26/012025

Predicción del riesgo de Diabetes Tipo II con Machine Learning

CIENCIA DE DATOS

RUBÉN GARRIDO HIDALGO

Descripción: Basándonos en trabajos estudiados y trabajados en clase de forma individual sobre modelos de machine Learning para la predicción de enfermedades cardiacas, voy a aplicar mis conocimientos en machine learning para realizar un modelo de predicción sobre pacientes de la diabetes tipo II. La diabetes tipo II es una enfermedad común en la sociedad y consiste en una afección crónica que afecta a la forma en la que el cuerpo procesa la glucosa generada por el propio cuerpo y adquirida en los alimentos. En este caso vamos a construir un modelo de aprendizaje automático que evaluará el riesgo de padecer diabetes de ciertos pacientes de los cuales se han registrado diversos datos médicos, personales y demográficos. El objetivo de este trabajo es modelar modelos de machine learning para obtener estrategias lo suficientemente confiables para prevenir la diabetes tipo II o saber con certeza si una persona va a sufrirla y tratarla a tiempo. Pues, aunque sea crónica puede regularse con medicamentos y bombas de insulina.

Sector: El sector principal es evidentemente el sector médico, trabajaremos con hospitales (públicos y privados), clínicas y centros de salud que buscan la integración de nuevas tecnologías para avanzar con el tratamiento de enfermedades crónicas, así como para programas de prevención de esta u otras enfermedades.

Descripción de los datos:

Tipo de fuente de datos:

Los datos que vamos a emplear para enfrentar nuestro modelo y realizar predicciones serán extraídos de las historias clínicas de los pacientes (datos médicos y demográficos), las cuales podemos encontrarlas en grandes bases de datos en los hospitales, encuestas personales realizadas a los pacientes del hospital sobre su estilo de vida y hábitos, así como pruebas realizadas de forma rutinaria a los pacientes como control de la glucosa, toma de la tensión, etcétera.

Volumen: Emplearemos unos 50.000 registros de pacientes recolectados durante ocho años en un hospital de Sevilla, este volumen de datos es lo suficientemente grande como para asegurar que podemos entrenar un modelo predictivo robusto.

Variables: Vamos a especificar las variables que vamos a emplear

1. **Demográficas:** Edad, género, antecedentes familiares de diabetes.
2. **Fisiológicas:** Glucosa en ayunas, índice de masa corporal (IMC), presión arterial, niveles de insulina.
3. **Estilo de vida:** Nivel de actividad física, consumo de tabaco, dieta rica en carbohidratos o grasas.
4. **Resultados médicos:** Historial de hipertensión, niveles de colesterol (HDL y LDL), resistencia a la insulina.

d) Algoritmos de aprendizaje automático.

Para modelar nuestro modelo de aprendizaje automático he probado empleando diferentes algoritmos de aprendizaje automático empleados en la clasificación de datos, entre ellos la regresión logística la cual como sabemos es un modelo estadístico que se emplea para calcular la probabilidad de que un caso pertenezca a una categoría específica. He intentado implementar este porque es sencillo de implementar y fácil de interpretar, menos propensos al sobreajuste y con un buen rendimiento en las predicciones. Pero en este caso no será el mejor algoritmo que podamos emplear, pues el tipo de código al ser tan variado y complejo incluirá felaciones no lineales en los datos y estas relaciones no son capturadas por la regresión logística.

Por lo tanto, basándome en las variables y el tipo de datos el mejor algoritmo de aprendizaje automático según mi criterio es el Random Forest. Este es un algoritmo de aprendizaje supervisado que combina múltiples árboles de decisión para realizar predicciones más precisas. Lo he seleccionado porque es muy robusto en las predicciones y captura relaciones no lineales. En este algoritmo cada árbol de decisión se entrena con una muestra diferente del conjunto de datos y selecciona la clase con más peso para tomarla comom la predicción final.

Aplicación en este caso:

En este caso entrenaremos el modelo Random Forest con datos como la edad de cada paciente, género, nivel de glucosa, BMI, nivel de colesterol y otras variables para poder clasificar a los pacientes en función del riesgo de padecer la diabetes tipo-II:

1. Bajo riesgo

2. Riesgo medio

3. Riesgo alto

Código encontrado: en cuanto a este caso no he encontrado código como tal, lo he diseñado yo. He implementado un modelo (Random Forest) de 100 árboles de decisión sobre una base de datos extraída de Kaggle que adjuntaré a la entrega.

Las características de este van como anotaciones entre las líneas de código, como la métrica para evaluar la capacidad de clasificación, la división del conjunto de entrenamiento y demás. Adjunto varias capturas del programa además de adjuntarlo con la entrega:

Datos:

[8]:
```
ruta = 'C:/Users/Rubén Garrido/Desktop/Ciencia de Datos/diabetes.csv'
data = pd.read_csv(ruta)
print(str(data))
```

```
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0              6      148             72             35        0  33.6
1              1       85             66             29        0  26.6
2              8      183             64              0        0  23.3
3              1       89             66             23       94  28.1
4              0      137             40             35      168  43.1
..           ...      ...            ...            ...      ...   ...
763           10      101             76             48      180  32.9
764            2      122             70             27        0  36.8
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age  Outcome
0                       0.627   50        1
1                       0.351   31        0
2                       0.672   32        1
3                       0.167   21        0
4                       2.288   33        1
..                        ...  ...      ...
763                     0.171   63        0
764                     0.340   27        0
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0

[768 rows x 9 columns]
```

Algoritmo de Random Forest:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

# Importamos los datos
ruta = 'C:/Users/Rubén Garrido/Desktop/Ciencia de Datos/diabetes.csv'
data = pd.read_csv(ruta)

# Definino las variables independientes (X) y la variable dependiente (y)
X = data[['Pregnancies', 'Glucose', 'BMI ', 'BloodPressure', 'Insulin']]  # Variables predictoras
y = data['DiabetesPedigreeFunction']
# Variable objetivo: 0 (bajo), 1 (medio), 2 (alto)

# División de los datos en conjuntos de entrenamiento (80%) y prueba (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Creación y entrenamiento del modelo de Random Forest con 100 árboles
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(X_train, y_train)

# Hacer predicciones en el conjunto de prueba
y_pred = model.predict(X_test)

# Evaluamos el modelo (precisión)
accuracy = accuracy_score(y_test, y_pred)
print(f"Precisión del modelo: {accuracy:.2f}")
print(classification_report(y_test, y_pred))

# Analizamos la importancia de las variables
importances = pd.DataFrame({'Variable': X.columns, 'Importancia': model.feature_importances_})
print(importances.sort_values(by='Importancia', ascending=False))
```

Herramientas que se usan:

Como lenguaje principal en el que hemos desarrollado el algoritmo tenemos Python, en mi caso particular estoy trabajando en un entorno de Jupyter Lab un entorno para desarrollar y documentar todo mi análisis. En Python he implementado librerías como:

- Pandas: para manejar y manipular los datos
- Scikit-learn: para entrenar y evaluar el modelo

Además, podríamos emplear la librería matplotlib para visualizar una visualización de los datos y resultados.

Como base de datos, como he mencionado antes he empleado un archivo CSV extraído de Kaggle, dicho archivo fue publicado hace siete años y es de dominio público.

Valor que aporta a la ciencia de datos:

Algunos de los aportes principales que hace a la ciencia de datos es personalizar los tratamientos a los pacientes, permitiendo que los médicos puedan evitar ciertas confusiones en los datos o realicen predicciones erróneas. Este tipo de algoritmo ayuda a corroborar las predicciones de los médicos e incluso abren nuevas líneas de pensamiento y de pronóstico de enfermedades que a simple vista un médico no piensa en ellas.

También en casos como el de la diabetes II, estos modelos ayudan a poder deducir los factores de riesgo y así poder diseñar campañas para concienciar a la gente sobre los peligros que suponen ciertos hábitos.

Referencias:

Scikit-learn Documentation:

(https://scikit-learn.org/stable/index.html)

Health Data Science Research: Diabetes Risk Prediction

(https://www.ncbi.nlm.nih.gov/)

Factores de riesgo de la Diabetes

(https://www.cdc.gov/diabetes/es/risk-factors/factores-de-riesgo-de-la-diabetes.html)

Beneficios de la ciencia de datos para la salud

(https://blogs.uoc.edu/informatica/es/beneficios-ciencia-datos-para-salud/)

Información Random Forest

(https://interactivechaos.com/es/wiki/random-forest)

Laboratorio_1:

Ejercicio 2: Realiza una tabla comparativa entre las plataformas de Big Data (AWS, Google Cloud, Microsoft Azure). Dicha tabla debe comparar los siguientes elementos:

a) Máquinas virtuales y servidores.

b) Escalabilidad.

c) Procesamiento.

d) Almacenamiento.

e) Servicios.

f) Puntos fuertes.

g) Precio.

h) Referencias.

Referencias

Información sobre Amazon AWS:

(Productos de Amazon AWS, s.f.)

https://aws.amazon.com/es/big-data/datalakes-and-analytics/

Información sobre Google Cloud:

(Google Cloud, s.f.)

https://cloud.google.com/?_gl=1*1nkcng5*_up*MQ..&gclid=c5b64d38c33d12ac48302aadb1cb7b32&gclsrc=3p.ds

Información de Microsoft Azure:

(Productos Microsoft Azure, s.f.) https://azure.microsoft.com/es-es/products

(Form.io, s.f.)

Webs que muestran comparaciones entre las plataformas:

(Tech Radar, s.f.)

https://global.techradar.com/es-es

(Petersen, s.f.)

https://www.linkedin.com/pulse/comparaci%C3%B3n-de-las-principales-plataformas-en-la-nube-juan-carlos-glowf/?originalSubdomain=es

Característica	AWS(Amazon)	Google Cloud	Microsoft Azure
Máquinas virtuales y servidores.	EC2: Ofrece instancias flexibles y configurables para ejecutar cargas de trabajo.	Compute Engine: Proporciona máquinas virtuales personalizables para diversas aplicaciones.	Azure Virtual Machines: Ofrece una amplia variedad de configuraciones, con enfoque en escalabilidad.
Escalabilidad.	Escalabilidad automática	Escalabilidad automática en Google Kubernetes Engine (GKE) y Google Compute Engine.	Escalabilidad automática con Azure Virtual Machines
Procesamiento	Amazon EMR: Proporciona un entorno escalable para Big Data, con soporte para Hadoop, Spark, etc.	Google Dataproc: Ofrece un servicio administrado para Hadoop y Spark.	Azure HDInsight
Almacenamiento	Amazon S3: Almacenamiento escalable y duradero, ideal para Big Data.	Google Cloud Storage: Almacenamiento duradero y escalable, con integración en toda la plataforma.	Azure Blob Storage: Almacenamiento masivo y económico, ideal para datos no estructurados.
Servicios	AWS Lambda, Amazon Redshift,	BigQuery: Análisis de datos sin necesidad de infraestructura, como Google Dataflow, Google Pub/Sub.	Azure SQL Data Warehouse, Azure Data Lake Analytics,
Puntos fuertes	Amplia gama de servicios y ecosistema maduro, con gran flexibilidad y soporte para múltiples tecnologías.	BigQuery: Innovación en análisis de datos y alta integración con IA y machine learning	Integración fluida con herramientas de Microsoft (Excel, Power BI) y amplio soporte para empresas con alta seguridad
Precio	Pago por uso, con precios competitivos con descuentos por uso reservado.	Precios basados en pago por uso y descuentos para uso a largo plazo	Precios basados en pago por uso y precios reservados