DATA VR

PRESENTADO POR: MIGUEL ÁNGEL HERNÁNDEZ VARGAS Y DIEGO ANDRES AVELLA ACEVEDO DE LA FACULTAD DE INGENIERÍA ESTADÍSTICA

PRESENTADA A: ANA MARIA GOMEZ LAMUS

UNIVERSIDAD ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

BOGOTÁ D, C

2024

**RESUMEN**

En el análisis multivariado, el Análisis de Componentes Principales (PCA, por sus siglas en inglés) es una técnica ampliamente utilizada para reducir la cantidad de variables a estudiar, facilitando su representación gráfica. Sin embargo, la simplificación en dos dimensiones puede ocasionar solapamiento de puntos, pérdida de información o dificultades en la clasificación clara de los datos.

Este proyecto presenta el desarrollo de una herramienta interactiva basada en realidad virtual (VR), diseñada para superar estas limitaciones mediante una visualización tridimensional intuitiva e inmersiva. La plataforma integra técnicas estadísticas avanzadas, incluyendo análisis exploratorio (EDA), reducción de dimensionalidad mediante PCA y modelado predictivo con regresión logística. Utilizando el framework A-Frame y bibliotecas especializadas como PapaParse, ml-pca y ml-logistic-regression, la herramienta genera gráficos tridimensionales interactivos, conservando las categorías originales definidas en los conjuntos de datos y mostrando contribuciones individuales y de variables a las componentes principales.

La herramienta permite la identificación automática de posibles errores de etiquetado mediante la comparación visual de clasificaciones reales y predicciones generadas por el modelo. Además, ofrece una navegación fluida tanto desde navegadores web como desde gafas de realidad virtual, proporcionando una experiencia inmersiva que mejora significativamente la comprensión y precisión del análisis visual de resultados complejos. Está dirigida principalmente a analistas e investigadores que buscan optimizar la interpretación y la toma de decisiones mediante la exploración visual interactiva y colaborativa de datos multivariados.

**INTRODUCCIÓN**

Este proyecto implementa una herramienta interactiva de visualización en realidad virtual, centrada en el análisis de datos descriptivo y predictivo. Se enfoca en el análisis de componentes principales (PCA) y en la evaluación y entrenamiento de un modelo de regresión logística, mostrando tanto información de los datos originales como los resultados del entrenamiento del modelo para la clasificación de un conjunto de datos en función de una variable de respuesta. La solución facilita la detección de posibles errores en la clasificación de los registros, ayudando a identificar puntos que podrían estar mal catalogados al observar su disparidad con la clasificación predicha por el modelo en comparación con su categoría original, así como el comportamiento general de los datos en su conjunto.

**OBJETIVOS**

**OBJETIVO GENERAL:**  
Desarrollar una herramienta interactiva en realidad virtual que facilite el análisis descriptivo y predictivo de datos, integrando técnicas como el análisis de componentes principales (PCA) y la regresión logística, para mejorar la clasificación de datos y detectar posibles errores de etiquetado en un conjunto de datos.

**OBJETIVOS ESPECIFICOS:**

* Implementar un modelo de regresión logística que clasifique un conjunto de datos en función de una variable de respuesta, evaluando su rendimiento mediante métricas estándar como la precisión y la matriz de confusión.
* Mostrar los resultados del entrenamiento del modelo, incluyendo las predicciones y las clasificaciones reales, para permitir la comparación visual y la identificación de posibles errores de clasificación.
* Desarrollar una funcionalidad para identificar y resaltar puntos en los datos que podrían estar mal etiquetados, comparando la clasificación predicha por el modelo con la categoría original de los datos.

**MATERIAL UTILIZADO**

1. **Introducción**
   1. La herramienta desarrollada permite realizar análisis multivariados avanzados, incluyendo reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) y clasificación con regresión logística, entre otros modelos estadísticos. Los resultados de estos análisis se visualizan en un entorno de realidad virtual (VR) inmersivo, que incluye una sala virtual donde el usuario puede interactuar con representaciones gráficas tridimensionales, facilitando la exploración y comprensión de los datos complejos.
2. **Requisitos del sistema**
   1. **Hardware:** Para el desarrollo, pruebas y uso final, se emplearon computadores con conexión a internet para acceder al repositorio y hosting del proyecto, junto con gafas de realidad virtual Meta Quest con soporte actualizado para el framework VR utilizado. Para pruebas sin gafas se utilizó también el emulador "Immersive Web Emulator for Meta Quest devices".
   2. **Software:** El backend y la lógica de análisis están implementados en JavaScript, mientras que el frontend se desarrolla en HTML y A-Frame para la generación del entorno VR. La gestión de datos y la integración con modelos estadísticos se realizan mediante bibliotecas JavaScript especializadas.
3. **Librerías y frameworks** 
   1. Framework principal: Para realizar el entorno virtual y el gráfico de dispersión se utilizó A-frame en su versión 1.2.0.
   2. Librerías:
      1. *PapaParse* v5.3.0 para la lectura y manejo eficiente de archivos CSV.
      2. *csv-parser* para la conversión entre formatos CSV y JSON, facilitando la interoperabilidad entre los datos y el entorno VR.
      3. *ml-pca* v2.0.0 para la ejecución del Análisis de Componentes Principales, garantizando compatibilidad con el resto del ecosistema de librerías.
      4. *ml-logistic-regression* y otras librerías de *machine learning* para la implementación de modelos estadísticos como regresión logística, los cuales se integran en el flujo de análisis para complementar la reducción de dimensionalidad con capacidades predictivas.
4. **METODOLOGIA**

El desarrollo del proyecto se fundamentó en la integración de visualizaciones en realidad virtual (VR) con modelos estadísticos de análisis multivariado, complementando inicialmente un entorno para Análisis de Componentes Principales (PCA) con la incorporación posterior de modelos de clasificación.

En lugar de utilizar plataformas online como Glitch, el proyecto fue implementado y ejecutado localmente para obtener mayor control y flexibilidad. Para poder visualizar el entorno VR generado con A-Frame, fue necesario establecer una conexión local mediante un servidor web que permitiera abrir la aplicación desde el navegador, ya que simplemente abrir los archivos HTML directamente no soporta adecuadamente las características WebXR ni el manejo de dependencias necesarias.

El entorno local facilitó la integración de dependencias y librerías para la manipulación avanzada de datos, como PapaParse para lectura de archivos CSV, ml-pca para la reducción de dimensionalidad y librerías para modelos estadísticos como regresión logística. Esto permitió que el procesamiento y análisis se hicieran de forma eficiente y segura dentro del flujo del proyecto.

Durante el desarrollo se diseñó un espacio virtual inmersivo (la "sala VR") donde el usuario puede interactuar con gráficos tridimensionales que reflejan los resultados de los análisis, facilitando la exploración intuitiva de conjuntos de datos complejos.

Entre los desafíos técnicos superados destacan:

* Configurar el servidor local para servir los archivos estáticos con soporte para WebXR y permitir la conexión con las gafas de realidad virtual Meta Quest.
* Manejar el flujo de datos desde archivos CSV hasta las visualizaciones en VR, garantizando la correcta conversión y procesamiento de los datos.
* Implementar y probar modelos estadísticos dentro del entorno, asegurando su correcta integración con la visualización.

Esta metodología garantizó un desarrollo flexible, adaptable y escalable, permitiendo ampliar el proyecto con nuevos modelos y funcionalidades futuras.

**Procesamiento y modelado estadístico**

Aunque inicialmente el núcleo de la aplicación se centraba en la reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA), el sistema fue evolucionando hacia una herramienta más completa de análisis multivariado. Esto incluyó la incorporación de modelos predictivos como la regresión logística, que aportan una dimensión práctica al análisis exploratorio, permitiendo no solo visualizar los datos, sino también tomar decisiones basadas en ellos.

**Aplicación del PCA**  
Una vez estructurado el entorno y automatizado el procesamiento de datos, se integró el paquete ml-pca para aplicar la reducción de dimensionalidad. Uno de los primeros desafíos fue garantizar que los datos estuvieran correctamente estructurados. Para ello, se implementó un módulo de lectura y limpieza de archivos CSV, que identifica automáticamente las variables numéricas requeridas por el PCA, y permite al usuario especificar las variables categóricas.

Este proceso de estandarización facilitó la aplicación sistemática del modelo y la extracción de resultados relevantes para la visualización tridimensional. Además del cálculo de las nuevas coordenadas proyectadas, se implementaron funciones personalizadas para profundizar en la interpretación del modelo:

1. **Contribución de los individuos a los componentes principales:** calculada con base en los valores proyectados, permitiendo identificar observaciones influyentes.
2. **Contribución de las variables a los componentes:** basada en las cargas obtenidas mediante getLoadings(), procesadas para obtener los porcentajes de influencia por variable en cada componente.

Estas funciones fueron desarrolladas desde cero, dado que el paquete no ofrece mecanismos directos como otros entornos (por ejemplo, R con FactoMineR), lo que requirió una comprensión teórica sólida del método.

**Implementación de la regresión logística**  
En la fase posterior del proyecto se integró un modelo de regresión logística utilizando librerías de machine learning en JavaScript. Este modelo se aplica a conjuntos de datos con una variable categórica binaria como objetivo.

El flujo para este análisis comienza con la selección de las variables predictoras y la variable de salida. El sistema permite al usuario ajustar parámetros como la proporción de entrenamiento y prueba, y posteriormente visualiza los resultados del modelo. Se muestran métricas clave como la **exactitud**, **precisión**, **recall**, y **F1-score**.

**Configuración del servidor**  
Para ejecutar la aplicación localmente con soporte completo para visualización en realidad virtual, fue necesario configurar un servidor local que permitiera servir los archivos de forma dinámica y habilitara el uso adecuado de dependencias y recursos WebXR.

Utilizamos un servidor basado en Node.js con Express para exponer tanto los archivos estáticos (HTML, scripts y entorno A-Frame) como los archivos de datos (CSV) desde una carpeta pública. Esta estructura nos permitió acceder a los datos a través de rutas locales bien definidas y mantener un flujo coherente entre la carga de datos, el procesamiento estadístico y la visualización en VR.

Adicionalmente, para que las gafas Meta Quest pudieran acceder correctamente al entorno VR, fue necesario abrir el servidor en una red local (LAN) y acceder mediante la dirección IP del equipo anfitrión. Esto se convirtió en un paso fundamental para permitir una visualización inmersiva directa desde el visor, ya que no era posible simplemente abrir los archivos desde el explorador local debido a restricciones del navegador con WebXR.

Esta configuración garantizó un entorno controlado, reproducible y escalable, ideal para pruebas con modelos estadísticos interactivos en entornos inmersivos.

**Creación del entorno virtual e integración de visualizaciones estadísticas**

La construcción del entorno virtual se realizó utilizando el framework **A-Frame**, que permite generar escenarios en WebXR directamente desde el navegador. A partir de esta base, se diseñó una **sala inmersiva interactiva**, donde el usuario puede explorar múltiples representaciones estadísticas en 3D, moverse libremente por el entorno y seleccionar elementos para obtener información contextual.

**Sala Virtual**

El espacio está construido como una habitación cerrada con paredes texturizadas, piso y techo, en la que se distribuyen distintos paneles temáticos. Estos paneles representan diferentes aspectos del análisis estadístico aplicado, incluyendo:

* **Análisis de Componentes Principales (PCA)**
* **Regresión Logística**
* **Resumen de Datos (EDA)**
* **Métricas del modelo**

La cámara se controla con teclado, mouse o mediante controladores de gafas VR (como Meta Quest). Se integraron movimientos horizontales, verticales y en profundidad, así como interacción mediante puntero láser o selección por clic.

**Visualización del PCA**

En el panel del PCA (pcaMini.js), se renderiza un **gráfico de dispersión tridimensional** con las tres componentes principales que acumulan mayor varianza explicada. Cada punto representa una observación, y su color refleja su clasificación original (por ejemplo, benigno o maligno). Al seleccionar un punto, se despliega una etiqueta flotante con sus coordenadas en el nuevo espacio dimensional, su ID, tipo y contribuciones porcentuales a cada componente.

También se incluyen paneles con **las variables que más aportaron a cada componente**, lo cual proporciona una guía interpretativa sobre qué características originales influyen en cada eje principal del espacio transformado.

**Visualización de Regresión Logística**

El módulo de regresión (regresionMini.js) permite seleccionar tres variables independientes (o el PCA completo) para construir un modelo de regresión logística binaria. Los puntos se visualizan en el mismo espacio 3D, coloreados según el resultado real y la predicción del modelo:

* **Rojo**: maligno clasificado correctamente
* **Verde**: benigno clasificado correctamente
* **Café o negro**: error de clasificación

Además, se muestran las métricas del modelo en un panel de resultados (results-panel), incluyendo **exactitud, precisión, recall, F1-score, especificidad, tasa de error y matriz de confusión**.

**Panel de EDA (Exploración de Datos)**

Se añadió un componente específico (edaMini.js) para mostrar de forma resumida:

* Estadísticos descriptivos de las variables numéricas
* Cantidad de valores faltantes por variable
* Identificación de outliers (por IQR)
* Correlaciones entre pares de variables

Este módulo se muestra como un panel interactivo, donde el usuario puede navegar entre secciones mediante botones virtuales y explorar las distintas capas del análisis exploratorio.

**Interacción y selección de variables**

Mediante el componente selectorVR.js, se integró un menú principal para seleccionar las variables a analizar. El usuario puede:

* Elegir variables independientes para X, Y y Z (o usar directamente el PCA)
* Habilitar o deshabilitar la estandarización de datos
* Iniciar la visualización con un botón interactivo
* Limpiar selecciones y reiniciar el análisis

**Experiencia inmersiva**

Todo el entorno es compatible con dispositivos VR. Se configuraron eventos específicos para los botones de los controles de mano de las gafas (por ejemplo, triggerdown, gripdown, abuttondown, xbuttondown) para garantizar una navegación fluida y una experiencia completamente inmersiva.

**RESULTADOS**

El proyecto culminó en el desarrollo de una **sala virtual interactiva en realidad virtual (VR)** que integra exhaustivamente los distintos componentes de análisis estadístico y visualización de datos:

**1. Exploración de Datos (EDA)**

Antes de aplicar modelos, la plataforma ofrece un **panel de EDA** que resume automáticamente:

* **Estadísticos descriptivos** (mínimo, máximo, media, desviación estándar) de cada variable numérica.
* **Conteo de valores faltantes** y **detección de outliers** mediante el método IQR.
* **Matriz de correlaciones** que resalta los pares de variables con mayor coeficiente de correlación.

Este prerequisito facilita identificar limpiezas de datos necesarias y relaciones lineales antes de proceder al modelado.

**2. Reducción de Dimensionalidad (PCA)**

En la sección de PCA:

* Cada observación se representa como una **esfera** en un espacio 3D cuyas coordenadas corresponden a las tres primeras componentes principales (PC1, PC2, PC3), que en conjunto explican la mayor parte de la varianza.
* Un **panel fijo** muestra:
  + El porcentaje de varianza explicada por cada componente.
  + El **top 3** de variables originales más influyentes en cada componente.
* Al **seleccionar** un punto (clic o “trigger” VR), aparece una etiqueta dinámica con:
  + **ID** y **categoría** del individuo (por ejemplo, benigno/maligno).
  + **Coordenadas numéricas** en PC1, PC2 y PC3.
  + **Contribución porcentual** de ese individuo a la varianza de cada componente.

**3. Modelado Predictivo (Regresión Logística)**

La plataforma incorpora un módulo de regresión logística que permite:

1. **Seleccionar** tres variables independientes (o emplear directamente las tres PCs).
2. **Definir** si se estandarizan los predictores.
3. **Entrenar** y **evaluar** el modelo sobre todo el conjunto de datos.

Los resultados se visualizan de dos maneras:

* En el espacio 3D, las mismas observaciones se colorean según el **resultado real** y la **predicción** del modelo:
  + **Rojo**: maligno correctamente clasificado
  + **Verde**: benigno correctamente clasificado
  + **Café/Negro**: clasificaciones erróneas
* Un **panel de métricas** muestra:
  + Exactitud, precisión, recall, especificidad, F1-score y tasa de error.
  + Matriz de confusión (TP, TN, FP, FN).

**4. Navegación y Experiencia Inmersiva**

Los usuarios pueden desplazarse libremente por la sala mediante:

* **Teclas** (W/A/S/D, Shift/Space) o **joysticks** en VR.
* **Controles de mano** para selección (triggerdown), restaurar opacidades (gripdown) y mostrar paneles (abutton, bbutton, xbutton, ybutton).

Este nivel de interacción ofrece una visión completa de la estructura de los datos, facilita la detección de clústeres o valores atípicos y refuerza la comprensión de cómo los distintos análisis se complementan entre sí.

**Logros alcanzados**

1. **Automatización del EDA y preprocesamiento**
   * Desarrollo de un módulo de EDA que calcula y expone automáticamente:
     + Estadísticos descriptivos (mín, máx, media, desviación estándar).
     + Conteo de valores faltantes y detección de outliers por IQR.
     + Matriz de correlaciones entre variables numéricas.
   * Integración de estos resultados en un panel interactivo (edaMini.js), con navegación paginada y textos dinámicos.
2. **Implementación avanzada del PCA**
   * Creación de funciones personalizadas para:
     + Calcular la contribución porcentual de cada individuo a la varianza de cada componente.
     + Determinar el top 3 de variables con mayor aporte a cada componente principal.
   * Exposición de estas contribuciones en paneles fijos y etiquetado dinámico al seleccionar puntos, todo dentro de pcaMini.js.
3. **Integración de regresión logística**
   * Inclusión de un módulo de regresión logística (regresionMini.js y serverRoom.js) que permite:
     + Selección de variables (o uso directo de PCs) y opcional estandarización.
     + Cálculo y visualización de métricas clave: exactitud, precisión, recall, especificidad, F1-score, tasa de error y matriz de confusión.
4. **Diseño de una sala VR inmersiva y multidisciplinar**
   * Construcción de un entorno cerrado con texturas personalizadas y paneles temáticos para PCA, regresión, EDA.
   * Configuración de un servidor Express local que sirve recursos estáticos y rutas /eda, /pca y /regresion, habilitando WebXR y acceso desde gafas Meta Quest en LAN.
5. **Interacción y navegación fluidas**
   * Movimiento 6DOF con teclado (W/A/S/D, Shift/Space) y joysticks de VR.
   * Selección de puntos con “trigger” y restauración de opacidades con “grip”.
   * Menú de selección de variables y botones de acción (“Graficar”, “Limpiar”) diseñados para mouse y controladores VR (selectorVR.js).
6. **Arquitectura modular y escalable**
   * Separación de la lógica en componentes independientes (edaMini.js, pcaMini.js, regresionMini.js, selectorVR.js, results.js, rocViewer.js).
   * Uso de localStorage para gestión de estado y parámetros de usuario.
   * Facilidad para incorporar nuevos modelos o datasets sin reescribir la base del entorno.

Este conjunto de logros demuestra la eficacia de combinar análisis estadístico automatizado, modelado predictivo y entornos VR interactivos para ofrecer una experiencia de exploración de datos integral, intuitiva y extensible.

**Conclusión de Resultados**  
La integración de EDA, PCA y regresión logística en un entorno de realidad virtual ha generado una plataforma potente y versátil para el análisis de datos multivariados. La experiencia inmersiva facilita la detección de patrones y anomalías, acelera la interpretación de resultados y fortalece la toma de decisiones al ofrecer un contacto directo con la información. Gracias a su arquitectura modular, el sistema puede escalarse fácilmente para incorporar nuevos algoritmos, datasets o métricas, lo que lo convierte en una base robusta para futuras extensiones (por ejemplo, clustering, clasificación multiclase o análisis de series temporales). En conjunto, este proyecto demuestra que la fusión de técnicas estadísticas avanzadas con tecnologías VR no solo mejora la comprensión de datos complejos, sino que también amplía el horizonte de la analítica interactiva hacia aplicaciones más dinámicas y colaborativas.

**CONCLUSIONES**

1. Se logró desarrollar una plataforma inmersiva en realidad virtual que integra de manera fluida análisis descriptivo (EDA), reducción de dimensionalidad (PCA) y modelado predictivo (regresión logística). Esta herramienta facilita la exploración interactiva de datos multivariados, mejora la comprensión de la estructura subyacente y permite detectar errores de etiquetado al contrastar visualmente las clasificaciones reales y predichas.

2. El módulo de regresión logística se incorporó exitosamente, entrenando el modelo sobre cualquier combinación de variables seleccionadas (o directamente sobre las tres componentes principales). Su rendimiento se cuantificó con métricas estándar, precisión, recall, especificidad, F1-score y matriz de confusión integradas en la interfaz VR. Esto demuestra la viabilidad de ejecutar análisis predictivos directamente en un entorno inmersivo, sin sacrificar rigor estadístico.

3. La visualización simultánea de predicciones y etiquetas reales en el espacio 3D permite comparar de un vistazo el desempeño del modelo. Al colorear los puntos según verdadero positivo, falso negativo, etc., el usuario identifica fácilmente patrones de acierto y error. Este enfoque gráfico no solo mejora la interpretación de métricas numéricas, sino que también acelera la detección de posibles inconsistencias en el dataset.

4. Se implementó un mecanismo automático que destaca aquellos puntos cuya clasificación predicha difiere de la etiqueta original. Al seleccionarlos en la sala VR o observar su color distintivo, se facilita la inspección detallada de casos conflictivos. Esta capacidad de identificación de etiquetas incorrectas contribuye a mejorar la calidad de los datos y refuerza el ciclo de limpieza y validación dentro del análisis.

**REFERENCIAS**

* Funciones. (s/f). MDN Web Docs.
* Rodrigo, J. A. (s/f). Análisis de Componentes Principales (PCA) y t-SNE.
* Algoritmo PCA: de lo complejo a lo sencillo. (2018). LIS Data Solutions.
* Camacho, J. A. (2018). Análisis del Componente Principal PCA.
* *Papa Parse Documentation.* Papa Parse es la librería que usas para cargar CSV en el navegador y en Node.js [PapaParse](https://www.papaparse.com/docs?utm_source=chatgpt.com)[PapaParse](https://www.papaparse.com/?utm_source=chatgpt.com).
* *A-Frame Documentation.* A-Frame es el framework central de tu sala VR [A-Frame](https://aframe.io/docs/?utm_source=chatgpt.com).
* *Express.js Guide.* Para la configuración del servidor local con rutas /eda, /pca y /regresion.
* *ml-pca NPM page.* Describe la versión y opciones que utilizas.
* *ml-logistic-regression NPM page.* Detalla configuración del modelo y métodos de entrenamiento.