Especialização em *Data Science* e Estatística Aplicada

Módulo IV – Análise Multivariada

Prof. Dr. Ricardo Limongi

Goiânia, 2025

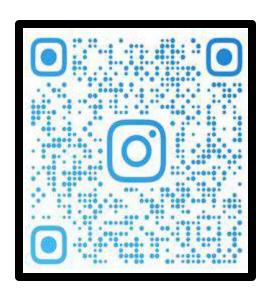








Ricardo Limongi



@limongi

Conteúdo Programático

- Introdução à análise estatística multivariada
- Análise fatorial
- Métodos de agrupamento (análise de clusters)
- Análise discriminante
- Aplicações em dados em linguagem R

l datacamp

Springer Texts in Statistics

Gareth James Daniela Witten Trevor Hastie Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition



Springer Series in Statistic

Trevor Hastie Robert Tibshirani Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

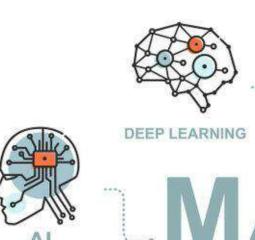
Second Edition

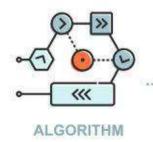


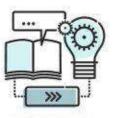
Springer









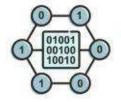


LEARNING

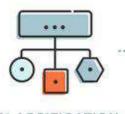
MACHINE. LEARNING



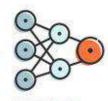




DATA MINING



CLASSIFICATION



NEURAL NETWORKS



AUTONOMUS

Tendências em ML



AUTOMATED MACHINE LEARNING



PRIVACIDADE DOS DADOS E SEGURANÇA



INTERPRETABILIDADE DOS MODELOS



COMPUTER VISION

Técnicas Multivariadas Exploratórias



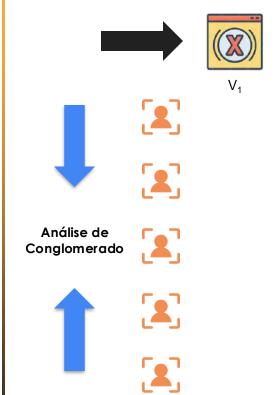
menti.com

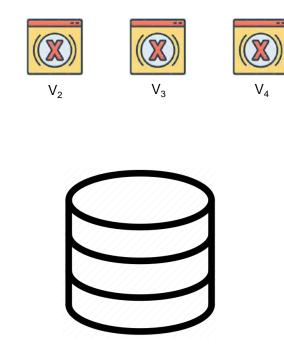
Code - 27398502

"Aprendizagem não supervisionada é, muitas vezes, muito mais desafiadora. O exercício tende a ser **mais subjetivo**, e não há um **objetivo simples** para a análise, como a previsão de uma resposta.

O aprendizado não supervisionado geralmente é realizado como parte de uma **análise exploratória de dados**."

Análise Fatorial





Data

Análise Fatorial

Análise Fatorial (AF)

AGENDA

- Introdução à Análise Fatorial
- Pacotes em R para análise multivariada
- Preparação dos dados
- Verificação de pressupostos
- Análise Fatorial Exploratória (EFA)
- Análise Fatorial Confirmatória (CFA)
- Técnicas avançadas e tratamento de problemas

Introdução

O que é Análise Fatorial?

- Técnica estatística multivariada para identificar padrões de correlação
- Reduz variáveis a um número menor de dimensões ou fatores latentes
- Descobre constructos não observáveis diretamente

Principais aplicações:

- Validação de escalas e questionários
- Identificação de dimensões subjacentes em dados complexos
- Redução de dimensionalidade para análises subsequentes

Análise Fatorial (AF) x Análise de Componentes Principais (ACP)

AF e ACP

AF:

- Identificar fatores latentes (não observáveis) que explicam correlações entre variáveis
- Assumes que as variáveis observadas são manifestações de constructos latentes (não observáveis)
- Foca na variância comum (comunalidade) entre variáveis

PCA:

- Encontrar componentes que maximizam a variância total nos dados
- PCA: Simples transformação matemática para criar novas variáveis não correlacionadas
- Considera toda a variância (comum e específica)

Tipos de AF

Análise Fatorial Exploratória (EFA)

- Descoberta de estruturas latentes sem hipóteses prévias
- Identifica agrupamentos naturais de variáveis

Análise Fatorial Confirmatória (CFA)

- Testa hipóteses sobre estruturas fatoriais predefinidas
- Verifica o ajuste de modelos teóricos
- Valida estruturas fatoriais propostas

Pacotes

```
"psych", # Análise fatorial exploratória
"GPArotation", # Métodos de rotação
"tidyverse",
               # Manipulação de dados
"corrplot",
               # Visualização de correlações
               # Análise fatorial confirmatória
"lavaan".
"semPlot",
               # Visualização de modelos
"car".
               # Transformações de dados
"MASS".
               # Métodos robustos
"mice"
               # Imputação de dados
```

Pressupostos

Teste de Bartlett:

- H0: matriz de correlação é uma matriz identidade
- Interpretação: p < 0.05 indica adequação para análise fatorial
- Sintaxe: cortest.bartlett(dados) do pacote psych

Índice KMO:

- Medida de adequação amostral (MSA)
- Valores entre 0 e 1 (>0.6 aceitável, >0.8 excelente)
- Sintaxe: KMO(dados) do pacote psych

Visualização de correlações:

- corrplot() oferece várias opções de visualização
- Parâmetros importantes: method ("color", "circle", "number")
- Alternativas: pairs.panels() do psych para visualização mais completa

Pressupostos

```
# Teste de Esfericidade de Bartlett
bartlett_teste <- cortest.bartlett(dados_saude)

# Índice Kaiser-Meyer-Olkin (KMO)
kmo_teste <- KMO(dados_saude)
```

EFA

Carga Fatorial: Representa a correlação entre a variável original e o fator extraído. Indica quanto cada variável está associada a um determinado fator latente.

Interpretação:

Valores entre -1 e 1

Quanto maior o valor absoluto, mais forte a associação

Critérios comuns de significância:

0.30: Mínimo aceitável

0.40: Importante

0.50: Significativo

0.70: Bem definido

Comunalidade

Proporção da variância de uma variável que é explicada pelos fatores comuns extraídos no modelo.

Interpretação:

- Varia de 0 a 1
- Quanto mais próximo de 1, melhor a variável é explicada pelos fatores
- Comunalidade baixa (< 0.5) sugere que a variável não está bem representada na solução fatorial
- Comunalidade alta (> 0.7) indica boa representação da variável pelos fatores comuns

Variância Explicada

Indica o quanto da variabilidade total dos dados originais é capturada por cada fator e pelo modelo

Interpretação:

- Expressa geralmente em porcentagem
- Quanto maior, melhor o modelo representa os dados originais

Regras práticas:

- Ciências sociais: considera-se aceitável 60% da variância total explicada
- Ciências naturais: geralmente exige-se 80% ou mais

A distribuição da variância entre os fatores indica a importância relativa de cada constructo latente

Interpretação da Análise Fatorial Exploratória

Cargas > 0,4 são consideradas significativas; Comunalidades > 0,5 são desejáveis

```
# Cargas fatoriais
modelo_efa$loadings

# Comunalidades
modelo_efa$communality

# Variância explicada
modelo_efa$Vaccounted

# Correlações entre fatores (para rotações oblíquas)
modelo_efa$Phi
```

Rotações

Critérios de Escolha

- Correlação entre fatores
- Base teórica
- Objetivo da análise
- Complexidade da estrutura
- Abordagem prática

Correlação entre fatores

Rotações ortogonais (quando os fatores não são correlacionados):

- Varimax: Maximiza a variância das cargas dentro dos fatores; cada variável tende a ter carga alta em um fator e baixas nos demais
- Quartimax: Simplifica as linhas da matriz de cargas; cada variável tende a carregar principalmente em um fator
- Equamax: Combinação de Varimax e Quartimax; equilibra a simplificação de variáveis e fatores

Rotações oblíquas (quando os fatores podem ser correlacionados):

- Promax: Rotação oblíqua eficiente computacionalmente
- Oblimin: Permite controle sobre o grau de correlação entre fatores
- Geomin: Útil para estruturas complexas ou quando há muitos fatores

Base Teórica

Conhecimento prévio do domínio:

- Se a teoria sugere que os constructos são independentes, use rotações ortogonais; se há razão para acreditar que os constructos são relacionados, use rotações oblíquas
- Em ciências sociais e saúde, fatores raramente são completamente independentes, tornando rotações oblíquas mais realistas

Objetivo da Análise

- Simplificação máxima: Varimax (ortogonal) ou Promax (oblíquo)
- Identificação de fator geral + específicos: Quartimax
- Equilíbrio entre estrutura simples e interpretabilidade: Equamax ou Geomin

Abordagem Prática

- Começar com uma rotação oblíqua (Promax ou Oblimin) para verificar a correlação entre fatores
- Se as correlações entre fatores forem baixas (< 0.32), considerar trocar para rotação ortogonal
- Comparar soluções usando diferentes métodos de rotação e avaliar qual oferece a estrutura mais interpretável
- Considerar o princípio da parcimônia a solução mais simples que explica adequadamente os dados

Extração e Rotação de Fatores

- Varimax: Maximiza variância das cargas (ortogonal)
- Promax: Permite correlação entre fatores (oblíqua)
- Quartimax, Oblimin: Alternativas para casos específicos

EFA

Definição do número de fatores

AFC

É preciso especificar:

- Quais fatores existem no modelo
- Quais variáveis são indicadores de cada fator
- Se os fatores são correlacionados entre si
- Se existem correlações entre erros de medida

AFC

Processo de Implementação

- Especificação do modelo: Definir exatamente quais variáveis pertencem a quais fatores
- Identificação: Garantir que o modelo tenha graus de liberdade suficientes
- Estimação: Geralmente usa Maximum Likelihood ou métodos robustos
- Avaliação do ajuste: Análise de índices de ajuste para verificar adequação do modelo
- Modificação (se necessário): Ajustes baseados em índices de modificação, com justificativa teórica

AFC

Índices de Ajuste

- Chi-quadrado (χ²): Idealmente não significativo (p > 0.05), mas sensível ao tamanho da amostra
- RMSEA (Root Mean Square Error of Approximation): < 0.06 excelente, < 0.08 aceitável
- CFI (Comparative Fit Index): > 0.95 excelente, > 0.90 aceitável
- TLI (Tucker-Lewis Index): Similar ao CFI, penaliza modelos complexos
- SRMR (Standardized Root Mean Square Residual): < 0.08 aceitável

Índices de Ajuste

Chi-quadrado (χ^2):

- Teste de ajuste absoluto (p > 0.05 desejável)
- Sensível ao tamanho da amostra
- Sintaxe para extração: fitMeasures(ajuste_cfa, "chisq")

Índices comparativos:

- CFI (Comparative Fit Index): >0.95 excelente, >0.90 aceitável
- TLI (Tucker-Lewis Index): similar ao CFI, penaliza modelos complexos
- Sintaxe: fitMeasures(ajuste_cfa, c("cfi", "tli"))

AFC

Índices baseados em resíduos:

- RMSEA (Root Mean Square Error of Approximation): <0.06 excelente,
 <0.08 aceitável
- SRMR (Standardized Root Mean Square Residual): <0.08 aceitável
- Sintaxe: fitMeasures(ajuste_cfa, c("rmsea", "srmr"))

Índices de informação:

- AIC, BIC: úteis para comparar modelos alternativos (menor = melhor)
- Sintaxe: fitMeasures(ajuste_cfa, c("aic", "bic"))

Boas Práticas e Recomendações

Validação cruzada

Dividir amostra: EFA em uma parte, CFA em outra

Interpretabilidade sobre ajuste

Fatores devem fazer sentido teórico/prático

Tamanho amostral adequado

5-10 observações por variável (mínimo)

Documentação completa

Decisões metodológicas devem ser justificadas

Validação convergente e discriminante

Correlação com construtos relacionados

Análise de Conglomerado

Afinal, o que é Análise de Cluster?

- Técnica de aprendizado não supervisionado
- Agrupa objetos similares em categorias (clusters)
- Objetivo: maximizar homogeneidade dentro dos grupos e heterogeneidade entre grupos
- Sem conhecimento prévio de classes/categorias

Aplicações

- Marketing: Segmentação de clientes
- Biologia: Agrupamento de genes similares
- Medicina: Identificação de subgrupos de pacientes
- Finanças: Detecção de anomalias
- Análise de texto: Agrupamento de documentos

Pacotes

```
tidyverse # Manipulação e visualização de dados
```

cluster # Algoritmos de clustering factoextra # Visualização de clusters

NbClust # Determinar número ótimo de clusters

fpc # Estatísticas de validação

mclust # Clustering baseado em modelos

dendextend # Comparação de dendrogramas

dbscan # Clustering baseado em densidade

Processo de Análise

- 1. Preparação dos dados
- 2. Determinação do número ótimo de clusters
- 3. Aplicação de algoritmos de clustering
- 4. Validação e avaliação
- 5. Interpretação dos resultados

Preparação dos dados

```
# Z-score (média = 0, desvio padrão = 1)
dados_padronizados <- scale(dados_cluster)
# Min-Max (0 a 1)
normalizar_minmax <- function(x) {</pre>
 return((x - min(x)) / (max(x) - min(x)))
dados_minmax <- as.data.frame(lapply(dados_cluster, normalizar_minmax))
# Robusta (baseada em mediana e MAD)
normalizar_robusto <- function(x) {
 return((x - median(x)) / mad(x))
dados_robustos <- as.data.frame(lapply(dados_cluster, normalizar_robusto))
```

Avaliação da Tendência

```
# Estatística de Hopkins
# > 0.5 indica tendência a clustering
tendencia <- hopkins(dados_padronizados, n = nrow(dados_padronizados) * 0.1)
tendencia_valor <- as.numeric(tendencia)

# Visualização com PCA
pca_resultado <- prcomp(dados_padronizados)
dados_pca <- as.data.frame(pca_resultado$x[, 1:2])
```

Cotovelo

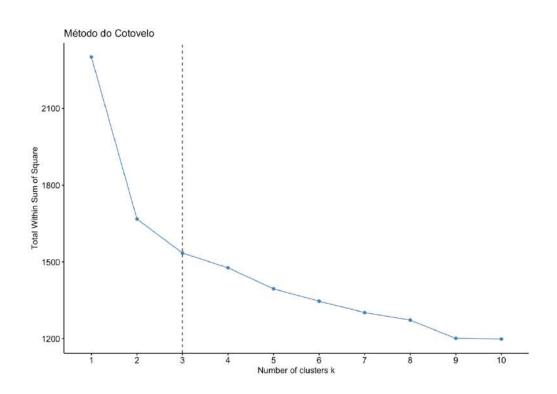
Avalia a variação dos dados conforme o número de agrupamentos

Silhueta

Interpretação e avaliação da consistência dentro do agrupamento

Método do Cotovelo

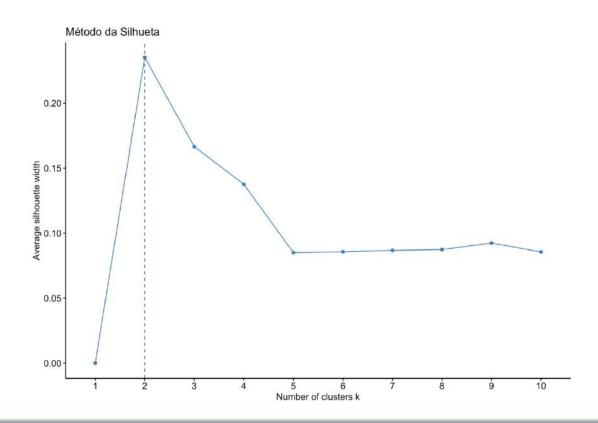
```
wss <- sapply(1:10, function(k) {
   kmeans(dados_padronizados, centers = k, nstart = 25)$tot.withinss
})
# Visualização com factoextra
fviz_nbclust(dados_padronizados, kmeans, method = "wss", k.max = 10)</pre>
```



Método da Silhueta

```
silhouette_scores <- sapply(2:10, function(k) {
   km <- kmeans(dados_padronizados, centers = k, nstart = 25)
   ss <- silhouette(km$cluster, dist(dados_padronizados))
   mean(ss[, 3])
})

# Visualização com factoextra
fviz_nbclust(dados_padronizados, kmeans, method = "silhouette", k.max = 10)
```



Principais Algoritmos

K-means

- Conceito: Minimiza a soma dos quadrados das distâncias aos centroides
- Vantagem: Simples, eficiente, escalável
- Desvantagem: Sensível a outliers, requer número predefinido de clusters, clusters esféricos

Principais Algoritmos

Clustering Hierárquico

- Conceito: Constrói hierarquia de clusters (aglomerativo ou divisivo)
- Vantagem: Visualização hierárquica, não requer k predefinido
- Desvantagem: Computacionalmente intensivo para grandes conjuntos

Método de Ligação na Abordagem Hierárquica

Complete

Distância entre elementos mais distantes

Single

Distância entre elementos mais próximos

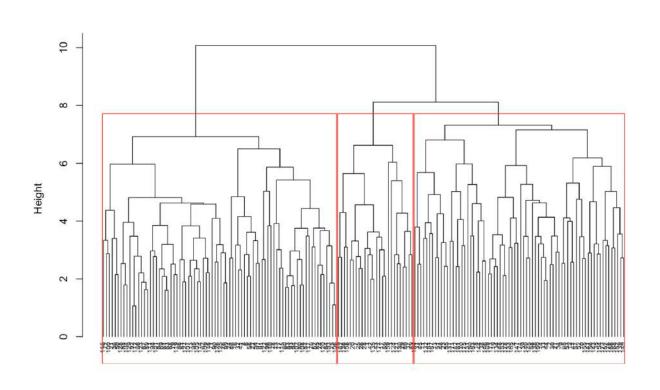
Average

Média das distâncias

Ward

Minimiza variância dentro dos clusters

Dendograma



Quadro Comparativo dos Algoritmos

Método	Vantagens	Desvantagens	Quando Usar
K-means	Simples, eficiente, escalável	Sensível a outliers, requer k, clusters esféricos	Dados grandes, clusters bem separados
Hierárquico	Visualização de relacionamentos, não requer k	Computacionalment e intensivo	Análise exploratória, clusters aninhados
DBSCAN	Detecta formas arbitrárias, identifica ruído	Sensível a parâmetros	Dados com ruído, clusters de formas irregulares
GMM	Flexível para diferentes formas, probabilidades	Supõe normalidade, computacionalment e intensivo	Clusters sobrepostos, incerteza relevante

Análise Discriminante

Introdução

- Técnica estatística desenvolvida por Ronald Fisher (1936)
- Busca encontrar combinações lineares de variáveis que melhor separam classes
- Técnica supervisionada (requer dados rotulados)
- Maximiza separação entre grupos e minimiza variação dentro dos grupos

Pressupostos

- Normalidade multivariada das variáveis independentes
- Homogeneidade de variâncias-covariâncias entre grupos
- Fronteiras de decisão lineares
- Amostras representativas para cada grupo

Bibliotecas

```
library(MASS)  # Para função Ida()
library(dplyr)  # Para manipulação de dados
library(ggplot2)  # Para visualização
library(caret)  # Para validação e métricas
library(klaR)  # Para visualização da LDA
```

Etapas

- Preparação dos dados
- Análise Exploratória
- Divisão Treino/Teste

Modelo LDA

```
# Sintaxe básica da LDA
modelo_lda <- Ida(grupo ~ idade + pressao_sistolica + colesterol_total +
hdl + triglicerideos + glicemia + imc,
data = dados_treino)
```

Avaliação do Modelo

```
# Predições no conjunto de teste
predicoes <- predict(modelo_lda, dados_teste)</pre>
confusao <- table(Predito = predicoes$class, Real = dados teste$grupo)
# Acurácia
acuracia <- sum(diag(confusao)) / sum(confusao)</pre>
# Métricas detalhadas
caret::confusionMatrix(predicoes$class, dados_teste$grupo)
# Probabilidades posteriores
head(predicoes$posterior)
```

Interpretação dos Resultados

- Análise dos coeficientes padronizados
- Identificação das variáveis importantes
- Entendimento das funções discriminantes

Limitações

- Pressupõe normalidade multivariada
- Assume homogeneidade das variâncias-covariâncias
- Sensível a outliers
- Fronteiras de decisão lineares
- Probabilidades posteriores dependem das proporções dos grupos de treino

Possíveis Extensões

- Análise Discriminante Quadrática (QDA)
- Flexível na suposição de homogeneidade de variâncias-covariâncias
- Permite fronteiras de decisão não-lineares
- Regularização para lidar com multicolinearidade
- Validação cruzada para estimar erro de generalização
- Combinação com outras técnicas de classificação

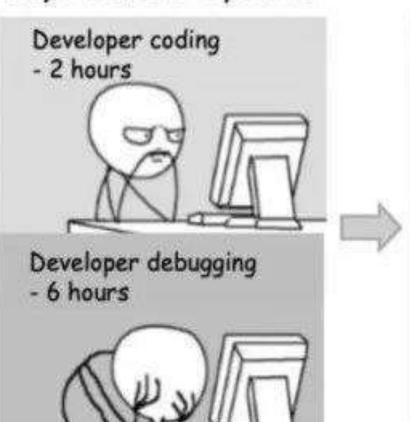
Inteligência Artificial



Gemini



Days before OpenAI



Days after OpenAI









Chat



LLM Chat

Interact with various LLMs (GPT, Claude, Grok,



Agent



Develop LLM agents, chatbots, and workflows



Autogen Studio

Create autonomous agent group chats to solve problems

Image



Flux (Schnell, Dev, Pro)



Generate images from texts from various Flux



Create visuals with hidden texts for various uses

SD with ControlNets ControlNets add precise control options for enhanced image generation



Stable Diffusion (SD, SDXL, SD3)

Generate images from texts using various Stable Diffusion models



SD with LoRA

LoRA models adjust image styles to enable unique artistic creations



DALL-E 3

Generate images from text descriptions using OpenAl DALL E 3



JULIUS.AI

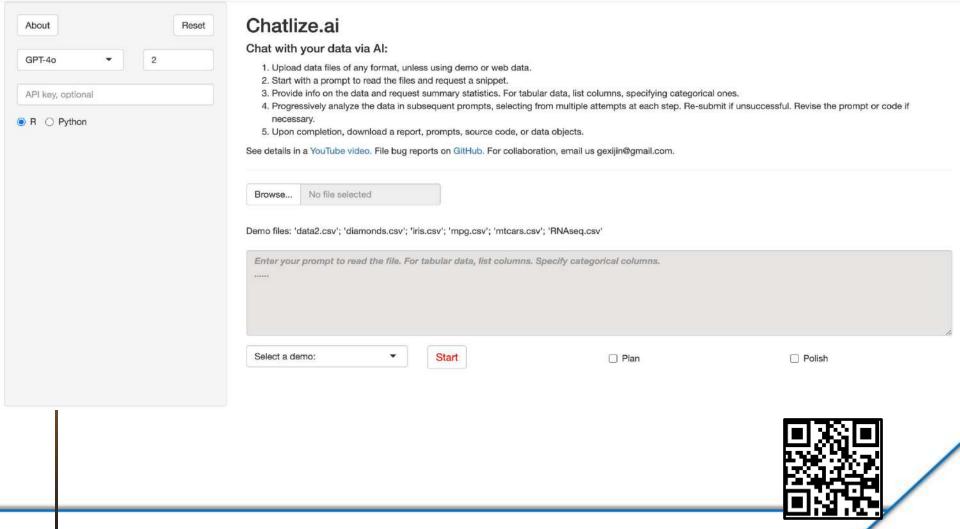
Julius.ai

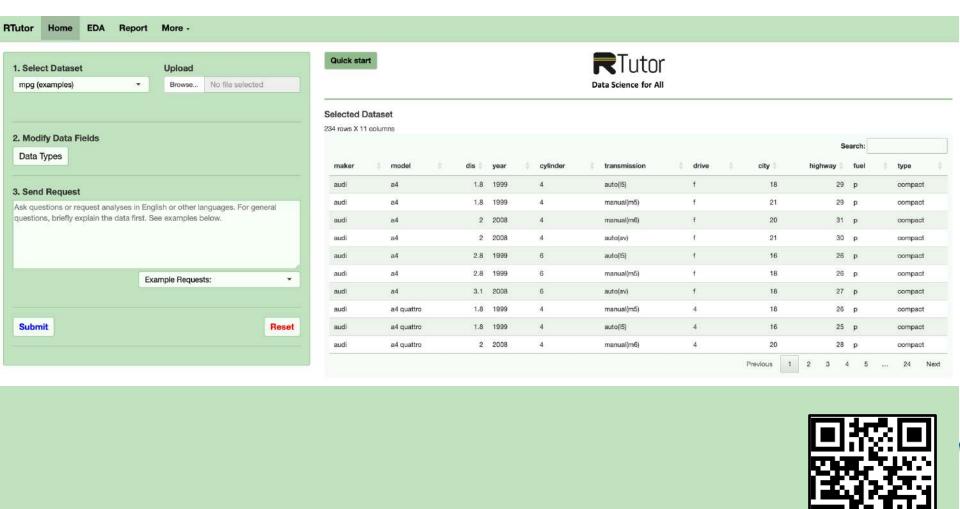
AIDATA ANALYST















Code only Concise

Verbose



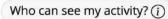


Hello, I'm Shiny Assistant! I'm here to help you with <u>Shiny</u>, a web framework for data driven apps. You can ask me questions about how to use Shiny, to explain how certain things work in Shiny, or even ask me to build a Shiny app for you.

Here are some examples:

- "How do I add a plot to an application?"
- "Create an app that shows a normal distribution."
- "Show me how make it so a table will update only after a button is clicked."
- Ask me, "Open the editor", then copy and paste your existing Shiny code into the editor, and then ask me to make changes to it.

Let's get started! 🚀





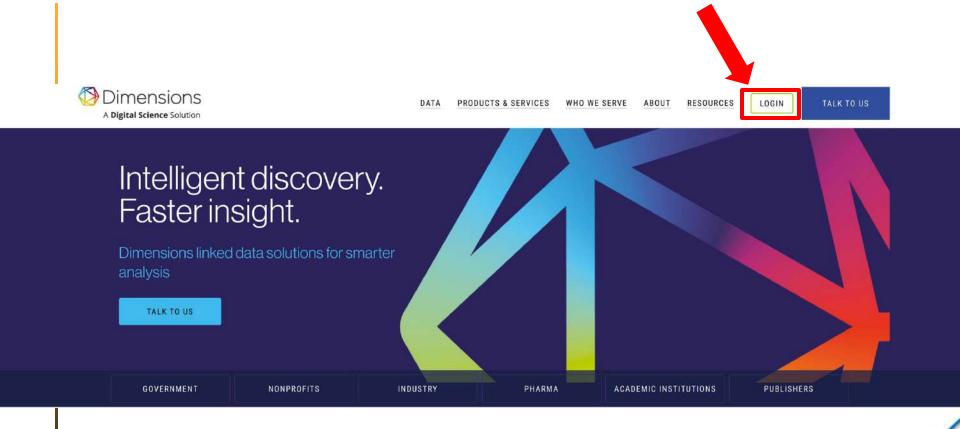
Enter a message...





Dimensions

A **Digital Science** Solution



https://www.dimensions.ai/

Sort by: Publication date ~

33 274 161

19,576,610

14:238.740

11.459.047

9.531,465

Dimensions

FAVORITES

> PUBLICATION YEAR

-) RESEARCHER

FILTERS

- > RESEARCH CATEGORIES
- > PUBLICATION TYPE
-) SOURCE TITLE
-) JOURNAL LIST
- OPEN ACCESS

Free Account

For personal, non-commercial use only.



About Dimensions - LinkedIn - X

Privacy policy · Cookie settings - Legal terms

@ 2024 Digital Science & Research Solutions, Inc.



Show abstract Title, Author(s), Bibliographic reference - About the metrics

Comparison of patient outcomes of anterior and posterior lumbar interbody fusions: A retrospective national database analysis

Cole Veliky, Paul Michael Alvarez, Hania Shahzad, Diego Martinez, Elizabeth Yu, Varun K. Singh 2025, Journal of Orthopaedics - Article

Background Lumbar interbody fusions are used to treat degenerative lumbar disease unresponsive to conservative treatment. This procedure may be divided into anterior lumbar interbody fusion (ALIF), an... more

View PDF ≡ Add to Library

Q e.g. plastic AND instrument

(+) Chat with PDF + Summarize

Early Release - Equine Encephalomyelitis Outbreak, Uruquay, 2023-2024 - Volume 31, Number 1-January 2025 - Emerging Infectious Diseases journal - CDC

Sandra Frabasile, Noelia Morel, Ramiro Pérez, Lucía Moreira Marrero, Analia Burgueño, María Noel Cortinas, Lucía ... 2025, Emerging Infectious Diseases - Article

We report the genomic analysis from early equine cases of the Western equine encephalitis virus outbreak during 2023-2024 in Uruguay. Sequences are related to a viral isolate from an outbreak in 1958 ... more

♠ Chat with PDF ♦ Summarize Altmetric 3 View PDF ≡ Add to Library

Turner syndrome and neuropsychological abnormalities; a review and case series

Bruna Baierle Guaraná, Marcela Rodrigues Nunes, Victória Feitosa Muniz, Bruna Lixinski Diniz, Maurício Rouvel Nu... 2025. Revista Paulista de Pediatria - Article

OBJECTIVE: The objective of this study was to establish the genotype-phenotype correlation between karyotype results and the neurological and psychiatric alterations presented in patients with Turner ... more

Altmetric 1 View PDF = Add to Library

(+) Chat with PDF + Summarize

ANALYTICAL VIEWS RESEARCH CATEGORIES 32 Biomedical and Clinical Sciences 40 Engineering 3202 Clinical Sciences 31 Biological Sciences 34 Chemical Sciences OVERVIEW Citations Citations (Mean) 2.03 B 13.53 10.000.000 7,500,000 5 000 000 2,500,000



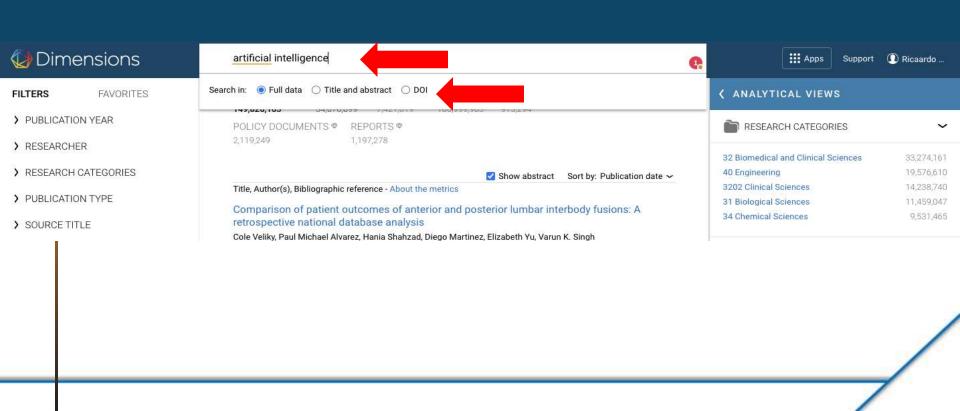
RESEARCHERS

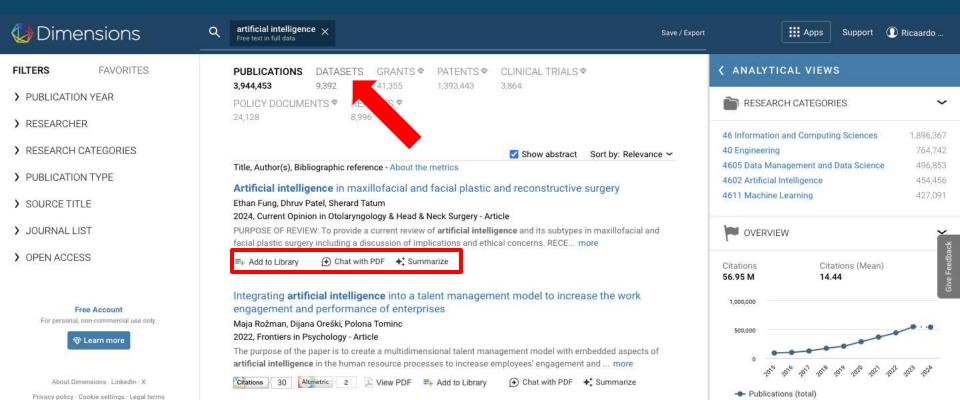
Publications (total)

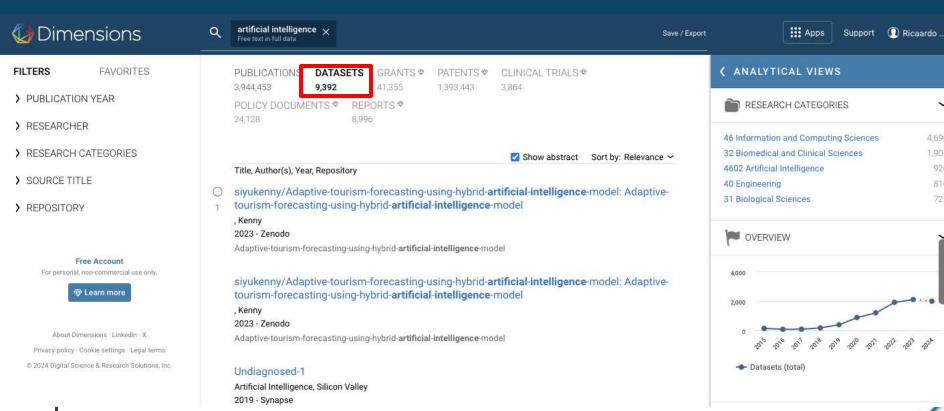
Ramchand T Pardasani Central University of Rajasthan, India

Pushpa Pardasani University of Rajasthan, India 17.574

17.540







4.690

1.901

926

810

725





Search

Current

Archives Call for Papers Announcements

Indexing and abstracting

Guide to authors

Home / Archives / Vol. 21 No. 3 (2024): Jul/Sep - 2024 / Research Articles

An Empirical Study of Digital Lending in India and the Variables Associated with its Adoption



Akhileshwari Asamani

Koneru Lakshmaiah Education Foundation, Department of Management Hyderabad, Telangana, India

https://orcid.org/0000-0002-7639-2376

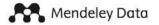
Javavani Maiumdar

Koneru Lakshmaiah Education Foundation, Department of Management Hyderabad, Telangana, India

https://orcid.org/0009-0000-0741-6672

Abstract

Objective: the digital lending platform is a significant and innovative business concept in the field of financial technology. It creates a direct connection between lenders and borrowers. Scientists studied the technique by which digital lending companies use leverage. This study proposes an updated technology acceptance model (TAM) to investigate the factors influencing consumer adoption of digital lending platforms. More specifically, it examines how service quality and perceived threat influence trust development. Methods: this study employs structural equation modeling (SEM) to investigate the potential links between the underlying variables. Results: the study found that customers' attitudes toward adoption of digital lending platforms are highly influenced by trust, perceived value, and perceived ease of use. The quality of service has a considerable impact on consumers' perceptions of risk while using digital lending services. The generated model corresponds to the findings of previous studies. Conclusions: the findings of the continuing research are important for optimizing platform marketing strategies and



You are viewing a previous version of this dataset

AN EMPIRICAL STUDY OF DIGITAL LENDING IN INDIA AND THE VARIABLES ASSOCIATED WITH ITS ADOPTION

Published: 27 August 2024 | Version 1 | DOI: 10.17632/syb6twd96n.1

Contributor: Akhileshwari Asamani

Description

As a fresh and trending occurrence in today's world, digital lending in India has piqued my curiosity for research. FinTech use has increased significantly in recent years, particularly during the pandemic. From the days when pawnshops would lend money in exchange for collateral to the present day, when sophisticated technology is utilised to carry out credit underwriting, the backdrop of lending organisations has changed dramatically over the last century. This research helps to get to know about numerous digital lending platforms that are popular among users, as well as the factors that influence digital lending platforms.

Download All 5.39 KB



Files



46 7 KB

46.7 KB

Categories

Technology Adoption

Licence

CC BY 4.0

Learn more

Google CO CO C



Especialização em *Data Science* e Estatística Aplicada

Módulo IV – Análise Multivariada

Prof. Dr. Ricardo Limongi

Goiânia, 2024





