Universidad Mayor de San Andrés
Facultad de Ciencias Puras y Naturales
Carrera de Informatica

INF 354 – INTELIGENCIA ARTIFICIAL

Docente: Ph. D. Moisés Silva Ch.

Apellidos y Nombre: Jael Lany Valeriano Baltazar
C.I.: 9945924 LP

Fecha: 18 – 06 – 2024

PROYECTO DE INTELIGENCIA ARTIFICIAL

TÓPICO: **Stroke Prediction Dataset (Predicción de accidente cerebrovascular)**

- Descripción clara del objetivo de investigación a partir del dataset elegido.

Objetivo Principal

El objetivo principal de esta investigación es desarrollar y evaluar modelos predictivos capaces de predecir la probabilidad de que un paciente sufra un accidente cerebrovascular (ACV) basándose en diversas características demográficas, clínicas y de estilo de vida.

Objetivos Específicos

- Identificar Factores de Riesgo:
  Analizar la relación entre variables como la edad, hipertensión, enfermedades del corazón, niveles de glucosa, índice de masa corporal (IMC), estado civil, tipo de trabajo, tipo de residencia y el estado de fumador con la ocurrencia de accidentes cerebrovasculares.
- Desarrollar un Modelo Predictivo:
  Crear y entrenar varios modelos de machine learning (como regresión logística, árboles de decisión, random forest, SVM, y redes neuronales) para predecir la ocurrencia de un accidente cerebrovascular.
- Evaluar el desempeño del modelo:
  Evaluar los modelos predictivos utilizando métricas de rendimiento como la precisión, recall, F1-score y el área bajo la curva ROC (AUC-ROC).
- Interpretación de Modelos:
  Interpretar los modelos para identificar qué variables tienen mayor influencia en la predicción de accidentes cerebrovasculares.
- Implementación Práctica:
  Proponer un sistema de alerta temprana para el personal médico basado en el modelo predictivo, que pueda utilizarse para identificar pacientes en riesgo y tomar medidas preventivas oportunas.

- Descripción detallada de los campos del dataset.

***Id**: Identificación única del paciente.*
- Tipo de dato: Numérico (entero). Ejemplo: 9046

***gender**: Género del paciente.*
- Tipo de dato: Categórico. Valores posibles: "Male" (Masculino), "Female" (Femenino), "Other" (Otro). Ejemplo: "Male"

***age**: Edad del paciente.*
- Tipo de dato: Numérico (flotante). Generalmente entre 0 y 82 años. Ejemplo: 67.0

***hypertension**: Si el paciente tiene hipertensión.*
- Tipo de dato: Binario (0 o 1). Valores posibles: 0 (No), 1 (Sí). Ejemplo: 0

***heart_disease**: Si el paciente tiene una enfermedad del corazón.*
- Tipo de dato: Binario (0 o 1). Valores posibles: 0 (No), 1 (Sí). Ejemplo: 1

***ever_married**: Si el paciente alguna vez estuvo casado.*
- Tipo de dato: Categórico. Valores posibles: "No" (No), "Yes" (Sí). Ejemplo: "Yes"

***work_type**: Tipo de trabajo del paciente.*
- Tipo de dato: Categórico. Valores posibles: "children" (Niño), "Govt_job" (Trabajo en el gobierno), "Never_worked" (Nunca ha trabajado), "Private" (Privado), "Self-employed" (Autónomo). Ejemplo: "Private"

***Residence_type**: Tipo de residencia del paciente.*
- Tipo de dato: Categórico. Valores posibles: "Urban" (Urbano), "Rural" (Rural). Ejemplo: "Urban"

***avg_glucose_level**: Nivel promedio de glucosa en la sangre del paciente.*
- Tipo de dato: Numérico (flotante). Ejemplo: 228.69

***bmi**: Índice de Masa Corporal (IMC) del paciente.*
- Tipo de dato: Numérico (flotante). Ejemplo: 36.6

***smoking_status**: Estado de fumador del paciente.*
- Tipo de dato: Categórico. Valores posibles: "formerly smoked" (Ha fumado anteriormente), "never smoked" (Nunca ha fumado), "smokes" (Fuma actualmente), "Unknown" (Desconocido). Ejemplo: "formerly smoked"

stroke: Indica si el paciente ha tenido un accidente cerebrovascular.
- Tipo de dato: Binario (0 o 1). Valores posibles: 0 (No), 1 (Sí). Ejemplo: 1

- Proceso básico de análisis de datos:

1. Preprocesamiento (al menos una valida, otros dos por ver los resultados, si no se aplica justifique porque). Balanceo de datos.

Justificación de Técnicas de Preprocesamiento y Balanceo
- Valores Faltantes: Se han manejado porque los valores faltantes pueden limitar el análisis y reducir la precisión del modelo.
- Codificación de Variables Categóricas: Necesaria para usar estas variables en modelos de machine learning.

- Normalización o Escalado: Algunos modelos de machine learning, como SVM y KNN, requieren que las características estén en la misma escala. (INVESTIGADO)
- Balanceo de Datos: El balanceo es crucial porque un dataset desbalanceado puede llevar a que el modelo se sesgue hacia la clase mayoritaria, ignorando la clase minoritaria, que en este caso es la de mayor interés.

2. Selección del clasificador (acorde a los datos supervisado, no supervisado). El clasificador puede, pero no necesariamente depender del preprocesamiento.
   a. Justificar el clasificador (máximo 2 planas con fuente ISBN, DOI).

Selección del Clasificador para Predicción de Accidentes Cerebrovasculares

Clasificadores Supervisados

Si el objetivo es predecir si un paciente sufrirá un accidente cerebrovascular (stroke) basado en características demográficas y clínicas, la naturaleza del problema es de clasificación supervisada. Los modelos supervisados son adecuados porque se tienen datos etiquetados donde la variable objetivo es conocida. A continuación, se presentan dos clasificadores populares en el contexto de la predicción de accidentes cerebrovasculares y justificaciones basadas en fuentes académicas.

1. **Random Forest**

Random Forest es un conjunto de árboles de decisión que operan como un equipo, donde cada árbol en el bosque emite un voto y la clase que obtiene la mayoría de los votos se convierte en la predicción del modelo. Este algoritmo es robusto, reduce el sobreajuste y maneja bien los datos no lineales.

Justificación

- Robustez y Precisión: Random Forest tiende a ser robusto contra el sobreajuste, especialmente en datasets de alta dimensionalidad y con características heterogéneas [ISBN: 978-3319212119].
- Manejo de Datos Desbalanceados: Con técnicas de ajuste como ponderación de clases y SMOTE, Random Forest puede manejar problemas de desequilibrio de clases, lo cual es crítico en este dataset [DOI: 10.1007/s10994-007-5022-7].
- Interpretabilidad: Aunque no es tan interpretable como modelos lineales, las importancias de las características proporcionadas por Random Forest pueden ayudar a entender las contribuciones de cada variable [ISBN: 978-1461471370].

2. **Support Vector Machine (SVM)**

SVM es un clasificador que busca un hiperplano en un espacio multidimensional que separa las clases de manera óptima. Es eficaz en espacios de alta dimensionalidad y puede utilizar diferentes funciones kernel para manejar relaciones no lineales.

Justificación

- Eficacia en Alta Dimensionalidad: SVM es eficaz en espacios de alta dimensionalidad, lo que es útil cuando se tienen muchas características [ISBN: 978-0387301737].

- Versatilidad: Utilizando diferentes kernels (lineal, polinomial, RBF), SVM puede modelar relaciones complejas y no lineales entre las características [DOI: 10.1023/A:1010933404324].
- Maximización del Margen: SVM maximiza el margen entre las clases, lo que a menudo lleva a una mejor generalización en conjuntos de datos no vistos [ISBN: 978-0262033589].

3. Primera ejecución: Confiabilidad, matriz de confusión.
4. Splits: al menos 100 asignaciones, la mediana de la confiabilidad.
   Académico (primera ejecución) 80(train)/20(test) – Investigación 50/50 (segunda ejecución).
5. Primer Código: Github, kaagle, codelab.
   Aplicar Componentes principales (PCA), determinar la cantidad óptima para mejorar o llegar al resultado anterior. Al menos unas 5 ejecuciones. (12 columnas, 10, 11, 9, 5, 3).
   a. Explicar cómo funciona PCA (álgebra lineal).
   b. Escribir un artículo de al menos 4 páginas.
6. Segundo Código: Github, kaagle, codelab.
   a. Artículo que resuma lo realizado (min 3 hojas).