BIFX-546: Machine Learning for Bioinformatics

Instructor: Dr. Sarangan (Ravi) Ravichandran, Ph.D.

Course Project: Deliverables & Grading Requirements

These project milestones constitute 100% of the course grade. Percentages shown below are percent of the course grade. Class numbers correspond to week numbers in the course schedule.

1) Self-Introduction & Learning Goals (5%, Due: Class Meeting 1)

What to Submit

- A short paragraph (150–200 words) posted to Blackboard:
  - Who you are (background, interests)
  - What do you hope to learn from this course
  - What skills do you want to gain or improve
  - Whether you prefer **solo** or **team of 2** for the project

2) Project Proposal (10%, Due: Class Meeting 3)

What to Include

A **1-page PDF** containing:

1. **Dataset**
   - Name
   - Source/URL
   - Why you chose it
2. **Problem statement**
   - What question are you trying to answer?
   - What outcome or insight do you expect?
3. **Methods you plan to use**
   - Choose from Weeks 1–7 concepts (e.g., visualization, summary stats, probability, correlation, optimization)
4. **Team members** (if applicable)

3) Progress Check-In #1 (10%, Due: A day before Class Meeting 8)

Submit preliminary EDA notebook:

- Dataset loaded

- Initial cleaning

- 2–3 visualizations

- At least one summary statistic

- Short progress note

4) Project Midterm Presentation (20%, Due: Class Meeting 8)

This is the Midterm milestone focused on EDA (20%): code demo + logic + key exploratory findings.

5–7 minute presentation including:

- Title + team members

- Dataset overview

- EDA (3 plots + summary)

- Early insights

- Next steps

5) Progress Check-In #2 (10%, Due: Class Meeting 13)

Submit 1–2 paragraphs + intermediate notebook:

- Updated analysis

- New visualizations or methods

6) Final Demo & Discussion (30%, Due: Last Class)

This is the Final milestone focused on Modeling (30%): model(s), validation, interpretation, limitations, and next steps.

10-minute final presentation:

- Project goal & methods

- Final visualizations and results

- Discussion, limitations, future work

- Live repo walkthrough

7) GitHub Repository Quality (15%, Due: 3 days before final demo)

Repo structure:

project-name/

notebooks/

data/

results/

src/

README.md

requirements.txt

README must include:

• Title, team, dataset, methods

• How to run the code

• Summary of findings

• List of plots

Project Deliverables

1. Github Repository (Required)

Each team must maintain a **GitHub repository** with the following structure:

```
project-name/
├── notebooks/        # All Jupyter notebooks
├── data/             # (optional) Small sample or link to data
├── results/          # Plots, tables, figures
├── src/              # Optional Python scripts
├── README.md         # Required (see below)
└── requirements.txt  # Minimal package list
```

2. README.md Requirements

Your README should include:

- 📄 **Title** and **Team Members**
- 🎯 **Project Goal** (what question are you answering?)
- 🧠 **Techniques Used** (from class: visualization, statistics, probability, etc.)
- 📊 **Dataset Source** (include citation or URL)
- ⚙️ **How to Run Your Code** (basic instructions)
- 📝 **Summary of Findings** (1–2 paragraphs)

✅ 3. Minimum Expectations

Each project must:

1. Use at least one real dataset (publicly available or self-collected).
2. Apply at least two techniques from class (e.g., visualization + hypothesis test). One technique should be for analyzing the data and the other for testing (e.g. regression or clustering etc.)
3. Produce at least three plots or visual summaries.
4. Include basic statistical or computational analysis (mean, correlation, simulation, regression, etc.).
5. Have well-documented, reproducible code that runs without modification on Colab or Jupyter.
6. Present a clear question, method, and conclusion (not just data exploration).

💡 4. Example Project Themes (Optional Ideas)

Domain	Example Question	Possible Data Source
Personal Health & Wearables	How do daily step counts relate to sleep duration or quality?	Fitbit-style datasets (Kaggle), Apple Health exports
Clinical Measurements	Are BMI and blood pressure correlated across age groups?	Public clinical datasets, CDC health indicators
Public Health & Policy	Do vaccination rates vary by income or education level?	CDC, data.gov
Environmental Health	Is air quality associated with asthma-related hospital visits?	EPA Air Quality Data, CDC
Healthcare Operations	What does the distribution of emergency room wait times look like?	Public hospital operations datasets, CMS synthetic data
Epidemiology	How variable are daily case counts across weeks or regions?	Our World in Data, CDC
Health Disparities	Are chronic disease rates different across demographic groups?	CDC BRFSS, public census-linked health data
Unstructured → Structured Health Data	What symptoms are most frequently mentioned in patient reviews?	Patient review datasets (Kaggle), public health reports
Clinical Text Analysis	What medical terms appear most often in de-identified clinical notes?	MIMIC-style demo datasets, synthetic clinical text
Medication & Treatment Trends	Are certain side effects reported more frequently for specific drugs?	FDA Adverse Event Reporting System (FAERS)

5. CommonDataset Sources

Choose public, clean datasets:

- Kaggle Datasets
- data.gov
- Our World in Data
- UCI Machine Learning Repository
- CDC Data Portal
- https://huggingface.co/

Dataset size: ideally **< 50 MB** and **< 200 k rows** — manageable in Colab or laptop Jupyter.

Evaluation (100 pts total)

Assignments	Grade %	Criteria	Focus
Self-Intro & Goals	5	Self-introduction and goals and expectations for the class	
Problem Definition & Project Proposal	10	Clear problem, dataset justification, feasible scope, plan	
Check-in #1 (Pre midterm): Data Handling, EDA, Analysis; Visualization & Communication	10	Proper data loading, cleaning, transformation, correct application of techniques; Clear, well-labeled plots, visual storytelling, interpretability; Use of appropriate methods from Weeks 1–7	
Midterm Presentation	20	Progress clarity, initial results, feedback incorporation; Reproducible notebook, good use of class concepts (WHAT ELSE I CAN ADD?)	Foundations:
Check-in #2 (Pre Final): Final Demo, Discussion & Analysis; Visualization & Communication	10	Use of appropriate methods from the class (e.g., summary stats, correlations, probability, modeling methods such as simple regression)	
Final Demo & Discussion	30	Professional presentation, insights, and ability to explain findings and limitations	Students apply one or more techniques (discussed in the course) to real world data.
Github Repo Quality	15	Organized structure, reproducibility, Notebooks should be able to run in COLAB; documentation, README	