Análisis Predictivo y Gestión de Datos

Sesión 1: Introducción al Análisis Predictivo

Oscar Leonardo Rincón León

24 de abril de 2025

Objetivos de la sesión

- Comprender qué es el análisis predictivo
- Introducir conceptos básicos de aprendizaje estadístico y Machine Learning
- Conocer los tipos de datos utilizados en predicción
- Identificar herramientas clave como Excel y Python
- Aplicar los conceptos en contextos sociales y financieros

¿Qué es el análisis predictivo?

- Es una disciplina que combina estadísticas, ciencia de datos y aprendizaje automático.
- Utiliza datos históricos para anticipar comportamientos o resultados futuros.
- Permite modelar relaciones entre variables independientes (predictoras) y una variable objetivo.

Ejemplos de aplicación

Contexto social:

- Predecir deserción escolar.
- Identificar hogares con riesgo de pobreza extrema.
- Priorizar beneficiarios de subsidios o programas sociales.

Contexto financiero y administrativo:

- Anticipar incumplimientos de pago (riesgo crediticio).
- Clasificar clientes según rentabilidad o riesgo.
- Proyectar demanda de servicios o productos.

¿Qué es el aprendizaje estadístico?

- Es un conjunto de herramientas que permite:
  - Comprender la estructura de los datos.
  - Estimar relaciones entre variables.
  - Realizar predicciones sobre nuevos datos.
- Está en la base de muchas técnicas modernas de Machine Learning.
- Su objetivo principal es estimar una función f tal que $Y \approx f(X)$.

Tipos de aprendizaje estadístico

- **Aprendizaje supervisado:**
  - El conjunto de datos incluye una variable objetivo Y conocida.
  - Se busca predecir Y a partir de X.
  - Ejemplos: regresión, clasificación.
- **Aprendizaje no supervisado:**
  - No se observa una variable objetivo.
  - El objetivo es encontrar patrones o estructuras internas.
  - Ejemplos: clustering, reducción de dimensionalidad.

¿Por qué usar aprendizaje estadístico?

- Permite construir modelos predictivos a partir de evidencia empírica.
- Se utiliza cuando no es posible modelar fenómenos con fórmulas teóricas exactas.
- Es útil en áreas como:
  - Salud, educación, banca, gobierno, mercadeo.
- También permite explorar grandes volúmenes de datos donde otras técnicas fallan.

Tipos de problemas en análisis predictivo

- En análisis predictivo, los problemas se dividen según el tipo de variable que queremos predecir.
- Esta clasificación determina:
  - El tipo de modelo a usar.
  - La métrica de evaluación adecuada.
  - El preprocesamiento necesario.
- Existen dos tipos principales: **regresión** y **clasificación**.

Problemas de regresión

- Ocurren cuando la variable objetivo Y es **numérica y continua**.
- El objetivo es predecir un valor aproximado dentro de un rango posible.
- **Ejemplos:**
  - Predecir el ingreso mensual de una persona.
  - Estimar el costo de un proyecto social.
  - Calcular la demanda futura de un producto o servicio.
- Modelos típicos: regresión lineal, regresión de árboles, redes neuronales regresivas.

Problemas de regresión

- El objetivo del análisis predictivo es estimar una función $f(X)$ que prediga con precisión una variable objetivo Y.
- **En regresión:** se busca minimizar el error cuadrático medio entre el valor real y la predicción:

$$\min_f \ \mathbb{E}\left[(Y - f(X))^2\right]$$

Problemas de clasificación

- Se presentan cuando la variable objetivo Y es **categórica o discreta**.
- El objetivo es asignar cada observación a una clase o grupo definido.
- **Ejemplos:**
  - Determinar si un estudiante abandonará o no la escuela.
  - Clasificar si una persona es elegible o no para un subsidio.
  - Detectar si una transacción financiera es fraudulenta.
- Modelos típicos: regresión logística, KNN, árboles de decisión, redes neuronales.

Problemas de clasificación

- **En clasificación:** se busca predecir la clase más probable:

$$\hat{Y} = \arg\max_{c} \; P(Y = c \mid X)$$

- Este planteamiento guía la elección del modelo y de la métrica de evaluación.

Tipos de datos en predicción

- **Numéricos continuos:**
  - Toman cualquier valor dentro de un rango.
  - Ejemplos: edad, ingreso mensual, horas de trabajo semanales.
- **Categóricos (nominales):**
  - No tienen orden intrínseco.
  - Ejemplos: género, municipio, tipo de programa.
- **Ordinales:**
  - Tienen un orden natural entre categorías.
  - Ejemplos: nivel educativo (primaria ¡ secundaria ¡ universidad), satisfacción del usuario (baja, media, alta).
- **Temporales:**
  - Representan el tiempo: fechas, años, trimestres.
  - Ejemplos: fecha de inscripción, año de nacimiento.

Importancia del tipo de dato en modelado

- Cada tipo de dato requiere un tratamiento específico para usarse correctamente en un modelo:
  - **Numéricos:** pueden usarse directamente o ser escalados.
  - **Categóricos:** necesitan ser codificados (por ejemplo, one-hot encoding).
  - **Ordinales:** pueden representarse con enteros, respetando su orden.
  - **Temporales:** pueden descomponerse en componentes útiles (año, mes, estacionalidad).
- Usar mal un tipo de dato puede distorsionar los resultados del modelo.
- El preprocesamiento adecuado es clave para un buen desempeño predictivo.

¿Cómo se clasifican los sistemas de Machine Learning?

- El aprendizaje automático (Machine Learning) abarca una gran variedad de métodos y algoritmos.
- Estos sistemas pueden clasificarse desde distintas perspectivas según:
  - El tipo de datos disponibles para entrenar el modelo.
  - La forma en que el modelo procesa los datos.
  - La estrategia que sigue para hacer predicciones.
- Entender estas diferencias nos ayuda a elegir el enfoque adecuado para cada problema.

Tipos de aprendizaje automático

- **Aprendizaje supervisado:**
  - El modelo aprende a partir de datos etiquetados (con una variable objetivo conocida).
  - Ejemplo: predecir si un hogar es vulnerable según sus características.
- **Aprendizaje no supervisado:**
  - No se dispone de una variable objetivo.
  - Se busca descubrir estructuras o patrones ocultos.
  - Ejemplo: segmentar municipios en grupos similares de forma automática.

Otras formas de clasificar los sistemas de ML

- **Por lotes (batch) vs. en línea (online):**
  - En batch, el modelo se entrena con todos los datos de una vez.
  - En línea, el modelo se actualiza progresivamente con cada nuevo dato.
- **Basado en instancias vs. basado en modelos:**
  - Los modelos basados en instancias (como KNN) guardan los datos y comparan nuevos casos.
  - Los modelos basados en funciones (como regresión) aprenden una regla general a partir de los datos.
- **Este curso:** usaremos principalmente aprendizaje **supervisado por lotes**, con modelos funcionales y de instancia.

Herramientas de software para análisis predictivo

- **Excel:**
  - Útil para exploración rápida y reportes básicos.
  - Limitado en volumen y capacidad para modelado complejo.
- **SQL (Structured Query Language):**
  - Lenguaje estándar para consultar bases de datos relacionales.
  - Permite seleccionar, agrupar, unir y filtrar datos antes de analizarlos.
  - Muy usado en contextos institucionales, financieros y públicos.
- **Python + Jupyter Notebook:**
  - Plataforma principal del curso.
  - Permite documentar análisis, ejecutar código y visualizar resultados.
  - Librerías clave:
    - `pandas`, `matplotlib`, `seaborn`, `scikit-learn`

Otras herramientas relevantes

- **R:** Lenguaje poderoso para análisis estadístico y visualización. Usado en investigación y salud pública.
- **Stata / SPSS:** Entornos amigables con menú. Muy usados en economía y ciencias sociales.
- **Power BI / Tableau:** Herramientas visuales para presentar resultados a públicos no técnicos.
- **SAS / KNIME / RapidMiner:** Plataformas de pago o gratuitas para modelado visual sin programación.
- **En este curso:** Usaremos Python + Jupyter como herramienta principal, pero reconoceremos el ecosistema completo.

Resumen de la sesión (1/2)

- El análisis predictivo permite anticipar resultados y apoyar decisiones estratégicas mediante el uso de datos históricos y modelos matemáticos.
- Comprender la naturaleza de los datos es esencial para su uso en modelos:
  - Numéricos, categóricos, ordinales, temporales.
  - Cada tipo requiere un tratamiento diferente.
- Los problemas que resolveremos se dividen en regresión y clasificación, según el tipo de variable objetivo.

Resumen de la sesión (2/2)

- Un modelo predictivo no solo debe ser preciso, también debe:
  - Ser útil en su contexto.
  - Ser comprensible para quienes toman decisiones.
  - Generar valor en la práctica.
- Durante el curso, construiremos soluciones aplicadas a contextos reales:
  - Proyectos sociales, programas públicos y administración financiera.

Cierre de la sesión y lo que viene

Conclusiones:

- Hoy dimos nuestros primeros pasos en análisis predictivo: qué es, para qué sirve, y cómo se relaciona con los datos.
- Exploramos un conjunto de datos reales para reconocer variables numéricas y categóricas, e interpretar visualmente su distribución.
- Comprobamos que observar los datos antes de modelar es esencial para detectar errores, entender patrones y tomar buenas decisiones.

En la próxima sesión:

- Nos enfocaremos en la **preparación y limpieza de datos**:
  - Tratamiento de valores faltantes
  - Codificación de variables categóricas
  - Escalado y normalización
- Estas transformaciones son fundamentales para que los modelos predictivos funcionen correctamente.