Data Science and Material Informatics

Course Syllabus

Course Code: MSE 589
Credits: 3
Instructor: Dr. Burhan Beycan
Duration: 14 weeks
Prerequisites: Materials Science, Statistics, Programming (Python/R), Linear Algebra

---

Course Description

This course provides a comprehensive introduction to data science methodologies specifically applied to materials informatics. Students will learn to extract insights from materials data, develop predictive models for materials properties, and apply statistical and computational methods to accelerate materials discovery and design. The course emphasizes practical applications using real materials datasets and industry-standard tools.

Learning Objectives

Upon completion of this course, students will be able to:

1. **Master data science workflows** for materials research applications
2. **Apply statistical methods** to analyze materials property relationships
3. **Develop predictive models** using machine learning techniques
4. **Design and implement** materials informatics pipelines
5. **Visualize and communicate** complex materials data insights
6. **Integrate experimental and computational** materials data effectively

Course Outline

Week 1-2: Introduction to Materials Informatics

- Historical perspective and current trends
- Materials data types and sources
- Data quality assessment and validation
- Introduction to materials informatics platforms

Week 2-3: Data Collection and Management

- Materials databases and repositories
- Web scraping for materials data
- Data integration from multiple sources
- Database design for materials research

Week 4-5: Exploratory Data Analysis

- Descriptive statistics for materials properties
- Data visualization techniques
- Correlation analysis and feature relationships
- Outlier detection and data cleaning

Week 6-7: Feature Engineering and Selection

- Materials descriptors and fingerprints
- Compositional and structural features
- Dimensionality reduction techniques
- Feature importance and selection methods

Week 8-9: Statistical Modeling

- Linear and nonlinear regression models

- Hypothesis testing for materials data
- ANOVA for experimental design
- Bayesian approaches to materials modeling

Week 10-11: Machine Learning Applications

- Supervised learning for property prediction
- Unsupervised learning for materials classification
- Model validation and cross-validation
- Ensemble methods and model optimization

Week 12-13: Advanced Analytics

- Time series analysis for process monitoring
- Survival analysis for materials degradation
- Network analysis for materials relationships
- Optimization algorithms for materials design

Week 14: Integration and Future Directions

- Automated materials discovery workflows
- Integration with experimental systems
- Uncertainty quantification and decision making
- Future trends in materials informatics

Assessment Methods

Component	Weight	Description
Weekly Assignments	30%	Data analysis and programming exercises
Midterm Project	25%	Comprehensive data analysis project
Final Project	30%	Independent materials informatics research
Participation & Quizzes	15%	Class engagement and knowledge assessments

Required Software and Tools

Programming Languages

- **Python** with scientific computing libraries
- **R** for statistical analysis (optional)
- **SQL** for database queries

Data Science Libraries

- **Pandas** for data manipulation
- **NumPy/SciPy** for numerical computing
- **Matplotlib/Seaborn** for visualization
- **Scikit-learn** for machine learning

Materials Science Tools

- **Pymatgen** for materials analysis
- **MATMINER** for feature generation
- **Materials Project API**
- **AFLOW** and **OQMD** databases

Specialized Software

- **Jupyter Notebook** for interactive analysis
- **Tableau** or **Power BI** for advanced visualization
- **Apache Spark** for big data processing

Textbooks and Resources

Primary Textbooks

- **"Introduction to Statistical Learning"** by James, Witten, Hastie, and Tibshirani
- **"Python for Data Analysis"** by Wes McKinney
- **"Materials Informatics"** by Krishna Rajan

Supplementary Resources

- Materials Project documentation and tutorials
- Kaggle materials science competitions
- GitHub repositories for materials informatics
- Recent papers from npj Computational Materials

Laboratory Exercises

Lab 1: Data Exploration and Visualization

Comprehensive exploration of a large materials dataset, including data cleaning, statistical summaries, and advanced visualization techniques.

Lab 2: Feature Engineering Workshop

Hands-on experience creating materials descriptors from crystal structures and compositions, including both traditional and modern approaches.

Lab 3: Predictive Modeling Challenge

Building and comparing multiple machine learning models for predicting specific materials properties, with emphasis on model evaluation and interpretation.

Lab 4: Materials Discovery Pipeline

Development of an end-to-end pipeline for materials discovery, from data collection through model deployment and validation.

Major Projects

Midterm Project: Materials Property Analysis

Students will conduct a comprehensive analysis of a specific materials property (e.g., thermal conductivity, elastic moduli) using multiple datasets and analytical approaches.

Final Project Options

Option A: Predictive Model Development

Develop and validate a novel predictive model for an important materials property, including comparison with existing approaches and discussion of practical applications.

Option B: Materials Database Analysis

Comprehensive analysis of a large materials database to identify trends, patterns, and opportunities for new materials discovery.

Option C: Experimental Data Integration

Design and implement a system for integrating experimental and computational materials data to improve predictive capabilities.

Industry Applications

Case Studies

- Automotive materials selection and optimization
- Aerospace alloy development
- Energy storage materials discovery
- Pharmaceutical materials design

Guest Lectures

- Industry data scientists from materials companies
- National laboratory researchers
- Materials informatics software developers

Grading Scale

Grade	Percentage	Description
A	90-100%	Exceptional analytical skills and insights
B	80-89%	Strong performance with good methodology
C	70-79%	Adequate work meeting basic requirements
D	60-69%	Below average with significant gaps
F	<60%	Unsatisfactory performance

Course Policies

Data Ethics and Privacy

Students must adhere to ethical guidelines for data usage, respect intellectual property rights, and maintain confidentiality of proprietary datasets.

Reproducible Research

All analyses must be documented and reproducible. Code should be well-commented and version-controlled using Git.

Collaboration Policy

Collaboration is encouraged for learning concepts and discussing approaches, but all submitted work must be individually completed.

Late Assignment Policy

Late assignments will be penalized 10% per day unless prior arrangements are made with the instructor.

Professional Development

Skills Development

- Technical writing for data science reports
- Presentation skills for communicating results
- Project management for data science projects
- Collaboration tools and practices

Career Preparation

- Portfolio development for data science positions
- Interview preparation for materials informatics roles
- Networking opportunities with industry professionals
- Guidance on graduate school applications

Research Opportunities

Undergraduate Research

Opportunities to participate in ongoing materials informatics research projects with faculty and graduate students.

Conference Participation

Support for presenting work at materials science and data science conferences, including funding opportunities.

Publication Opportunities

Potential for contributing to peer-reviewed publications based on exceptional course projects.

Contact Information:
Dr. Burhan Beycan
Email: burhanbeycan@hotmail.com
Office Hours: Wednesdays and Fridays, 2:00-4:00 PM
Location: Materials Informatics Lab, Room 205