# **Trabajo Práctico: Implementación y Análisis de Algoritmos de Regresión Lineal**

**Información General**

**Modalidad:** Individual  
**Herramientas requeridas:** Python, bibliotecas numpy, pandas, matplotlib, seaborn, scikit-learn

**Objetivos**

Al finalizar este trabajo práctico, el estudiante será capaz de:

* Distinguir y clasificar los diferentes tipos de algoritmos de aprendizaje supervisado
* Comprender los fundamentos teóricos de la regresión lineal simple y múltiple
* Implementar desde cero un algoritmo de regresión lineal simple
* Aplicar técnicas de preparación y evaluación de datos
* Comparar resultados entre implementación propia y bibliotecas especializadas
* Analizar e interpretar resultados de modelos de regresión

**Contexto del Problema**

Una empresa agroexportadora peruana especializada en productos como palta, quinua, café y arándanos desea optimizar la predicción de rendimientos de cultivos y precios de exportación. La empresa opera en diferentes regiones del país (costa, sierra y selva) y necesita un sistema predictivo que le permita tomar decisiones estratégicas sobre qué cultivos priorizar, en qué regiones invertir, y cómo establecer precios competitivos en mercados internacionales.

Como científico de datos, debes desarrollar modelos predictivos que consideren variables como condiciones climáticas, características del suelo, altitud, precipitaciones, y datos históricos de producción para predecir tanto el rendimiento por hectárea como los precios de exportación.

**Parte 1: Marco Teórico y Análisis Conceptual (1 hora)**

**1.1 Taxonomía del Aprendizaje Supervisado**

Elabore un documento que incluya:

* **Definición completa** del aprendizaje supervisado y sus características principales
* **Clasificación detallada** de los algoritmos de aprendizaje supervisado:
  + Algoritmos de clasificación: definición, características, ejemplos de uso en agricultura
  + Algoritmos de regresión: definición, características, ejemplos de uso en agroexportación
* **Comparación** entre clasificación y regresión: similitudes, diferencias, criterios de selección
* **Ejemplos prácticos** específicos del sector agroindustrial peruano y latinoamericano

**1.2 Fundamentos de Regresión Lineal**

Desarrolle una explicación comprehensiva que cubra:

* **Regresión lineal simple**: definición matemática, asunciones del modelo, interpretación geométrica
* **Regresión lineal múltiple**: definición, ventajas sobre la simple, casos de aplicación en agricultura
* **Diferencias clave** entre ambos tipos de regresión
* **Limitaciones y restricciones** de los modelos lineales en contextos agrícolas
* **Métricas de evaluación**: R², MSE, MAE, RMSE (definición e interpretación en el contexto agroindustrial)

**Parte 2: Análisis Exploratorio y Preparación de Datos (1.5 horas)**

**2.1 Obtención y Exploración del Dataset**

Utilizando un dataset relacionado con agricultura/agroexportación (puede ser sintético o adaptado):

**Variables sugeridas para el dataset:**

* Rendimiento por hectárea (variable objetivo para modelo 1)
* Precio de exportación USD/kg (variable objetivo para modelo 2)
* Precipitación anual (mm)
* Temperatura promedio (°C)
* Altitud (msnm)
* Tipo de suelo (categórica: arcilloso, arenoso, franco)
* Región (categórica: Costa, Sierra, Selva)
* Hectáreas cultivadas
* Inversión en fertilizantes (USD/hectárea)
* Días de riego
* pH del suelo

**Tareas a realizar:**

* **Cargar y examinar** la estructura del dataset
* **Identificar** variables dependientes e independientes
* **Analizar** tipos de datos, valores faltantes, outliers
* **Generar estadísticas descriptivas** completas considerando las particularidades regionales del Perú

**2.2 Análisis Exploratorio de Datos (EDA)**

Realizar un análisis visual comprehensivo:

* **Distribuciones** de variables numéricas mediante histogramas y boxplots
* **Análisis por regiones** (Costa, Sierra, Selva) y sus características distintivas
* **Matriz de correlación** entre todas las variables
* **Gráficos de dispersión** entre variables predictoras y las variables objetivo
* **Análisis estacional** si hay datos temporales disponibles
* **Identificación de patrones** específicos del sector agroindustrial peruano

**2.3 Preparación y Limpieza de Datos**

Implementar las siguientes tareas de preprocessamiento:

* **Tratamiento de valores faltantes** considerando la naturaleza de datos agrícolas
* **Codificación** de variables categóricas (región, tipo de suelo)
* **Normalización o estandarización** de variables (justificar considerando las diferentes escalas)
* **División del dataset** en conjuntos de entrenamiento y prueba (70-30 o 80-20)
* **Documentar** todas las decisiones tomadas durante la preparación

**Parte 3: Implementación de Regresión Lineal Simple (2 horas)**

**3.1 Implementación desde Cero**

Desarrollar una implementación completa del algoritmo:

* **Seleccionar** una variable predictora con mayor correlación (ej: precipitación vs rendimiento)
* **Implementar** el cálculo de coeficientes usando el método de mínimos cuadrados
* **Crear funciones** para:
  + Entrenamiento del modelo
  + Realización de predicciones
  + Cálculo de métricas de evaluación
* **Visualizar** la línea de regresión sobre los datos de entrenamiento

**3.2 Evaluación del Modelo Simple**

Evaluar exhaustivamente el modelo implementado:

* **Calcular métricas** en conjuntos de entrenamiento y prueba
* **Generar gráficos** de valores reales vs predichos
* **Analizar residuos**: distribución, patrones, homocedasticidad
* **Interpretar** los coeficientes obtenidos en el contexto agroindustrial peruano
* **Evaluar** la viabilidad práctica para decisiones comerciales

**Parte 4: Regresión Lineal Múltiple y Comparación (1.5 horas)**

**4.1 Implementación de Regresión Múltiple**

Desarrollar un modelo con múltiples variables:

* **Seleccionar** 4-6 variables predictoras más relevantes para el contexto peruano
* **Implementar** regresión múltiple usando bibliotecas especializadas (scikit-learn)
* **Considerar** interacciones entre variables (ej: altitud × precipitación)
* **Entrenar** modelos separados para diferentes cultivos o regiones si es relevante
* **Validar** usando técnicas de validación cruzada

**4.2 Análisis Comparativo**

Realizar una comparación detallada:

* **Contrastar métricas** entre regresión simple y múltiple
* **Evaluar** la mejora en capacidad predictiva
* **Analizar** el problema de sobreajuste en el contexto de variabilidad climática
* **Interpretar** la importancia relativa de cada variable en el contexto agroindustrial
* **Considerar** la estacionalidad y variabilidad climática del Perú

**Parte 5: Análisis de Resultados y Conclusiones (30 minutos)**

**5.1 Interpretación Integral**

Desarrollar un análisis comprehensivo que incluya:

* **Evaluación** de la calidad predictiva considerando la variabilidad inherente del sector agrícola
* **Identificación** de limitaciones específicas para el contexto peruano
* **Recomendaciones estratégicas** para la empresa agroexportadora:
  + Decisiones de inversión por regiones
  + Estrategias de diversificación de cultivos
  + Optimización de recursos hídricos y fertilizantes
  + Estrategias de pricing para exportación

**5.2 Contextualización Nacional**

Incluir una sección específica sobre:

* **Factores externos** que afectan el modelo (fenómenos climáticos como El Niño/La Niña)
* **Consideraciones** de políticas gubernamentales de apoyo al agro
* **Impacto** del cambio climático en la predictibilidad de los modelos
* **Oportunidades** de mejora considerando la digitalización del agro peruano

**5.3 Reflexión Crítica**

Incluir una sección de reflexión personal sobre:

* **Desafíos específicos** del machine learning aplicado al sector agroindustrial
* **Consideraciones éticas** del uso de estos modelos para decisiones que afectan productores rurales
* **Alternativas** que podrían explorarse (modelos más sofisticados, datos satelitales, IoT)
* **Sostenibilidad** y responsabilidad social empresarial en las recomendaciones

**Entregables**

1. **Documento técnico** (formato PDF) con todo el marco teórico y análisis contextualizado
2. **Notebook de Jupyter** con código comentado, incluyendo interpretaciones en español
3. **Dashboard de visualizaciones** que simule un reporte ejecutivo para gerencia
4. **Dataset procesado** con documentación de variables
5. **Resumen ejecutivo** (2 páginas) con recomendaciones estratégicas para la agroexportadora
6. **Anexo** con glosario de términos agroindustriales utilizados

**Criterios de Evaluación**

|  |  |  |
| --- | --- | --- |
| Criterio | Porcentaje | Descripción |
| Completitud teórica | 20% | Profundidad y precisión del marco conceptual |
| Calidad del código | 25% | Claridad, documentación y funcionalidad |
| Análisis contextual | 20% | Comprensión del sector agroindustrial peruano |
| Evaluación de modelos | 20% | Rigor en la evaluación y comparación |
| Aplicabilidad práctica | 15% | Calidad de recomendaciones y viabilidad comercial |

**Recursos Adicionales Sugeridos**

* Datos de MINAGRI (Ministerio de Agricultura del Perú)
* Información de SENASA sobre exportaciones
* Reportes de ADEX sobre agroexportación peruana
* Datos climáticos de SENAMHI
* Estudios de caso de empresas agroexportadoras peruanas exitosas

**Nota:** Se valorará especialmente la capacidad de conectar los conceptos técnicos con la realidad del sector agroindustrial peruano, considerando sus desafíos específicos como la variabilidad climática, la geografía diversa, y las oportunidades de crecimiento en mercados internacionales.