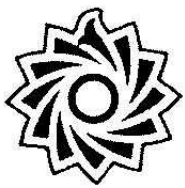
بسم الله الرحمن الرحیم



دانشگاه تربیت دبیر شهید رجائی

پاک‌سازی داده‌ها

رفع ناسازگاری، تکرار، مقدار ازدست‌رفته و اغتشاش

تألیف:

دکتر نگین دانشپور

عضو هیأت علمی دانشگاه تربیت دبیر شهید رجائی

مهدیه عطاییان

سرشناسه	:	دانشپور، نگین، ۱۳۵۵-
عنوان و نام پدیدآور	:	پاک‌سازی داده‌ها: رفع ناسازگاری، تکرار، مغدار از دست‌رفته و اغتشاش/ تالیف نگین دانشپور، مهدیه عطاییان؛ ویراستار ادبی ساغر سلمانی‌نژاد.
مشخصات نشر	:	تهران: دانشگاه تربیت دبیر شهید رجایی، ۱۴۰۰.
مشخصات ظاهری	:	۲۲۰ ص.: جدول.
شابک	:	۹۷۸-۶۲۲-۶۵۸۹-۲۰-۸
وضعیت فهرست نویسی	:	فیپا
یادداشت	:	واژه‌نامه.
یادداشت	:	کتابنامه: ص. [۱۸۶] - ۱۸۸.
یادداشت	:	نمایه.
موضوع	:	داده‌پردازی
موضوع	:	Electronic data processing
موضوع	:	داده‌کاوی
موضوع	:	Data mining
شناسه افزوده	:	عطاییان، مهدیه، ۱۳۷۱-
شناسه افزوده	:	دانشگاه تربیت دبیر شهید رجایی
شناسه افزوده	:	Shahid Rajaee Teacher Training University
رده بندی کنگره	:	QA۷۶
رده بندی دیویی	:	۰۰۵
شماره کتابشناسی ملی	:	۸۵۲۱۹۸۸
اطلاعات رکورد کتابشناسی	:	فیپا

عنوان	:	پاک‌سازی داده‌ها: رفع ناسازگاری، تکرار، مغدار از دست رفته و اغتشاش
تألیف	:	دکتر نگین دانشپور، عضو هیأت علمی دانشگاه تربیت دبیر شهید رجائی/ مهدیه عطاییان
ویراستار ادبی	:	دکتر ساغر سلمانی‌نژاد
نوبت چاپ	:	اول- زمستان ۱۴۰۰
انتشارات	:	دانشگاه تربیت دبیر شهید رجائی
لیتوگرافی، چاپ	:	رجاء نقشینه، شریف
طراح جلد	:	محمد معتمدی‌نژاد
ناظر چاپ	:	محمد معتمدی‌نژاد
صفحه‌آرا	:	نیره فیروزی
شمارگان	:	۱۰۰ جلد
قیمت	:	۶۰۰،۰۰۰ ریال
شابک	:	۹۷۸-۶۲۲-۶۵۸۹-۲۰-۸

ISBN: 978-622-6589-20-8

نشانی: تهران، لویزان، کد پستی ۱۶۷۸۸-۱۵۸۱۱، صندوق پستی ۱۶۳ – ۱۶۷۸۵، تلفن: ۹ (۲۶۳۲ – ۲۲۹۷۰۰۶۰)،
۲۲۹۷۰۰۷۰، تلفکس: ۲۲۹۷۰۰۴۲، پست الکترونیکی: publish@sru.ac.ir، وب سایت: http://publish.sru.ac.ir

پیش‌گفتار

داده‌های جمع‌آوری‌شده از منابع مختلف و توزیع‌شده برای اتخاذ تصمیمات مفید و سودمند باید به اطلاعات و دانش تبدیل شوند. به‌طور سنتی استخراج دانش را تحلیل‌گران داده انجام می‌دهند اما با توجه به رشد روزافزون داده‌ها، نیازمند روش‌هایی مبتنی بر رایانه هستیم. فرآیندهای مبتنی بر رایانه برای کشف دانش را «داده‌کاوی» می‌گویند. کیفیت داده‌های مورد بررسی در هنگام استخراج دانش، اهمیت بسزایی دارد. صحت، کامل‌بودن و سازگاری از جمله معیارهای مورد بررسی در بحث کیفیت داده‌ها هستند. داده‌های دنیای واقعی ممکن است به دلایل مختلفی دچار مشکل عدم کیفیت شوند. مقادیر جاافتاده، داده‌های مغشوش، ناسازگاری و مقادیر تکراری از جمله مشکلات عمده داده‌ها هستند. از این‌رو پیش‌پردازش که اغلب به منظور رفع این مشکلات انجام می‌شود، یکی از مهم‌ترین مراحل کشف دانش است.

پیش‌پردازش فرایندی زمان‌بر است اما با توجه به هزینه‌هایی که داده‌های فاقد کیفیت برای سازمان‌ها ایجاد می‌کنند، فرآیندی اجتناب‌ناپذیر است. درنهایت که دانش مورد نیاز برای اتخاذ تصمیمات، از داده‌های فاقد کیفیت استخراج شود، منجر به شکست تصمیمات می‌شود. پاک‌سازی و یکپارچه‌سازی از مراحل پیش‌پردازش هستند. در این کتاب تمرکز بر روی روش‌های جدید پاک‌سازی داده‌ها است. روش‌های مورد بررسی بر روی داده‌های عددی، ترتیبی و اسمی متمرکز هستند. در سال‌های اخیر پاک‌سازی داده‌ها از اهمیت ویژه‌ای برخوردار شده است، زیرا الگوریتم‌های داده‌کاوی برای عملکرد صحیح و ارائه دانش مفید، پیش‌بینی‌ها با توصیف‌ها، نیاز به داده‌های صحیح، کامل و سازگار دارند. در فرایند پاک‌سازی داده‌ها هدف، تشخیص داده‌های مغشوش، تکراری، پرت و مقادیر جاافتاده است.

در این کتاب به معرفی مفاهیم اولیه پیش‌پردازش داده‌ها و ابعاد مختلف آن می‌پردازیم. همچنین مشکلات کیفیت داده‌ها به صورت اجمالی بیان می‌شوند و راهکارهای ابتدایی برای پاک‌سازی داده‌ها به‌طور مختصر معرفی می‌شوند. تمرکز اصلی در این کتاب بر روی شرح الگوریتم‌های پیشنهادی نویسندگان در حوزه پاک‌سازی داده‌ها است. کارایی این الگوریتم‌ها به دقت ارزیابی شده و برتری آن‌ها در مقایسه با کارهای معتبر جدید و قابل قیاس در آن حوزه تأیید شده است.

به این منظور بر روی این الگوریتم‌ها آزمایش‌های متعدد در شرایط مختلف انجام داده‌ایم. بر اساس آزمایش‌ها و ارزیابی‌های انجام‌شده، این نتیجه حاصل شد که این الگوریتم‌ها نسبت به الگوریتم‌های مشابه عملکرد بهتری دارند و بر این اساس در این کتاب ارائه کردیم.

با توجه به این‌که پاک‌سازی داده‌ها امری اجتناب‌ناپذیر در تحلیل داده‌ها است و در سال‌های اخیر این امر از اهمیت ویژه‌ای برخوردار شده است، هدف اصلی از ویرایش این کتاب ارائه یک نسخه فارسی متمرکز در این حوزه است. اگرچه منابع انگلیسی متعددی در زمینه پیش‌پردازش و کیفیت داده‌ها موجود است، اما منبع فارسی متمرکزی در حوزه پاک‌سازی داده‌ها موجود نیست. این کتاب می‌تواند برای محققان یا تحلیلگران داده در دانشگاه‌ها و مراکز تحقیقاتی، مفید واقع شود.

فهرست مطالب