Universidad Tecnológica Metropolitana (UTEM)

Facultad de Ingeniería — Departamento de Informática

Informe de Prueba 1 y Anteproyecto del Proyecto Integrador

Predicción de la calidad del aire en Santiago utilizando ciencia de datos

Helen Maureira Barrenechea

Asignatura: Herramientas para Ciencia de Datos
Profesor: Dr. Ing. Michael Miranda Sandoval

Octubre 2025

Índice

1. Resumen

El presente documento corresponde al desarrollo de la **Primera Prueba del Segundo Semestre 2025** para la asignatura *Herramientas para Ciencia de Datos*, junto con la presentación del **anteproyecto del Proyecto Semestral Integrador**.

Durante la prueba se abordaron cinco ejercicios prácticos orientados al manejo de datos estructurados y no estructurados, el uso de librerías distribuidas como *Dask* y *PySpark*, técnicas de visualización en Python y la implementación de un modelo de perceptrón desde cero.

El anteproyecto propone la creación de un modelo predictivo de calidad del aire en Santiago, basado en datos históricos y variables meteorológicas, aplicando técnicas de limpieza, modelado y visualización aprendidas en la asignatura.

2. Desarrollo de la Prueba

2.1. Ejercicio 1: Comparación de Datasets Estructurados y No Estructurados

El primer ejercicio tuvo como propósito distinguir entre los diferentes tipos de datos que se utilizan en ciencia de datos y analizar sus implicancias prácticas. Se trabajó con un dataset estructurado —el clásico conjunto *Iris*— y con un ejemplo de datos no estructurados, compuesto por fragmentos de texto y descripciones meteorológicas.

Los datos estructurados se caracterizan por su organización en filas y columnas, lo que facilita su análisis mediante herramientas como *Pandas*, consultas SQL o visualizaciones tabulares. En cambio, los datos no estructurados, como texto libre o imágenes, requieren técnicas más avanzadas de procesamiento, tales como minería de texto o redes neuronales convolucionales.

A través de esta comparación se comprendió que el tipo de estructura de los datos define directamente las estrategias de limpieza, transformación y análisis que deben emplearse. Los resultados mostraron que los datasets estructurados permiten un análisis más rápido y directo, mientras que los no estructurados ofrecen información más rica, aunque demandan mayor poder computacional y técnicas especializadas.

2.2. Ejercicio 2: Pipeline de Ingestión con Pandas y Dask

Se utilizó el dataset público *AirQualityUCI* desde el repositorio UCI. El objetivo fue construir un flujo de procesamiento (pipeline) para la lectura y manipulación de datos utilizando las librerías **Pandas** y **Dask**.

Primero, se cargaron los datos con Pandas, observando tiempos de lectura aceptables pero un consumo de memoria alto al trabajar con más de 9000 registros. Posteriormente, se repitió el procedimiento con Dask, observándose una clara mejora en rendimiento gracias al procesamiento en paralelo y la posibilidad de manejar el dataset por fragmentos.

Se aplicaron tareas de limpieza como eliminación de valores nulos, conversión de tipos de datos y generación de nuevas columnas a partir de variables meteorológicas. Los resultados se midieron en tiempo de ejecución y uso de CPU, concluyéndose que Dask es más eficiente en entornos con grandes volúmenes de información o cuando se requiere escalabilidad.

2.3. Ejercicio 3: Comparación de Pandas y PySpark

Se contrastaron ambas librerías en tareas de filtrado, agrupamiento y completado de datos faltantes. Pandas resultó ideal para análisis exploratorios y entornos locales, destacando su sencillez y versatilidad. Por otro lado, PySpark mostró un desempeño superior cuando se simularon operaciones sobre grandes volúmenes de datos, gracias a su capacidad de procesamiento distribuido basado en Apache Spark.

Se observó que, aunque PySpark requiere mayor configuración inicial, su rendimiento en entornos de clúster lo convierte en una herramienta clave para proyectos que requieren análisis masivos. La comparación permitió comprender que la elección de la librería depende del contexto: Pandas para exploración rápida, PySpark para proyectos escalables en producción.

2.4. Ejercicio 4: Comparación de Librerías de Visualización

Se realizaron visualizaciones con **Matplotlib**, **Seaborn** y **Plotly**. El objetivo fue comprender las diferencias entre cada librería en términos de funcionalidad, personalización e interactividad.

Se generaron gráficos de dispersión, histogramas y diagramas de correlación. Matplotlib demostró ser la más flexible, permitiendo un control detallado de cada elemento gráfico, aunque requiere mayor cantidad de código. Seaborn simplificó la creación de gráficos estadísticos, integrando fácilmente análisis de distribución y regresión. Plotly destacó por su carácter interactivo, posibilitando crear dashboards dinámicos útiles para informes y presentaciones.

Los resultados confirmaron que la selección de la herramienta debe basarse en el propósito: Matplotlib para precisión técnica, Seaborn para análisis rápidos y Plotly para visualizaciones interactivas y comunicativas.

2.5. Ejercicio 5: Implementación de un Perceptrón desde Cero

Se implementó un perceptrón binario con **NumPy** usando el dataset *Iris*. El modelo se entrenó ajustando pesos y sesgos mediante la regla de aprendizaje del perceptrón. Se graficó la frontera de decisión y la evolución del error durante las épocas, observándose una rápida convergencia. El modelo alcanzó una precisión superior al 95 %, confirmando su capacidad para clasificar datos linealmente separables.

Este ejercicio permitió comprender los fundamentos de las redes neuronales y la importancia de parámetros como la tasa de aprendizaje, además de servir como puente

conceptual hacia modelos más complejos de aprendizaje profundo.

2.6. Conclusiones Generales

Los ejercicios permitieron aplicar herramientas esenciales de ciencia de datos, desde el manejo eficiente de datasets hasta la implementación de algoritmos de aprendizaje supervisado. El uso de diferentes librerías evidenció la importancia de elegir herramientas adecuadas según el tamaño del dataset, la disponibilidad de recursos y el objetivo analítico. En conjunto, los cinco ejercicios constituyen un recorrido completo por el flujo de trabajo en ciencia de datos: desde la comprensión de los tipos de datos, hasta la implementación de modelos predictivos básicos. La experiencia reforzó la importancia de la reproducibilidad, la eficiencia computacional y la interpretación clara de resultados, habilidades clave en el perfil de un científico de datos moderno.

3. Anteproyecto del Proyecto Integrador

Predicción de la calidad del aire en Santiago utilizando ciencia de datos.

3.1. Resumen

El proyecto busca desarrollar un modelo predictivo de contaminación atmosférica (niveles de CO, NO_2 y PM10) en la ciudad de Santiago, empleando técnicas de procesamiento y modelado de datos. Se pretende construir un pipeline que permita recolectar, limpiar y analizar datos provenientes de fuentes públicas, para generar visualizaciones y predicciones que apoyen la toma de decisiones ambientales.

3.2. Justificación

La contaminación del aire constituye uno de los principales problemas ambientales en Chile. Predecir sus niveles mediante modelos de ciencia de datos permitirá anticipar episodios críticos y apoyar políticas públicas. El uso de técnicas aprendidas en el curso (Pandas, Dask, PySpark, visualización y aprendizaje automático) garantiza la viabilidad técnica del proyecto.

3.3. Objetivos

Objetivo General:

- Desarrollar un modelo predictivo que estime la calidad del aire en Santiago a partir de datos históricos y meteorológicos.

Objetivos Específicos:

1. Recolectar y unificar datos históricos de contaminación desde fuentes públicas (SINCA, Kaggle).

2. Aplicar técnicas de limpieza y análisis exploratorio para preparar los datos.
3. Implementar modelos de predicción utilizando algoritmos de regresión y clasificación.
4. Evaluar el desempeño de los modelos mediante métricas estándar (RMSE, R^2).
5. Generar visualizaciones interactivas que permitan interpretar los resultados.

3.4. Metodología Propuesta

- **Fuentes de datos:** Repositorios abiertos (UCI, Kaggle, SINCA).
- **Herramientas:** Python, Pandas, Dask, PySpark, Scikit-learn, Plotly.
- **Etapas:**
  1. Extracción y limpieza de datos.
  2. Análisis exploratorio y visualización.
  3. Entrenamiento y validación de modelos.
  4. Interpretación y documentación de resultados.

3.5. Estado del Arte

Se han desarrollado modelos de predicción de calidad del aire en distintas ciudades utilizando regresión lineal, bosques aleatorios y redes neuronales. Entre los trabajos más recientes destacan:

- Sharma et al. (2023): "Air Quality Forecasting using Machine Learning Techniques", *Environmental Informatics Journal*.
- Li y Chen (2022): "Deep Learning for PM2.5 Prediction in Urban Areas", *IEEE Access*.
- Castro et al. (2021): "Aplicación de modelos predictivos para estimar contaminantes atmosféricos en Chile", *Revista Ingeniería Ambiental*.

Este proyecto se diferencia por centrarse en Santiago y combinar librerías distribuidas y visualización interactiva.

3.6. Cronograma Tentativo

Semana	Actividad Principal
1–2	Recolección y exploración de datos
3–4	Limpieza y normalización
5–6	Modelado predictivo inicial
7–8	Evaluación y ajuste de modelos
9	Visualización y documentación final
10	Presentación de resultados y entrega final

3.7. Conclusión General del Informe

El desarrollo de esta prueba me permitió comprender cómo las herramientas de ciencia de datos se integran de forma práctica en un flujo de trabajo real. Desde la ingesta de datos distribuidos hasta la visualización y modelado, cada ejercicio reforzó habilidades distintas pero complementarias. El proyecto integrador propuesto surge naturalmente de esta experiencia, aplicando los mismos principios para un problema ambiental relevante en Chile. Considero que este tipo de evaluaciones, que combinan práctica técnica y reflexión aplicada, fortalecen mi capacidad para abordar proyectos de ciencia de datos de forma autónoma, estructurada y responsable.

3.8. Referencias

- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
- Kaggle – Air Quality Dataset: https://www.kaggle.com/datasets
- SINCA (Sistema de Información Nacional de Calidad del Aire): https://sinca.mma.gob.cl/