

# Documento de planificación inicial

## 1.1 Propósito del Proyecto

Basado en el análisis del Proyecto Integrador33, este proyecto tiene como objetivo principal desarrollar un sistema integral de análisis de datos que combine técnicas de procesamiento, modelado predictivo y visualización para resolver problemas complejos de ciencia de datos.

## 1.2 Alcance del Proyecto

- **Incluye:** Procesamiento de datos, análisis exploratorio, modelado predictivo, evaluación de modelos y visualización de resultados
- **Excluye:** Implementación en producción, mantenimiento continuo, integración con sistemas externos

## 2. Objetivos del Proyecto

### 2.1 Objetivos Principales

#### 1. Análisis Exploratorio de Datos (EDA)

- Comprensión profunda de la estructura de datos
- Identificación de patrones y relaciones
- Detección de valores atípicos y anomalías

#### 2. Desarrollo de Modelos Predictivos

- Implementación de múltiples algoritmos de machine learning
- Optimización de hiperparámetros
- Validación cruzada de modelos

#### 3. Evaluación de Resultados

- Comparación sistemática de modelos
- Análisis de métricas de rendimiento
- Interpretación de resultados

### **3. Metodología de Trabajo**

#### **3.1 Fases del Proyecto**

##### **Fase 1: Preparación de Datos (Semanas 1-2)**

- Carga y exploración inicial de datos
- Limpieza y preprocesamiento
- Ingeniería de características

##### **Fase 2: Análisis Exploratorio (Semanas 3-4)**

- Análisis univariado y multivariado
- Visualizaciones exploratorias
- Identificación de correlaciones

##### **Fase 3: Modelado (Semanas 5-7)**

- Selección de algoritmos
- Entrenamiento de modelos
- Optimización de parámetros

##### **Fase 4: Evaluación (Semanas 8-9)**

- Validación de modelos
- Análisis comparativo
- Interpretación de resultados

##### **Fase 5: Documentación (Semana 10)**

- Documentación técnica
- Preparación de reportes
- Presentación de hallazgos

## **4. Recursos Requeridos**

### **4.1 Recursos Técnicos**

- **Lenguaje de Programación:** Python 3.8+
- **Entorno:** Google Colap
- **Librerías Principales:**
  - pandas, numpy (procesamiento de datos)
  - scikit-learn (machine learning)
  - matplotlib, seaborn (visualización)
  - scipy (análisis estadístico)

### **4.2 Recursos Humanos**

- Científico de Datos Principal
- Analista de Datos (opcional)
- Revisor de Métodos Estadísticos

## **5. Entregables Principales**

### **5.1 Documentación Técnica**

- Notebook completo con código ejecutable
- Documentación de métodos y algoritmos
- Reporte de resultados y hallazgos

### **5.2 Resultados del Análisis**

- Modelos entrenados y evaluados
- Métricas de rendimiento comparativas
- Visualizaciones clave de insights

### **5.3 Recomendaciones**

- Guías de implementación
- Sugerencias para mejoras futuras
- Limitaciones identificadas

## 6. Criterios de Éxito

### 6.1 Métricas de Calidad

- **Precisión de Modelos:** > 85% de accuracy (dependiendo del problema)
- **Reproducibilidad:** Código 100% ejecutable
- **Documentación:** Completa y clara
- **Tiempo de Ejecución:** Optimizado para eficiencia

### 6.2 Criterios de Aceptación

- Todos los análisis ejecutados sin errores
- Resultados consistentes y validados
- Documentación completa y comprensible

## 7. Gestión de Riesgos

### 7.1 Riesgos Identificados

#### 1. Calidad de Datos

- **Impacto:** Alto
- **Mitigación:** Validación temprana, limpieza robusta

#### 2. Rendimiento de Modelos

- **Impacto:** Medio
- **Mitigación:** Múltiples algoritmos, tuning exhaustivo

#### 3. Recursos Computacionales

- **Impacto:** Bajo
- **Mitigación:** Optimización de código, uso de muestras

### 7.2 Plan de Contingencia

- Backup diario del trabajo
- Versiones alternativas de modelos
- Documentación progresiva

## 8. Cronograma Tentativo

Fase	Duración	Hitos Principales
Preparación	2 semanas	Datos limpios y listos
Análisis	2 semanas	Insights iniciales
Modelado	3 semanas	Modelos entrenados
Evaluación	2 semanas	Resultados validados
Documentación	1 semana	Entrega final

## 9. Consideraciones Éticas y de Privacidad

### 9.1 Aspectos Éticos

- Uso responsable de datos
- Transparencia en metodologías
- Validación estadística apropiada

### 9.2 Privacidad de Datos

- Anonimización cuando sea necesario
- Cumplimiento de regulaciones aplicables
- Almacenamiento seguro de información

## 10. Próximos Pasos

1. **Inmediatos** (Semana 1):
  - Configuración del entorno
  - Revisión inicial de datos
  - Establecimiento de métricas base
2. **Corto Plazo** (Semanas 2-4):
  - Análisis exploratorio completo
  - Definición de estrategia de modelado
3. **Mediano Plazo** (Semanas 5-8):
  - Desarrollo y optimización de modelos
- Validación de resultados