انجام اتصال و به دست آوردن یک پایگاه داده یکپارچه

✓ ارزیابی بعد از اتصال

۷۵ درصد

۵ درصد

۲۰ درصد

استانداردسازی قالب‌های
مختلف: 29/10/99 و با
10/29/99

مشکلات و خطاهای رایج در اتصال رکوردی

افراد و سازمان‌های مختلف هرگز به یک صورت به گردآوری داده نمی‌پردازند

نبود یک کدشناسایی واحد و یا وجود خطا در کد: عدم وجود کد شناسایی به علت گم‌شدگی، خطا در ورود کد درست هنگام وارد کردن داده در پایگاه داده، وجود بیش از یک کد شناسایی برای واحدهای مختلف (شماره شناسنامه و کد ملی) و ...

نام خانوادگی: تغییر نام یا نام خانوادگی طی زمان، املاهای متفاوت یک نام و یا نام خانوادگی (به عنوان مثال مهرورز و مهرور)

نام: املاهای متنوع نام (به عنوان مثال: سیاوش و سیاووش) و همچنین استفاده از نام مخفف شده (استفاده از نام علی به جای نام کامل تر علیرضا)

ادغام القاب در نام و یا نام خانوادگی: به عنوان مثال نوشتن دکتر، خانم، آقا، سرکار و ... در ابتدای نام و یا نام خانوادگی ممکن است به عنوان نام به کار برده است.

آدرس: آدرس در صورتی که دقیق و درست نوشته شده باشد یک متغیر بسیار مهم است. اما در بیشتر مواقع به علت عوض شدن آدرس، ناکامل بودن آدرس در پایگاه داده مختلف نمیتوان به آن اتکا کرد. گاه آدرس محل سکونت به جای آدرس محل کار و یا بالعکس در پرسشنامه ها نوشته میشود.

جنسیت: جنسیت نیز به علت ثبات آن در طول عمر یک متغیر قابل اعتماد است. استفاده از علائم مختلف برای جنسیت و یا عدم درج این متغیر در پایگاه داده‌ها

تاریخ تولد: تاریخ تولد نیز مانند جنسیت به علت ثباتی که در طول دوران زندگی افراد دارد، یک متغیر قابل اعتماد به شمار می‌آید. اما این متغیر هم دارای اشکالاتی است. به عنوان مثال در بعضی پایگاه داده ها فقط سال و یا فقط ماه و سال تولد پرسیده میشود. در بعضی پرسشنامه ها ماه تولد به حرف و در بعضی به عدد نوشته میشود (29 آذر و یا 29/9) ترتیب نوشتن ماه و روز هم در صورتی که در پایگاه های داده مختلف یکسان نباشد میتواند ایجاد مشکل کند (8/7/1399 و یا 7/8/1399)

سن: استفاده از حرف و یا عدد برای نوشتن سن

تغییر شغل و تغییر محل کار

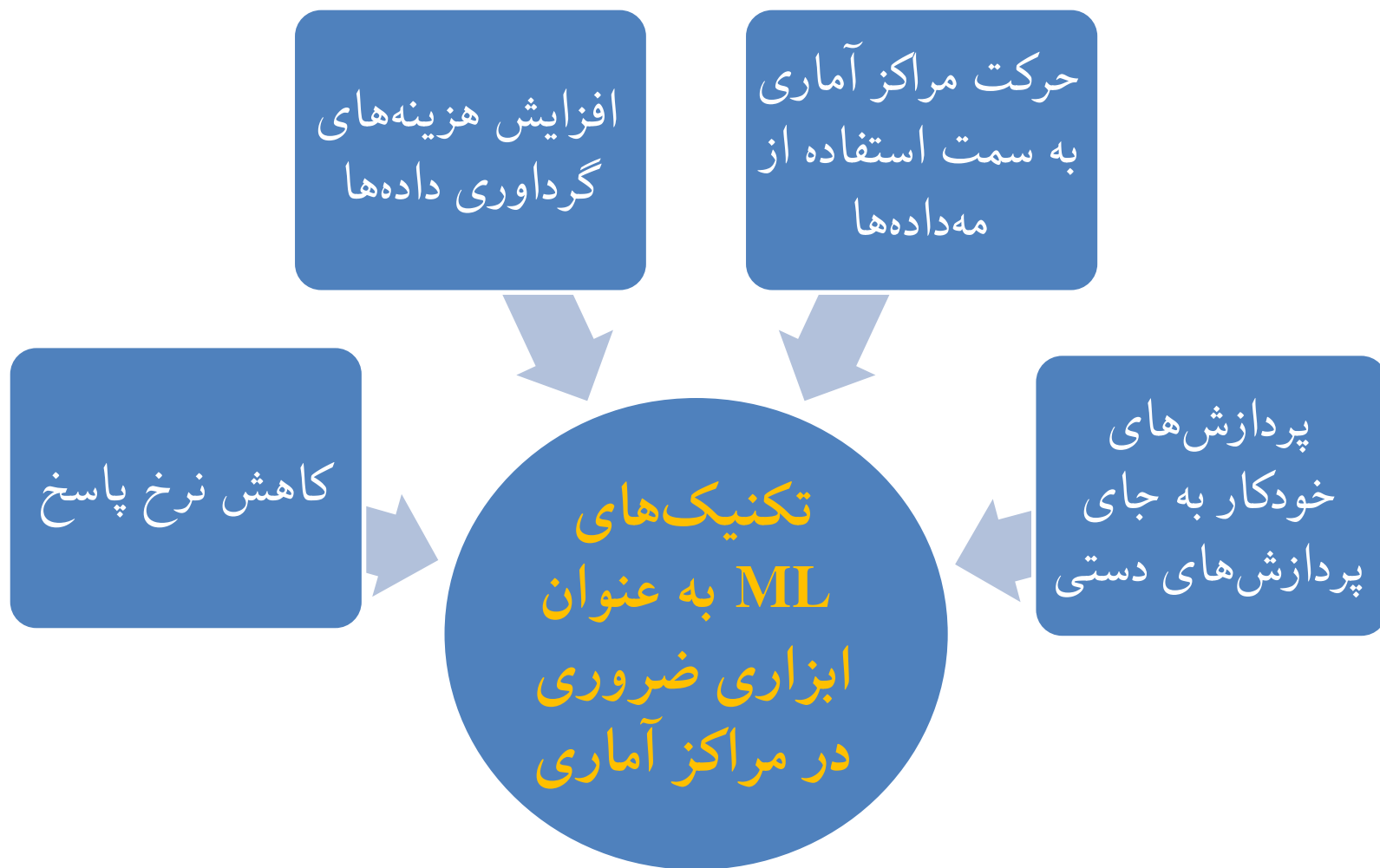
تغییر وضع تاهل، تعداد فرزندان، تغییر تعداد اعضای خانواده و ...

یادگیری ماشین در اتصال رکوردی

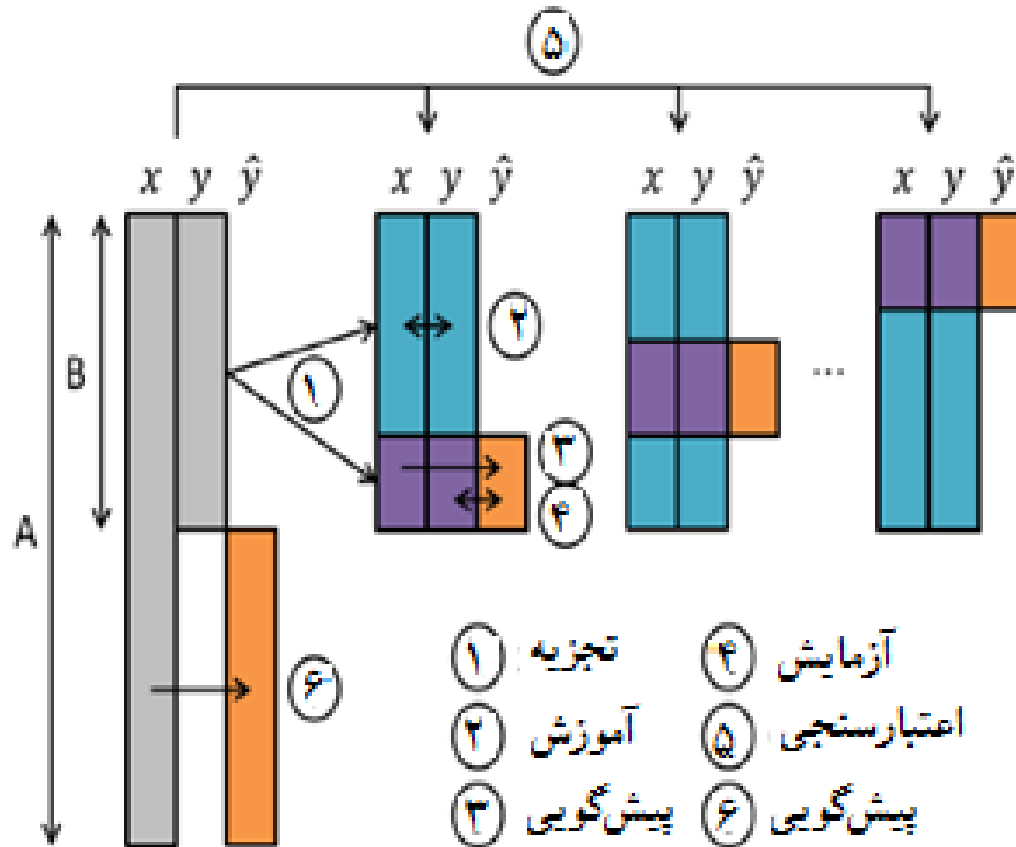
در سالهای گذشته با پیشرفت علم داده‌ها و اهمیت تصمیم‌گیری‌های داده‌محور، تعداد داده‌های سازمان‌های مختلف روز به روز بیشتر شده و اتصال رکوردی با روش‌های کلاسیک پاسخ‌گوی سیل عظیم پایگاه داده‌های مختلف که هر کدام شامل میلیاردها داده هستند نمی‌باشد.

روش‌های یادگیری ماشین که در اواخر دهه ۱۹۷۰ پدید آمدند، روش‌هایی برای فهم پایگاه داده‌های پیچیده‌تر را ارائه دادند. روش‌های یادگیری ماشین به کار تحلیل داده‌ها، در تمامی حوزه‌ها سرعت و دقت بخشیده است و اتصال رکوردی هم از این قاعده مستثنا نیست. روش‌های یادگیری ماشین، عمل یکپارچه‌سازی را به صورت خودکار انجام می‌دهد و به همین دلیل باعث کاهش هزینه و افزایش دقت می‌شود. برای مثال می‌توان به روش‌های رده‌بندی داده‌ها، بلوک‌بندی، استانداردسازی داده‌ها و... اشاره کرد. اتصال رکوردی با استفاده از روش‌های یادگیری ماشین مصداق بارز یک همکاری موثر است.

چرا یادگیری آماری در آمار رسمی مهم شده است



یادگیری آماری



شکل ۱- فرایند یادگیری آماری نیمه‌راهنماییده براساس داده‌های آموزشی و آزمایشی

مدل عمومی فرایند کسب و کار آماری GSBPM

ارزیابی	انتشار و اطلاع رسانی	تحلیل	پردازش	گردآوری	ساخت	طراحی	تعین نیازها
۱-۸: گردآوری اطلاعات برای ارزیابی	۱-۷: بهنگام سازی نظام های خروجی	۱-۶: اضافه کردن پیش نویس خروجی ها	۱-۵: جمع داده ها	۱-۴: ساخت چارچوب و انتخاب نمونه	۱-۳: ساخت ابزار گردآوری داده ها	۱-۲: طراحی خروجی ها	۱-۱: شناسایی نیازها
۲-۸: انجام ارزیابی	۲-۷: تولید محصولات انتشاراتی و اطلاع رسانی	۲-۶: تأیید خروجی ها	۲-۵: طبقه بندی و کدگذاری	۲-۴: تهیه و تنظیم برنامه ی گردآوری	۲-۳: ساخت یا ارتقای مؤلفه های پردازش	۲-۲: توصیف متغیرها	۲-۱: مشاوره و تأیید نهایی نیازها
۳-۸: ایجاد تفاهم در برنامه ی اقدام	۳-۷: مدیریت نشر محصولات	۳-۶: تفسیر و توضیح خروجی ها	۳-۵: بازیابی و اعتبارسنجی	۳-۴: گردآوری	۳-۳: ساخت یا ارتقای مؤلفه های انتشار و	۳-۲: طراحی ابزار و روش های گردآوری داده ها	۳-۱: تعیین خروجی ها
	۴-۷: ارتقای محصولات انتشاراتی و	۴-۶: به کارگیری روش های کنترل	۴-۵: ادیت و جانپی	۴-۴: نهایی کردن گردآوری	۴-۳: تنظیم جریان کار	۴-۲: طراحی چارچوب و نمونه	۴-۱: شناسایی مفاهیم
	۵-۷: مدیریت پشتیبانی از کاربران	۵-۶: نهایی کردن خروجی ها	۵-۵: استخراج متغیرها و واحدهای		۵-۳: آزمایش نظام تولید	۵-۲: طراحی پردازش و تحلیل های آماری	۵-۱: بررسی در دسترس بودن داده ها
			۶-۵: محاسبه وزن ها		۶-۳: آزمایش فرایندهای کسب و کار آماری	۶-۲: طراحی نظام تولید و جریان کار	۶-۱: تهیه ی طرح توجیهی کسب و کار
			۷-۵: محاسبه انبوه ها		۷-۳: نهایی کردن نظام تولید		
			۸-۵: نهایی کردن فایل های داده ای				

یادگیری آماری در آمار رسمی

کاربرد یادگیری آماری در فعالیت‌های مرتبط با داده‌های اولیه

فعالیت	روش‌های یادگیری آماری	زیرفرایند GSBPM
اتصال رکوردها	خوشه‌بندی	۱-۵، ۴-۲
کدگذاری	رده‌بندی	۲-۵، ۳-۴، ۴-۲
شناسایی داده‌های دورافتاده	خوشه‌بندی	۲-۶، ۱-۵، ۳-۴، ۴-۲
طبقه‌بندی	رده‌بندی	۶-۵، ۴-۵، ۳-۴، ۱-۴
برآورد	رگرسیون-رده‌بندی	۳-۴
جانبی	رگرسیون-رده‌بندی	۴-۵
کالبدین	رگرسیون-رده‌بندی	۶-۵
کنترل افشا	رگرسیون-رده‌بندی	۴-۶

کاربرد یادگیری آماری در فعالیتهای مرکز آمار ایران

مثال ۱: تحلیل و پردازش احساسات و عقاید مردم در نظرسنجی‌ها

مثال ۲: تعیین بار معنایی مثبت یا منفی پرسش‌های باز در آمارگیری‌های نمونه‌ای

مثال ۳: تشخیص واجد شرایط بودن یا نبودن کارگاه‌ها از متن‌کاوی پرسش باز «سایر دلایل عدم تکمیل پرسشنامه»

مثال ۴: تخصیص کد صحیح ISIC یا ISCO یا هر کد دیگر به پرسش‌های باز

مثال ۵: اتصال کدپستی با آدرس آماری به کمک روش‌های یادگیری آماری

نتیجه‌گیری

- با توجه به ضرورت مدرن‌سازی نظام آماری کشورها، از آن‌جا که استفاده از روش‌های یادگیری آماری در مدل GSBPM هم در داده‌های اولیه و هم در داده‌های ثانویه کاربرد دارد، امید است مرکز آمار ایران از روش‌های یادگیری آماری در هر یک مراحل مدل عمومی فرایند کسب و کار آماری، استفاده کند.
- کاربردهای مختلفی از روش‌های یادگیری آماری در آمار رسمی مانند اتصال رکوری، کدگذاری پرسش‌های باز، جورسازی داده‌ها، چارچوب‌سازی وجود دارد.
- بررسی تجربیات کشورهای مختلف و توصیه‌های بین‌المللی برای انجام سرشماری‌های ثبتي مبنا
- بررسی پایگاه‌داده‌های مختلف و تعیین اقلام نهایی ثبتي برای سال ۱۴۰۵
- آشنایی مرکز آمار ایران با روش‌شناسی اتصال رکوردها
- تشکیل تیم‌های کاری، هسته‌های پژوهشی، تعریف پروژه‌های مجزا با مشارکت مرکز آمار ایران و دانشگاه‌ها
- بررسی پایگاه‌داده‌های مختلف، پاکسازی داده‌ها، استانداردسازی، شناسایی ناسازگاری‌ها و رفع آنها و یا ارائه پیشنهاد برای رفع ناسازگاری‌ها، شناسی و حذف واحدهای تکراری
- تهیه دستورالعمل واحد گردآوری داده‌ها برای کاهش و جلوگیری از بروز مجدد خطاهای موجود در پایگاه داده‌ها
- انجام فرایند اتصال رکودی بین منابع اطلاعاتی و ارزیابی فرایند اتصال رکورد

References

زهرا رضائی قهرودی، حسن رنجی و علیرضا رضایی (۱۳۹۹). یادگیری آماری و کاربردهای آن در آمار رسمی. پژوهشکده آمار.

1-UNECE Machine Learning Team. (2018). The Use of Machine Learning in Official Statistics.

2- Gweon H., Schonlau M., Steiner S., Kaczmirek L. and Blohm M. (2017). Three Methods for Occupation Coding Based on Statistical Learning. Journal of Official Statistics, 33 (1), 101–122, <http://dx.doi.org/10.1515/JOS-2017-0006>.

3- Comber, S.; Arribas-Bel, D. (2019) Machine learning innovations in address matching: A practical comparison of word2vec and CRFs. Transactions in GIS, 23 (2): 334–348.

4- Newcombe, H. B., Kennedy, J., Axford, S. and James, A., 1959. Automatic linkage of vital records. Science, 130(3381), 954–959.

5- Diana Beltadze (2020). Developing methodology for the register-based census in Estonia, Statistical Journal of the IAOS. 36 , 159–164, DOI 10.3233/SJI-190585. IOS Press.

با سپاس از توجه شما

راه ارتباطی

z.rezaeigh@ut.ac.ir
zrezaeighahroodi@gmail.com