Project 5

Advanced Analytics

Introduction to Data Science with Python

Computer Science Program

Kutaisi International University

Project Information

Total Weight:	6% of Final Grade
Task Breakdown:	3 Tasks × 2% each
Due Date:	End of Week 13 (Sunday, 23:59 GT)
Format:	Individual Assignment
Deliverable:	Jupyter Notebook (.ipynb) or Python Script (.py)
Approach:	Classification OR Clustering (your choice!)

Contents

1 Project Overview

1.1 Introduction

Welcome to the final project! You've mastered data manipulation, visualization, and regression. Now it's time to tackle advanced machine learning: classification and clustering. This project offers flexibility—you choose whether to solve a classification problem (predicting customer churn) or a clustering problem (segmenting customers), or attempt both for bonus points!

You'll work with e-commerce customer data containing behavioral, demographic, and satisfaction metrics. Your goal is to build a working model, interpret its results, and provide actionable business recommendations that could save the company money or improve customer targeting.

Project Approach Options

Choose Your Adventure:
Option A: Classification (Churn Prediction)

- Predict which customers will churn (leave)
- Use supervised learning algorithms
- Evaluate with precision, recall, F1-score
- Identify at-risk customers for retention campaigns

Option B: Clustering (Customer Segmentation)

- Discover natural customer segments
- Use unsupervised learning algorithms
- Evaluate with silhouette score, elbow method
- Create personas for targeted marketing

Option C: Both (+Bonus points up to +5%)

- Complete both approaches
- Analyze churn patterns within segments
- Provide comprehensive strategy

Important: Data Files

You will be provided with a Python script (`generate_project5_data.py`) that generates comprehensive customer data with 1000 customers and 30+ features including demographics, purchase behavior, engagement metrics, and satisfaction scores.
Download the data generator script from:

- LMS course materials folder
- Or instructor will provide via email/announcement

Run the script: `python generate_project5_data.py`
This will create: `customer_data.csv`

1.2 Learning Objectives

Upon completion of this project, you will be able to:

- **Classification:** Build and evaluate binary classifiers for business problems
- **Clustering:** Discover and interpret customer segments using unsupervised learning
- **Model Evaluation:** Use appropriate metrics for different problem types
- **Feature Importance:** Identify key drivers of customer behavior
- **Business Translation:** Convert technical findings into actionable recommendations
- **Advanced Algorithms:** Implement tree-based methods, SVM, K-Means, Hierarchical clustering
- **Class Imbalance:** Handle imbalanced datasets (for classification)
- **Cluster Validation:** Determine optimal number of clusters

1.3 Project Structure

Task	Focus Area	Weight
Task 1	Data Preparation & EDA	2%
Task 2	Model Implementation	2%
Task 3	Business Insights & Recommendations	2%
Total		**6%**

Critical Deadline Information

Academic Integrity Policy:

- All work must be your own individual effort
- You may consult official documentation (Scikit-learn, Pandas) and course materials
- Copying code from online sources, AI tools, or other students will result in zero points
- If you reference any external resources for concepts, cite them in comments or markdown cells

2 Task Specifications

2.1 Task 1: Data Preparation & Exploratory Analysis (2%)

Objective

Prepare the data and conduct exploratory analysis to understand customer patterns and relationships between features.

Requirements

Part A: Data Loading & Initial Exploration (25%)

1. **Load and examine the dataset:**
   - Load `customer_data.csv`
   - Display basic info: shape, columns, data types
   - Show statistical summary
   - Identify missing values
   - Check class distribution (if doing classification)
2. **Target variable analysis:**
   - **For Classification:** Display churn distribution, check for class imbalance
   - **For Clustering:** Note that you'll remove the 'Churned' column

Part B: Exploratory Data Analysis (40%)
Create visualizations and analysis to understand the data:

1. **Distribution Analysis:**
   - Visualize distributions of key numerical features
   - Identify potential outliers
   - Check for skewed distributions
2. **Relationship Analysis:**
   - Create correlation heatmap for numerical features
   - **For Classification:** Compare feature distributions between churned and active customers
   - **For Clustering:** Look for natural groupings in scatter plots
   - Identify top 5 features most correlated with churn (classification) or showing variation (clustering)
3. **Categorical Analysis:**
   - Analyze membership type distribution
   - Compare behavior across customer segments
   - Visualize engagement metrics by category
4. **Key Insights:**

- Write 3-5 bullet points summarizing patterns discovered
- Hypothesize what might drive churn (classification) or segment differences (clustering)

Part C: Data Preprocessing (35%)
Prepare data for modeling:

1. **Handle Missing Values:**
   - Fill missing numerical values with median or mean
   - Document your strategy
2. **Feature Selection:**
   - Remove ID column (`Customer_ID`)
   - Remove `Churn_Risk_Score` (data leakage - calculated from target)
   - **For Clustering:** Also remove `Churned` column
   - Keep all other relevant features initially
3. **Encode Categorical Variables:**
   - One-hot encode: Gender, Education, Location_Type, Membership_Type, Payment_Method, Favorite_Category
   - Use `pd.get_dummies()` or sklearn encoders
4. **Feature Scaling:**
   - Apply StandardScaler to all numerical features
   - **For Classification:** Fit on train set only
   - **For Clustering:** Fit on entire dataset
5. **Train-Test Split (Classification only):**
   - Split 80/20 with stratification on target
   - Set random_state=42
   - Use `stratify=y` to maintain class balance

Evaluation Criteria

- Comprehensive data exploration (35%)
- Quality of visualizations and insights (30%)
- Proper preprocessing (25%)
- Code documentation (10%)

2.2 Task 2: Model Implementation (2%)

Objective

Implement and evaluate machine learning models appropriate for your chosen approach.
Complete EITHER Section A (Classification) OR Section B (Clustering)

SECTION A: Classification Approach

Part A1: Baseline Classification Models (35%)
Implement at least 3 classification algorithms:

1. **Logistic Regression:**
   - Train baseline logistic regression
   - Evaluate on test set
   - Print confusion matrix and classification report
2. **Decision Tree Classifier:**
   - Train with max_depth=5 initially
   - Visualize the tree (optional but recommended)
   - Extract feature importance
3. **Random Forest Classifier:**
   - Train with n_estimators=100
   - Extract feature importance
   - Compare with single decision tree
4. **Additional Model (choose one):**
   - Support Vector Machine (SVM)
   - Gradient Boosting
   - K-Nearest Neighbors (KNN)

Part A2: Handle Class Imbalance (25%)
If classes are imbalanced (>60/40 split):

1. **Technique 1 - Class Weights:**
   - Use `class_weight='balanced'` in models
   - Compare results with unweighted models
2. **Technique 2 - Resampling (Optional):**
   - Try oversampling minority class OR
   - Undersampling majority class
   - Compare with baseline

Part A3: Model Evaluation & Comparison (40%)

1. **Calculate Metrics:**
    - Accuracy, Precision, Recall, F1-Score for each model
    - ROC-AUC score
    - Confusion matrix
2. **Model Comparison Table:**
    - Create DataFrame comparing all models
    - Include all metrics
    - Identify best model (justify your choice)
3. **ROC Curve:**
    - Plot ROC curves for all models on same plot
    - Include AUC scores in legend
    - Interpret which model has best discrimination
4. **Feature Importance:**
    - Extract from best tree-based model
    - Visualize top 10 most important features
    - Interpret: What drives churn?

SECTION B: Clustering Approach

Part B1: K-Means Clustering (35%)

1. **Elbow Method:**
   - Test k from 2 to 10
   - Plot inertia (within-cluster sum of squares)
   - Identify the ”elbow” point
2. **Silhouette Analysis:**
   - Calculate silhouette score for k from 2 to 10
   - Plot scores
   - Identify k with highest silhouette score
3. **Choose Optimal k:**
   - Based on elbow method and silhouette scores
   - Justify your choice (typically k=3 to 5)
   - Train K-Means with optimal k

Part B2: Alternative Clustering Methods (25%)

Implement at least one additional clustering algorithm:

1. **Hierarchical Clustering:**
   - Create dendrogram
   - Use same k as K-Means
   - Compare cluster assignments

OR

2. **DBSCAN:**
   - Experiment with eps and min_samples
   - Identify core points and outliers
   - Count number of clusters found

Part B3: Cluster Analysis & Interpretation (40%)

1. **Cluster Profiles:**
   - Calculate mean values for each feature per cluster
   - Create DataFrame showing cluster characteristics
   - Identify what makes each cluster unique
2. **Visualization:**
   - Use PCA to reduce to 2D for visualization
   - Create scatter plot of clusters in 2D space
   - Color points by cluster

- Include cluster centers

3. **Cluster Naming:**
   - Give each cluster a descriptive name (e.g., "High-Value Loyalists", "Price-Sensitive Shoppers")
   - Write 2-3 sentence description of each segment
   - Include size of each cluster

4. **Churn Analysis within Clusters (Optional Bonus):**
   - Map original 'Churned' labels to clusters
   - Calculate churn rate per cluster
   - Identify which segments have highest churn risk

Evaluation Criteria

- Correct implementation of algorithms (40%)
- Proper evaluation methodology (30%)
- Quality of interpretation (20%)
- Code organization and documentation (10%)

Key Scikit-learn Imports

Classification:

```
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score,
    recall_score, f1_score, roc_auc_score, roc_curve, confusion_matrix
    , classification_report
```

Clustering:

```
from sklearn.cluster import KMeans, AgglomerativeClustering, DBSCAN
from sklearn.metrics import silhouette_score, silhouette_samples
from sklearn.decomposition import PCA
from scipy.cluster.hierarchy import dendrogram, linkage
```

2.3 Task 3: Business Insights & Recommendations (2%)

Objective

Translate technical findings into actionable business recommendations that can drive real-world decisions.

Requirements

Part A: Executive Summary (25%)
Write a non-technical summary (200-250 words) covering:

1. **Problem Statement:**
   - What business problem did you solve?
   - Why is this important to the company?
2. **Approach:**
   - What method did you use? (classification or clustering)
   - What data did you analyze?
3. **Key Findings:**
   - 3-5 most important discoveries
   - Quantify where possible (e.g., "30% of customers...")
4. **Bottom Line:**
   - What's the main takeaway for executives?
   - What action should the company take?

Part B: Detailed Analysis (35%)
Provide in-depth analysis based on your approach:
For Classification:

1. **Churn Drivers:**
   - Which factors most strongly predict churn?
   - Are there surprising findings?
   - What's the typical profile of a churning customer?
2. **Model Performance:**
   - How accurate is your best model?
   - What's the precision/recall tradeoff?
   - Can the company trust these predictions?
3. **At-Risk Customers:**
   - How many customers are predicted to churn?
   - What's the potential revenue at risk?
   - Which customer characteristics indicate highest risk?

For Clustering:

1. **Segment Profiles:**
   - Describe each customer segment in detail
   - What's the size and value of each segment?
   - What are the distinct behaviors of each group?
2. **Segment Opportunities:**
   - Which segment has highest lifetime value?
   - Which segment is most engaged?
   - Are there underserved segments?
3. **Segment Risks:**
   - Do any segments show warning signs?
   - Which segments have high churn rates?
   - Where should the company focus retention efforts?

Part C: Actionable Recommendations (40%)
Provide specific, actionable recommendations:

1. **Immediate Actions (next 30 days):**
   - 3-4 specific actions the company should take now
   - Prioritize by impact and feasibility
   - Example: "Launch retention campaign targeting customers with >100 days since last purchase"
2. **Strategic Initiatives (next 6 months):**
   - 2-3 longer-term strategies
   - Link to segment characteristics or churn factors
   - Example: "Develop loyalty program enhancements for Premium members"
3. **Personalized Approaches:**
   - **Classification:** Different retention tactics for different churn risk levels
   - **Clustering:** Segment-specific marketing and engagement strategies
   - Include specific examples for each group
4. **ROI Estimation:**
   - Estimate potential financial impact
   - Example: "Retaining 20% of at-risk customers = $X in saved revenue"
   - Show your calculations
5. **Implementation Plan:**
   - Who should own these initiatives?
   - What resources are needed?
   - How will success be measured?

Evaluation Criteria

- Clarity of executive summary (20%)
- Depth of analysis (30%)
- Quality and specificity of recommendations (35%)
- ROI/business impact consideration (10%)
- Professional presentation (5%)

Excellence Indicators: To achieve >95%, demonstrate:

- Deep understanding of business implications
- Specific, actionable recommendations (not generic)
- Quantified impact estimates
- Creative solutions beyond obvious approaches
- Clear communication suitable for non-technical audience
- Evidence-based argumentation

3 Submission Guidelines

3.1 Deliverable Format

Submit **ONE** file in your preferred format:

- Jupyter Notebook: `Project5_FirstName_LastName.ipynb`, OR
- Python Script: `Project5_FirstName_LastName.py`

Required File Structure

For Jupyter Notebook:

1. **Title Cell:**
   - Project title and number
   - Your full name and Student ID
   - Approach chosen: "Classification" OR "Clustering" OR "Both"
   - Submission date
   - Honor code: "I certify this work is my own"
2. **For Each Task:**
   - Clear headers
   - Code with outputs
   - Interpretations and insights
3. **Business Report Section:**
   - Use markdown for Task 3
   - Professional formatting
   - Include visualizations where helpful

3.2 Submission Methods

Choose **ONE** submission method:

Option 1: Direct LMS Upload

1. Ensure all code runs without errors
2. Create ZIP: `Project5_FirstName_LastName.zip`
3. Include notebook/script + CSV file
4. Upload to LMS before deadline

Option 2: GitHub Repository

1. Create repo: `KIU-DS-Project5-FirstName-LastName`
2. Include all files + README + requirements.txt
3. Submit URL via LMS before deadline
4. No commits after deadline!

Critical Deadline Information

Deadline: End of Week 13 - Sunday, 23:59 Georgian Time

- Late submissions not accepted
- Commits after deadline = 0 points
- Submit 2+ hours early to avoid issues

3.3 Pre-Submission Checklist

- ☐ All three tasks completed
- ☐ Approach clearly stated (Classification/Clustering/Both)
- ☐ Code runs without errors
- ☐ Models trained and evaluated
- ☐ Visualizations included
- ☐ Business recommendations written
- ☐ CSV file included
- ☐ Random states set (reproducibility)
- ☐ Professional presentation
- ☐ Submitting before deadline

4 Grading Rubric

4.1 Grade Distribution

Component	Points	% of Final Grade
Task 1: Data Preparation & EDA	2.0	2%
Task 2: Model Implementation	2.0	2%
Task 3: Business Insights	2.0	2%
Project Total	**6.0**	**6%**

4.2 Bonus Opportunities (Up to +5%)

- **+5%:** Complete BOTH classification and clustering approaches
- **+3%:** Advanced techniques (ensemble methods, hyperparameter tuning)
- **+2%:** Exceptional business insights with quantified ROI
- **+2%:** Outstanding visualizations and presentation

Note: Maximum score capped at 6%

5 Resources & Support

5.1 Official Documentation

- Scikit-learn Classification: `https://scikit-learn.org/stable/supervised_learning.html`
- Scikit-learn Clustering: `https://scikit-learn.org/stable/modules/clustering.html`
- Metrics: `https://scikit-learn.org/stable/modules/model_evaluation.html`

5.2 Getting Help

1. Review course materials (Weeks 10-13)
2. Check Scikit-learn documentation
3. Email: Nika Gagua at `Nika.Gagua@kiu.edu.ge`

5.3 Common Pitfalls

- Forgetting to remove Churn_Risk_Score (data leakage)
- Not scaling features for clustering
- Focusing only on accuracy for imbalanced classification
- Not interpreting model results
- Generic recommendations without specifics
- Starting too late (this is a complex project!)

Congratulations on reaching the final project!

This is your opportunity to showcase everything you've learned.
Focus on both technical excellence and business value.
Make your recommendations actionable and impactful.

We're excited to see your insights!

- Your Data Science Instructors