# CAP6768 Data Analytics - Final Project Guidelines

**Fall 2025**

# Overview

The final project (30% of grade) requires applying data analytics techniques from the course to solve a clearly defined business problem. Groups of 4-5 students (or less) will prepare a 5-8 page report demonstrating their analysis.

**How to find a group:**

1) Use the **"People"** tab on Canvas to directly contact the colleague(s).
   **OR**
2) Use the **"Finding a Group for the Project"** discussion on Canvas. There, you can share your ideas or a dataset you plan to use for the final project. You can also let other students know that you are looking for a group, or express your interest in joining someone who already has a dataset.

# Dataset Options

## Option 1: Professional Dataset

- Use data from a team member's workplace or professional context
- Must be real business data (properly anonymized)
- Requires employer permission if applicable
- **NOT ALLOWED**: Public repository datasets (Kaggle, UCI, GitHub, etc.)

## Option 2: Instructor-Provided Datasets

- Two business datasets will be posted on Canvas
- Each includes business context and suggested problems
- There is no restriction on multiple teams choosing the same dataset

# Project Requirements

## 1. Problem Definition (MANDATORY)

Before starting any analysis, you must clearly define:

- **The specific business problem** you are solving

- **The type of analytics problem**:
    - Classification (predicting categories)
    - Regression (predicting values)
    - Clustering (finding groups)
    - Time series forecasting
    - Other (specify)
- **The proposed solution approach** using course techniques
- **Expected business value** from your solution

### 2. Required Techniques

Although you should use some of the techniques covered in the course, you don't need to restrict your analysis just to the course content. You are more than welcome to expand it and use different techniques.

- Descriptive Data Mining
- Linear/Logistic Regression
- Classification algorithms
- Clustering methods
- Time series analysis
- Statistical inference
- Predictive data mining
- Dimensionality reduction

### 3. Analysis Components

- Data preparation
- Data summarization, visualization and descriptive data mining (if applicable)
- Method implementation
- Model evaluation
- Business insights and recommendations with more data visualization

# Deliverables

## Project Proposal (Due: Week 8 – October 12th)

**1-2 pages containing:**

- Selected dataset option
- **Clear problem statement**
- **Problem type** (classification/regression/clustering/etc.)
- **Proposed solution methods** (which techniques and why)
- Expected outcomes
- Team member names

### Final Report (Due: Week 16 - December 7th)

**5-8 pages including:**

- Executive summary
- Problem definition and business context
- Data description
- Methodology and analysis
- Results and evaluation
- Business recommendations
- References

### Supporting Materials

- Documented code (R, Python, or other)
- Clean data files (if using Option 1)

# Grading Focus

- Clear problem definition (20%)
- Appropriate technique selection (20%)
- Correct implementation (20%)
- Result interpretation (20%)
- Business insights (20%)

# Important Rules

❌ **NO public datasets** from Kaggle, UCI, GitHub, data.gov, or similar repositories

❌ **NO toy/academic datasets** (iris, titanic, boston housing, etc.)

✅ **ONLY** real business data or instructor-provided datasets

✅ **MUST** have clear business problem and proposed solution

# Timeline

- **Weeks 4-8: Form teams and work on proposal**
- **Week 8 (Oct 12th): Submit proposal with problem definition**
- **Week 16 (Dec 7th): Submit final report**

# Questions?

- Canvas discussion forum

- Office hours: Tu/Th 10am-12pm, Wed 12pm-2:30pm
- Email: allan.quadros@unf.edu

---

# APPENDIX A: Sample Project Proposal

## Customer Churn Prediction for TelecomCo

**Team Members:** Sarah Chen, Mike Rodriguez, Anna Patel, James Wilson, Lisa Thompson

### Dataset Selection

We will use data from TelecomCo (Sarah's employer), comprising 18 months of customer data including demographics, service usage, billing, and churn status for 50,000 customers.

### Problem Statement

TelecomCo experiences 23% annual customer churn, costing approximately $45M in lost revenue. We need to identify customers likely to churn within the next 90 days to enable targeted retention campaigns.

### Problem Type

**Classification Problem** - Binary classification predicting whether a customer will churn (Yes/No) within 90 days.

### Proposed Solution Methods

1. **Logistic Regression** - Baseline model for interpretability and identifying key churn drivers
2. **Random Forest** - Capture non-linear relationships and feature interactions
3. **Time Series Analysis** - Analyze usage patterns over time to identify churn signals
4. **K-means Clustering** - Segment customers to understand different churn profiles

### Expected Outcomes

- Prediction model with >80% accuracy and >75% recall for churn cases

- Identification of top 5 churn predictors
- Customer segments with specific retention strategies
- Cost-benefit analysis of retention campaigns

## Data Availability

Data is anonymized and approved by TelecomCo's legal department. Includes 45 features across customer demographics, account information, service usage, and payment history.

---