



به نام خدا

# ملزومات آماری در علم داده ها

شامل بیش از ۵۰ کد R و پایتون

مؤلفان:

**ساره هرمرزان** (کارشناسی ارشد علوم کامپیوتر از دانشگاه تهران)

**محمد رضاپور** (عضو هیأت علمی وزارت علوم، تحقیقات و فناوری)

**سینا توکلی بنیزی** (کارشناسی مهندسی صنایع)



هرگونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می گیرند.

## ◀ عنوان کتاب: ملزومات آماری در علم داده ها

شامل بیش از ۵۰ کد R و پایتون

◀ مولفان: ساره هرمزان - محمد رضاپور - سینا توکلی بنیزی

◀ ناشر: موسسه فرهنگی هنری دیباگران تهران

◀ ویراستار: نرگس مهرید

◀ صفحه آرایی: نازنین نصیری

◀ طراح جلد: داریوش فرسای

◀ نوبت چاپ: اول

◀ تاریخ نشر: ۱۴۰۱

◀ چاپ و صحافی: صدف

◀ تیراژ: ۱۰۰ جلد

◀ قیمت: ۲۴۰۰۰۰۰ ریال

◀ شابک: ۹۷۸-۶۲۲-۲۱۸-۶۰۰-۵

نشانی واحد فروش: تهران، میدان انقلاب،

خ کارگر جنوبی، روبروی پاساژ مهستان،

پلاک ۱۲۵۱-تلفن: ۶۶۴۱۰۰۴۶-۲۲۰۸۵۱۱۱

فروشگاههای اینترنتی دیباگران تهران :

[WWW.MFTBOOK.IR](http://WWW.MFTBOOK.IR)

[www.dibagarantehran.com](http://www.dibagarantehran.com)

سرشناسه: هرمزان، ساره، ۱۳۶۸-

عنوان و نام پدیدآور: ملزومات آماری در علم داده ها: شامل بیش از ۵۰ کد R و پایتون / مولفان: ساره هرمزان، محمد رضاپور، سینا توکلی بنیزی؛ ویراستار: نرگس مهرید.

مشخصات نشر: تهران: دیباگران تهران: ۱۴۰۱

مشخصات ظاهری: ۲۵۴ ص: مصور، جدول، نمودار

شابک: ۹۷۸-۶۲۲-۲۱۸-۶۰۰-۵

وضعیت فهرست نویسی: فیبا

یادداشت: ص.ع. به انگلیسی Sareh Hormozan, Mohammad Rezapour, Sina Tavakoli

Essentials of statistical in data science with more than ۵۰ codes in R and Python.

یادداشت: کتابنامه: ص. ۳۵۲-۳۵۳

موضوع: آمار ریاضی - داده پرداز

موضوع: mathematical statistics- data processing

موضوع: نرم افزار آر (computer software) R

موضوع: پایتون (زبان برنامه نویسی کامپیوتر)

موضوع: python (computer program language)

شناسه افزودن: رضاپور، محمد، ۱۳۵۷-

شناسه افزوده: توکلی بنیزی، سینا، ۱۳۷۷-

رده بندی کنگره: ۲۷۶/۶ QA

رده بندی دیویی: ۵۱۹/۵۰۲۸۵

شماره کتابشناسی ملی: ۸۹۶۰۵۵۱

نشانی تلگرام: @mftbook      نشانی اینستاگرام دیبا dibagaran\_publishing

هر کتاب دیباگران، یک فرصت جدید علمی و شغلی.

هر گوشه همراه، یک فروشگاه کتاب دیباگران تهران.

از طریق سایتهای دیباگران، در هر جای ایران به کتابهای ما دسترسی دارید.

این کتاب با کاغذ حمایتی منتشر شده است

## فهرست مطالب

مقدمه ناشر ..... ۱۲

پیش‌گفتاری بر «ملزومات آماری در علم داده‌ها» ..... ۱۳

### فصل صفر

کتاب‌شناسی و تقدیم به ... ..... ۱۵

قواعد مورد استفاده در این کتاب ..... ۱۶

اصطلاحات کلیدی و نمادهای استفاده‌شده در کتاب ..... ۱۷

استفاده از مثال‌های کد ..... ۱۷

تقدیم به ..... ۱۸

### فصل اول

تحلیل اکتشافی داده‌ها ..... ۱۹

عناصر داده‌های ساخت‌یافته ..... ۲۰

مطالعه بیش‌تر ..... ۲۲

داده‌های مستطیلی ..... ۲۳

دیتافریم‌ها و اندیس‌ها ..... ۲۴

ساختارهای داده‌های غیرمستطیلی ..... ۲۵

برآورد مکان ..... ۲۶

خلاصه‌سازی داده‌ها ..... ۲۷

میانگین ..... ۲۷

میان‌ه و برآوردهای مقاوم ..... ۲۸

داده‌های پرت ..... ۲۹

مثال: برآوردهای مکان جمعیت و نرخ جرم و جنایت ..... ۳۰

مطالعه بیش‌تر ..... ۳۱

برآورد پراکندگی ..... ۳۱

انحراف استاندارد و برآوردهای مربوط به آن ..... ۳۳

برآوردهای مبتنی بر صدک ..... ۳۵

مثال: برآوردهای پراکندگی جمعیت ایالتی ..... ۳۶

مطالعه بیش‌تر ..... ۳۷

۳۷.....	کشف توزیع داده
۳۸.....	صدک‌ها و نمودارهای جعبه‌ای
۴۰.....	جداول فراوانی و هیستوگرام‌ها
۴۲.....	نمودارهای چگالی و برآوردها
۴۴.....	مطالعه بیش‌تر
۴۵.....	کشف داده‌های دودویی و دسته‌ای
۴۷.....	داده‌های «عددی» به‌عنوان داده‌های «دسته‌ای»
۴۷.....	مد
۴۷.....	امید ریاضی
۴۸.....	احتمال
۴۸.....	همبستگی
۵۲.....	نمودارهای نقطه‌ای
۵۳.....	مطالعه بیش‌تر
۵۳.....	اکتشاف دو یا چند متغیر
۵۴.....	انبارک شش ضلعی و کانتورها (رسم داده‌های عددی در مقابل داده‌های عددی)
۵۶.....	دو متغیر دسته‌ای
۵۸.....	داده‌های دسته‌ای و عددی
۶۰.....	نمایش چند متغیر
۶۲.....	مطالعه بیش‌تر
۶۲.....	خلاصه

## فصل دوم

### ۶۳..... توزیع داده و نمونه‌گیری

۶۵.....	نمونه‌گیری تصادفی و بایاس نمونه
۶۶.....	بایاس نمونه‌گیری خودگزین
۶۷.....	بایاس
۶۸.....	انتخاب تصادفی
۶۸.....	اندازه در مقابل کیفیت: چه زمانی اندازه مهم است؟
۶۹.....	میانگین نمونه در مقابل میانگین جمعیت
۷۰.....	مطالعه بیش‌تر
۷۰.....	بایاس انتخاب
۷۲.....	رگرسیون به میانگین
۷۳.....	مطالعه بیش‌تر
۷۳.....	توزیع نمونه‌گیری یک آماره
۷۶.....	نظریه حد مرکزی

۷۶	خطای استاندارد.....
۷۷	مطالعه بیش تر .....
۷۸	بوت استرپ .....
۸۱	نمونه‌گیری مجدد در مقابل بوت استرپ .....
۸۱	مطالعه بیش تر .....
۸۲	بازه‌های اطمینان .....
۸۴	توزیع نرمال .....
۸۵	نمودارهای QQ و نرمال استاندارد .....
۸۷	توزیع‌های دم‌طولانی .....
۸۹	مطالعه بیش تر .....
۸۹	توزیع t دانش‌آموز .....
۹۱	مطالعه بیش تر .....
۹۱	توزیع دو جمله‌ای .....
۹۴	مطالعه بیش تر .....
۹۴	توزیع مربع-کای .....
۹۵	مطالعه بیش تر .....
۹۵	توزیع F .....
۹۵	مطالعه بیش تر .....
۹۶	پواسون و توزیع‌های مربوطه .....
۹۶	توزیع پواسون .....
۹۷	توزیع نمایی .....
۹۷	تخمین نرخ شکست (خرابی) .....
۹۸	توزیع ویبول .....
۹۹	مطالعه بیش تر .....
۹۹	خلاصه .....

## فصل سوم

۱۰۰	آزمایش‌های علمی و آزمون معناداری .....
۱۰۱	آزمون A/B .....
۱۰۴	چرا باید گروه کنترل داشته باشیم؟ .....
۱۰۵	چرا فقط A/B؟ چرا C/D، ... نه؟ .....
۱۰۶	مطالعه بیش تر .....
۱۰۶	آزمون‌های فرضیه .....
۱۰۷	فرض صفر .....
۱۰۸	فرضیه جایگزین .....

۱۰۸	آزمون‌های یک‌طرفه در مقابل آزمون‌های دوطرفه
۱۰۹	مطالعه بیشتر
۱۰۹	نمونه‌گیری مجدد
۱۱۰	آزمون جایگشت
۱۱۱	مثال: جذابیت وب
۱۱۵	آزمون‌های جایگشت بوت استرپ و کامل
۱۱۵	آزمون‌های جایگشت: خط پایان علوم داده
۱۱۶	مطالعه بیشتر
۱۱۶	معناداری آماری و مقادیر $p$
۱۱۹	مقدار $p$
۱۲۰	آلفا
۱۲۱	جدال بر سر مقدار $p$
۱۲۲	معناداری کاربردی
۱۲۲	خطاهای نوع ۱ و نوع ۲
۱۲۲	علوم داده و مقدار $p$
۱۲۳	مطالعه بیشتر
۱۲۳	آزمون‌های $t$
۱۲۵	مطالعه بیشتر
۱۲۵	آزمون چندگانه
۱۲۹	مطالعه بیشتر
۱۲۹	درجه آزادی
۱۳۰	ANOVA
۱۳۴	آماره $F$
۱۳۵	ANOVA دوطرفه
۱۳۶	مطالعه بیشتر
۱۳۶	آزمون مربع کای
۱۳۶	آزمون مربع کای: یک روش نمونه‌گیری مجدد
۱۳۹	آزمون مربع کای: نظریه آماری
۱۴۰	آزمون دقیق فیشر
۱۴۲	ارتباط برای علوم داده
۱۴۳	مطالعه بیشتر
۱۴۳	الگوریتم راهزن چنددست
۱۴۶	مطالعه بیشتر
۱۴۶	قدرت و اندازه نمونه
۱۴۸	اندازه نمونه

۱۵۰	مطالعه بیش تر .....
۱۵۰	خلاصه .....

## فصل چهارم

### رگرسیون و پیش‌بینی ..... ۱۵۱

۱۵۲	رگرسیون خطی ساده .....
۱۵۳	معادله رگرسیون .....
۱۵۶	مقادیر برازش شده و مانده‌ها .....
۱۵۸	حداقل مربعات .....
۱۵۹	پیش‌بینی در مقابل توضیح (پروفایلینگ) .....
۱۵۹	رگرسیون خطی چندگانه .....
۱۶۰	مثال: داده‌های خانه‌های کینگ کانتی .....
۱۶۲	ارزیابی مدل .....
۱۶۶	اعتبارسنجی متقابل .....
۱۶۷	انتخاب مدل و رگرسیون گام‌به‌گام .....
۱۷۱	رگرسیون وزن دار .....
۱۷۳	مطالعه بیش تر .....
۱۷۳	پیش‌بینی با استفاده از رگرسیون .....
۱۷۳	خطرات برون‌یابی .....
۱۷۴	بازه‌های اطمینان و پیش‌بینی .....
۱۷۵	متغیرهای عامل در رگرسیون .....
۱۷۶	نمایش متغیرهای ساختگی .....
۱۷۹	متغیرهای عامل با چندین سطح .....
۱۸۱	متغیرهای عاملی مرتب .....
۱۸۲	تفسیر معادله رگرسیون .....
۱۸۲	پیشگوهای همبسته .....
۱۸۴	هم‌خطی چندگانه .....
۱۸۴	متغیرهای مختلط .....
۱۸۶	تعامل‌ها و اثرات مهم .....
۱۸۸	تشخیص‌های رگرسیون .....
۱۸۹	داده‌های پرت .....
۱۹۱	مقادیر تأثیرگذار .....
۱۹۳	ناهم‌واربانی، نرمال نبودن و خطاهای همبسته .....
۱۹۶	نمودارهای مانده جزئی و غیرخطی بودن .....
۱۹۸	رگرسیون اسپیلاین و چندجمله‌ای .....

۱۹۹	چند جمله‌ای
۲۰۱	اسپیلاین
۲۰۲	مدل‌های جمعی تعمیم یافته
۲۰۴	مطالعه بیش تر
۲۰۴	خلاصه

## فصل پنجم

### طبقه‌بندی ..... ۲۰۵

۲۰۷	نایو بیز
۲۰۸	چرا طبقه‌بندی بیزین دقیق غیر کاربردی است؟
۲۰۸	راه حل نایو
۲۱۱	متغیرهای پیشگوی عددی
۲۱۲	مطالعه بیش تر
۲۱۲	تحلیل افتراقی
۲۱۳	ماتریس کوواریانس
۲۱۴	تفکیک کننده خطی فیشر
۲۱۴	مثال
۲۱۷	مطالعه بیش تر
۲۱۸	رگرسیون لجستیک
۲۱۸	تابع پاسخ لجستیک و لاجیت
۲۲۰	رگرسیون لجستیک و GLM
۲۲۱	مدل‌های خطی تعمیم یافته
۲۲۲	مقادیر پیش بینی شده رگرسیون لجستیک
۲۲۳	تفسیر ضرایب و نسبت شانس
۲۲۴	رگرسیون خطی و لجستیک: شباهت‌ها و تفاوت‌ها
۲۲۴	برازش مدل
۲۲۵	ارزیابی مدل
۲۲۷	تحلیل مانده‌ها
۲۲۹	مطالعه بیش تر
۲۲۹	ارزیابی مدل‌های طبقه‌بندی
۲۳۰	ماتریس اغتشاش
۲۳۲	مسأله کلاس نادر
۲۳۳	صحت، یادآوری و ویژگی
۲۳۴	منحنی ROC
۲۳۶	AUC



۲۳۷	لیفت
۲۳۹	مطالعه بیش تر
۲۳۹	استراتژی‌هایی برای داده‌های نامتوازن
۲۴۰	نمونه‌برداری کاهش‌ی
۲۴۲	نمونه‌برداری افزایشی یا وزن‌دهی افزایشی/کاهش‌ی
۲۴۳	تولید داده
۲۴۴	طبقه‌بندی مبتنی بر هزینه
۲۴۴	بررسی پیش‌بینی‌ها
۲۴۶	خلاصه

## فصل ششم

### ۲۴۷ یادگیری ماشین آماری

۲۴۹	k- نزدیک‌ترین همسایه
۲۵۰	مثال: پیش‌بینی نکول
۲۵۲	معیارهای فاصله
۲۵۳	رمزگذار وان هات
۲۵۴	استانداردسازی (نرمال‌سازی، امتیاز z)
۲۵۷	انتخاب K
۲۵۸	KNN به‌عنوان موتور ویژگی
۲۶۰	مدل‌های درختی
۲۶۲	مثال
۲۶۴	الگوریتم پارتیشن‌بندی بازگشتی
۲۶۶	اندازه‌گیری همگن بودن یا خلوص
۲۶۸	جلوگیری از رشد درخت
۲۶۸	کنترل پیچیدگی در R
۲۶۹	کنترل پیچیدگی درخت در پایتون
۲۶۹	پیش‌بینی یک مقدار پیوسته
۲۷۰	درخت‌ها چگونه استفاده می‌شوند؟
۲۷۱	بگینگ و جنگل تصادفی
۲۷۲	بگینگ
۲۷۳	جنگل تصادفی
۲۷۷	اهمیت متغیر
۲۸۰	فراپارامترها
۲۸۱	بوستینگ
۲۸۲	الگوریتم بوستینگ

۲۸۳	..... XGBOOST
۲۸۶	..... lightgbm
۲۸۸	..... Catboost
۲۸۹	..... اجتناب از بیش‌برازش
۲۹۴	..... فرآپارامترها و اعتبارسنجی متقابل
۲۹۷	..... خلاصه

## فصل هفتم

### ۲۹۸..... یادگیری بدون نظارت

۳۰۰	..... تحلیل مؤلفه اساسی
۳۰۱	..... مثال
۳۰۴	..... محاسبه مؤلفه‌های اساسی
۳۰۷	..... تحلیل تناظر
۳۰۹	..... خوشه‌بندی
۳۰۹	..... k-means خوشه‌بندی
۳۱۰	..... مثال
۳۱۳	..... الگوریتم k-means
۳۱۴	..... تفسیر خوشه‌ها
۳۱۵	..... انتخاب تعداد خوشه‌ها
۳۱۷	..... ارزیابی خوشه‌بندی
۳۱۸	..... الگوریتم Kmeans++
۳۲۰	..... الگوریتم Mini-batch kmeans
۳۲۱	..... خوشه‌بندی DBSCAN
۳۳۰	..... خوشه‌بندی سلسله‌مراتبی
۳۳۱	..... مثال
۳۳۲	..... دندوگرام
۳۳۴	..... الگوریتم تجمعی
۳۳۴	..... معیارهای عدم تشابه
۳۳۶	..... خوشه‌بندی مبتنی بر مدل
۳۳۶	..... توزیع نرمال چندمتغیره
۳۳۸	..... ترکیب نرمال‌ها
۳۴۰	..... انتخاب تعداد خوشه‌ها
۳۴۳	..... مقیاس‌بندی و متغیرهای دسته‌ای
۳۴۳	..... روش‌های مختلف برای مقیاس‌بندی
۳۴۴	..... تأثیر مقیاس‌بندی خوشه‌بندی

۳۴۵	.....	متغیرهای غالب
۳۴۷	.....	داده‌های دسته‌ای و فاصله گاور
۳۵۰	.....	مشکلات خوشه‌بندی داده‌های ترکیبی
۳۵۱	.....	خلاصه
۳۵۲	.....	منابع و مأخذ

خط‌مشی انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب‌هایی با کیفیت عالی است که بتواند  
خواسته‌های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.  
هر کتاب دیباگران تهران، یک فرصت جدید شغلی و علمی

حمد و سپاس ایزد منان را که با الطاف بی‌کران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی و آموزشی گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گسترده‌گی علوم و سرعت توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه، نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش برجسته نموده است.

در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران و محققان در زمینه‌های گوناگون و مورد نیاز جامعه تلاش نموده برای رفع کمبودها و نیازهای موجود، منابعی پُر بار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهد.

کتابی که در دست‌دارید تألیف "سرکارخانم ساره هرمزان و آقایان محمدرضا پور- و سینا توکلی بنیزی" است که با تلاش همکاران ما در نشر دیباگران تهران منتشر گشته و شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

**با نظرات خود مشوق و راهنمای ما باشید**

با ارائه نظرات و پیشنهادات و خواسته‌های خود، به ما کمک کنید تا بهتر و دقیق‌تر در جهت رفع نیازهای علمی و آموزشی کشورمان قدم برداریم. برای رساندن پیام‌هایتان به ما از رسانه‌های دیباگران تهران شامل سایتهای فروشگاهی و صفحه اینستاگرام و شماره‌های تماس که در صفحه شناسنامه کتاب آمده استفاده نمایید.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران  
dibagaran@mftplus.com

## پیش‌گفتاری بر «ملزومات آماری در علم داده‌ها»

روش‌های آماری بخش زیربنایی علوم داده هستند؛ اما با این وجود، تعداد کمی از دانشمندان علوم داده آموزش رسمی در این زمینه دیده‌اند. از طرف دیگر، سرفصل‌های درس‌هایی؛ مانند آمار مهندسی یا آمار و احتمالات و کتاب‌های آماری نیز، به ندرت موضوع‌های آماری را برای تکمیل نیازمندی‌های یک دانشمند علوم داده‌ها و از نگرش داده‌کاوی پوشش می‌دهند. این کتاب، با پر کردن این جای خالی، فاصله ظاهراً قهرآمیز آمار و تحلیل داده‌ها با زبان‌های روز را پُر کرده و خواننده با شروع خواندن اولین فصل از کتاب درمی‌یابد، که تحلیلگران آماری فقط خود را به SPSS محدود نمی‌کنند و داده‌کاوان هم با وجود تسلط بر کدنویسی در زبان‌های تحلیل داده‌ها، از دانش آماری مربوطه بی‌نیاز نیستند.

این پیوند بین آمار و علوم داده‌ها را با مثال‌های جامعی در دو زبان R و Python به کتاب آورده‌ایم تا راهنمای کاربردی برای استفاده از روش‌های آماری برای علوم داده بوده و به شما کمک کند تا از این روش‌ها در جای مناسب استفاده کنید و بالأخره به شما نشان دهد که کدام روش مهم است و کدام یک از این منظر چندان اهمیتی ندارد! بی‌آنکه بخواهیم زحمات نویسندگان دیگر را کم‌ارزش جلوه دهیم؛ اما این مهم را قاطعانه مطرح کرده و در پی رفع آن برآمده‌ایم؛ اینکه بسیاری از منابع علوم داده از روش‌های آماری استفاده می‌کنند؛ اما فاقد چشم‌انداز آماری عمیق و درخور داور علمی هستند. پس اگر شما تا کنون با یکی از زبان‌های R یا پایتون هم آشنایی دارید و از بعضی از آماره‌های آن استفاده کرده‌اید، این راهنمای سریع می‌تواند شکاف موجود را به‌طور خوانا و دسترس‌پذیر پُر نماید. در این کتاب می‌آموزید:

- چرا تحلیل داده‌های توصیفی یک گام کلیدی در علوم داده است.
- چگونه نمونه‌گیری تصادفی می‌تواند بایاس را کاهش داده و به مجموعه داده‌ای باکیفیت‌تر حتی هنگام کار با کلان‌داده‌ها شود.
- چگونه اصول طراحی آزمایشی به جواب‌های قطعی برای سوالات منجر می‌شود.
- چگونه از رگرسیون برای تخمین خروجی‌ها و تشخیص ناهنجاری‌ها استفاده می‌شود.
- روش‌های یادگیری ماشین آماری که از روی داده‌ها «یاد می‌گیرند».
- روش‌های یادگیری نظارت‌نشده برای استخراج مفاهیم از داده‌های بدون برچسب.

این کتاب نه یک کتاب آمار و نه یک راهنمای یادگیری ماشین است، بلکه تلفیقی از این دو است: مفاهیم مهم و ضروری آماری با اصطلاحات داده‌کاوی که در زبان‌های امروزی تحلیل داده‌ها متداول هستند، در این کتاب تلفیق و تبیین می‌شوند، آن‌هم همراه با توضیحات واضح و مثال‌های فراوان. بی‌اغراق بگوییم، متنی آماده شده که برای مبتدیان علوم داده لازم و برای افراد باتجربه هم لذیذ است!

لذت کتاب صرفاً به خاطر مرور دانسته‌های دانشی نیست؛ کتاب مبنای آماری دقیق از کاربردهای مهندسی داده‌ها را شرح می‌دهد که یکی از فواید آن بسنده نکردن به خروجی نرم‌افزارها و ارائه پژوهش‌های قابل پذیرش در داورهای سخت‌گیرانه است!

در واقع، یکی از مشکلات سال‌های اخیر که گریبان‌گیر پژوهشگران شده، کار با نرم‌افزارهای محبوبی است که تحلیلگری داده را انجام داده و افراد بیش از آنکه به زیربنای آماری الگوریتم‌ها توجه داشته باشند، به نتیجه فوری آن‌ها بسنده می‌کنند و نتیجه این فرایند، عدم پذیرش مقالات برخی از فارغ‌التحصیلان عزیز ما که روی پای خود ایستاده‌اند، توسط داوران گرامی و متبخر برخی مجلات معتبر گردیده‌است و این کتاب بخش قابل توجهی از این معضل را با ترویج «علم مفید» جبران می‌نماید و امید است گام کوچکی در جهت تعبیر این دعای پرمعنای پیامبر گرامی<sup>(ص)</sup> باشد که:

«اللهم إني أعوذ بك من علمٍ لا ينفع»

«خدایا به تو پناه می‌برم از علمی که سودی نمی‌بخشد!».

مخاطبان کتاب را می‌توان دانشجویان مقاطع کارشناسی و ارشد رشته‌های علوم کامپیوتر، فناوری اطلاعات، مهندسی نرم‌افزار، مهندسی صنایع، مدیریت فناوری، ریاضی کاربردی، آمار، مدیریت و تمامی علاقمندان به مباحث آماری و تحلیلگری داده‌ها دانست، که ضمن تقدیم مطالب به خوانندگان گرامی، درخواست دارد ما را از نظرات سازنده خویش دریغ نفرمایند.