Intelligence Academy

The Center for Research and Development

Statistical Machine Learning

1. Overview & Logistics

- 👩‍🏫 **Instructor**: Mejbah Ahammad
- 🗓️ **Semester**: Spring Semester
- ⏰ **Class Time**: 8:00 PM – 10:00 PM
- 📆 **Class Days**: Saturday and Wednesday
- 💻 **Class Mode**: Remote (Zoom)
- 💰 **Course Fee**: ৳4000
- ☎️ **Contact Number**: +8801874603631
- ⏳ **Lessons & Time**: **20 Lessons**, **40 ঘন্টা 20 মিনিট** total
- 📧 **Email**: hello@softwareintelligence.ai
- 🌐 **Website**: http://softwareintelligence.ai/

2. Course Description

Statistical Machine Learning combines:

- 🎓 **Statistical Foundations**: Probability, distributions, estimation, hypothesis testing
- 📈 **Machine Learning Techniques**: Regression, classification, clustering, ensemble methods
- 💻 **Practical Coding**: Python (`NumPy` , `pandas` , `scikit-learn`) for model building and evaluation
- 🔎 **Advanced Topics**: Bayesian inference, neural networks, SVMs, interpretability, ethics
- 🗣️ **Communication**: Presentation, reporting, real-world problem-solving

By course end, participants will build robust ML pipelines using **statistical rigor** and **cutting-edge techniques**, culminating in a **capstone-style** final project or demonstration.

3. Learning Outcomes

1. 🎓 **Foundational Skills (Beginner)**
    - 👉 Recognize basic probability concepts, distributions, and inference methods.
    - 👉 Perform initial data cleaning, wrangling, and EDA in Python.
2. 📈 **Intermediate Skills**
    - 👉 Develop supervised ML models (linear/logistic regression, ensembles).
    - 👉 Understand regularization, model tuning, and cross-validation strategies.
3. 💡 **Advanced Skills**
    - 👉 Implement Bayesian methods, SVMs, neural networks, or advanced ensemble strategies.
    - 👉 Critically analyze model assumptions, interpret results, and address ethical concerns.
4. 🗣️ **Professional Communication**
    - 👉 Present findings to diverse audiences with clarity (visuals, reports).
    - 👉 Collaborate effectively, incorporating peer/instructor feedback.

4. Prerequisites

- 🎓 **Mathematics & Statistics**:
    - Basic probability (Normal, Binomial), inferential stats (t-tests, p-values).
    - Some exposure to linear algebra (matrix operations) and calculus (derivatives).
- 💻 **Programming**:
    - Familiarity with Python (lists, loops, functions).
    - Basic usage of `pandas`, `NumPy`, `matplotlib`, or similar.
- 💼 **Logistics & Tools**:
    - Stable internet for Zoom.
    - Python environment (Anaconda recommended).
    - Willingness to install additional Python packages (e.g., `scikit-learn`).

5. Course Materials

A. Required Texts/Readings

1. 📗 ***An Introduction to Statistical Learning (ISL)*** by James, Witten, Hastie, Tibshirani (Springer).
2. 📕 ***The Elements of Statistical Learning (ESL)*** by Hastie, Tibshirani, Friedman (Springer).

B. Recommended

- 📒 ***Pattern Recognition and Machine Learning*** by Christopher Bishop (Springer).
- 📕 ***Bayesian Data Analysis*** by Gelman et al. (CRC Press).
- Official Python & scikit-learn documentation.

C. Software

- 💻 **Python 3.x** (Anaconda)
- 📓 **Jupyter Notebook** / IDE (VSCode, PyCharm)
- 🖥️ **Zoom** for remote sessions

6. Schedule & Lessons (20 Classes, 40 Hours 20 Minutes)

Each **lesson** is designed for approximately **2 hours** (some might slightly exceed to total 40:20). Classes blend **theory** and **hands-on** demos or exercises.

Lesson	Topic	Level	Key Focus
1	📍 Course Introduction & Probability Review	Beginner	Syllabus, environment setup, distributions (Bernoulli, Normal), random variables
2	📍 Parameter Estimation (MLE & MAP)	Beginner → Intermediate	Likelihood functions, Bayesian vs. frequentist approaches, small coding demos
3	📍 Exploratory Data Analysis & Hypothesis Testing	Beginner → Intermediate	Data wrangling, missing values, EDA, t-tests, p-values, confidence intervals
4	📍 Linear Regression & Regularization	Intermediate	OLS, Ridge, Lasso, cross-validation, bias-variance trade-off
5	📍 Logistic Regression & Classification Metrics	Intermediate	Confusion matrix, precision/recall, ROC-AUC, cross-entropy

Lesson	Topic	Level	Key Focus
6	📍 Feature Engineering & Model Diagnostics	Intermediate	Categorical encoding, polynomial features, residual analysis, error estimation
7	📍 Bayesian Methods & Conjugate Priors	Intermediate → Advanced	Posterior updates, Beta-Bernoulli, Normal-Normal, MCMC basics
8	📍 Decision Trees & Ensemble Methods (Bagging, RF)	Intermediate → Advanced	CART, random forests, OOB errors, feature importance, bagging strategies
9	📍 Boosting & Advanced Ensemble (AdaBoost, XGBoost)	Intermediate → Advanced	Gradient boosting, sequential error correction, hyperparameter tuning
10	📍 Support Vector Machines (Foundations)	Intermediate → Advanced	Margin maximization, kernel trick, soft/hard margins
11	📍 SVM in Practice & Tuning	Advanced	RBF, polynomial kernels, grid/random search, practical pitfalls
12	📍 Dimensionality Reduction (PCA, LDA)	Intermediate → Advanced	Covariance, eigen-decomposition, linear discriminants, advanced manifold methods (optional)
13	📍 Clustering (K-means, Hierarchical, DBSCAN)	Intermediate → Advanced	Unsupervised basics, cluster validation (silhouette), dendrograms, density-based methods
14	📍 Probabilistic Graphical Models (Bayesian Networks)	Advanced	Conditional independence, factorization, small examples, potential software for inference
15	📍 Hidden Markov Models & Sequential Data	Advanced	Markov chains, forward-backward algorithm, Viterbi decoding, time-series aspects
16	📍 Neural Networks (MLP Intro)	Advanced	Perceptron, activation functions, backprop basics, capacity vs. data requirements

7. Detailed Lesson Descriptions

Lesson 1 (2 Hours)

📌 **Topic**: Course Introduction & Probability Review

- 🔎 **Focus**: Syllabus overview, environment check, distributions (Bernoulli, Normal), sampling basics
- 👉 **Assignment**:
  - Install libraries (NumPy, pandas, scikit-learn).
  - Short quiz on probability concepts.
- 💼 **Professional Insight**:
  - Probability underpins **risk modeling** (finance, insurance) and **quality control** (manufacturing).
  - Proper environment setup mirrors **DevOps** best practices for reproducible data science.

Lesson 2 (2 Hours)

📌 **Topic**: Parameter Estimation (MLE & MAP)

- 🔎 **Focus**: Likelihood functions, frequentist vs. Bayesian viewpoint, small coding demos
- 👉 **Assignment**:
  - Compare MLE and MAP estimates on a simple dataset (e.g., coin toss).
- 💼 **Professional Insight**:
  - MLE vs. MAP is critical in **marketing analytics** (conversion rates) or **medical testing** (disease prevalence).
  - Choosing an appropriate prior (MAP) can incorporate **domain knowledge** in real-world deployments.

Lesson 3 (2 Hours)

📌 **Topic**: Exploratory Data Analysis & Hypothesis Testing

- 🔎 **Focus**: Data wrangling, missing value handling, outlier detection, t-tests, p-values
- 👉 **Assignment**:
  - Clean a real dataset, run basic hypothesis tests (A/B style).
- 💼 **Professional Insight**:
  - **EDA** is ~80% of real data science tasks. Quick hypothesis tests guide **business decisions** (new product vs. old).
  - Communicating results in non-technical terms fosters **stakeholder trust**.

Lesson 4 (2 Hours)

📌 **Topic**: Linear Regression & Regularization (Ridge, Lasso)

- 🔎 **Focus**: OLS assumptions, bias-variance, cross-validation, controlling overfitting
- 👉 **Assignment**:
  - Compare OLS vs. Ridge vs. Lasso on a regression problem (e.g., housing).
- 💼 **Professional Insight**:
  - Common approach for **pricing strategy**, **sales forecasting**, and **resource planning**.
  - Regularization ensures stability in production, saving compute costs by preventing overfitting.

Lesson 5 (2 Hours)

📌 **Topic**: Logistic Regression & Classification Metrics

- 🔎 **Focus**: Confusion matrix, precision/recall, ROC-AUC, threshold tuning
- 👉 **Assignment**:
  - Classify Titanic-like data; interpret different metrics.
- 💼 **Professional Insight**:
  - Logistic regression is key in **credit scoring**, **churn prediction**, and **medical diagnosis**.
  - Understanding metrics aligns models with **business objectives** (precision vs. recall trade-offs).

Lesson 6 (2 Hours)

📌 **Topic**: Feature Engineering & Model Diagnostics

- 🔎 **Focus**: Encoding categorical variables, polynomial transformations, residual/error analysis
- 👉 **Assignment**:
  - Enhance feature set, compare performance gains, analyze errors thoroughly.
- 💼 **Professional Insight**:
  - In real-world ML, **feature engineering** often trumps fancy algorithms in terms of improvement.
  - Detailed error analysis helps refine **future data collection** and domain strategies.

Lesson 7 (2 Hours)

📌 **Topic**: Bayesian Methods & Conjugate Priors

- 🔎 **Focus**: Posterior derivation, Beta-Bernoulli, Normal-Normal, intro to MCMC tools
- 👉 **Assignment**:
  - Perform Bayesian inference on a small dataset; compare to frequentist results.
- 💼 **Professional Insight**:
  - Bayesian methods handle **low-data** or **high-uncertainty** environments (startups, medical research).
  - MCMC used in **complex risk** modeling (insurance, environmental studies).

Lesson 8 (2 Hours)

📌 **Topic**: Decision Trees & Ensemble Methods (Bagging, RF)

- 🔎 **Focus**: CART, random forests, OOB errors, feature importance
- 👉 **Assignment**:
  - Fit a decision tree & a random forest, compare error rates & interpret features.
- 💼 **Professional Insight**:
  - **Random forests** are widely used in finance, healthcare, e-commerce for their interpretability & performance.
  - Bagging strategies often reduce variance in high-stakes fields (credit risk, fraud detection).

Lesson 9 (2 Hours)

📌 **Topic**: Boosting & Advanced Ensemble (AdaBoost, XGBoost)

- 🔎 **Focus**: Sequential error correction, gradient boosting, hyperparameter tuning
- 👉 **Assignment**:
  - Evaluate AdaBoost vs. XGBoost on a classification/regression dataset; tune parameters.
- 💼 **Professional Insight**:
  - **XGBoost** dominates Kaggle competitions & corporate environments (marketing, sales forecasting).
  - Boosting algorithms can **quickly overfit** if not carefully tuned—an important skill in production ML.

Lesson 10 (2 Hours)

📌 **Topic**: Support Vector Machines (Foundations)

- 🔎 **Focus**: Margin maximization, kernel trick, soft/hard margins
- 👉 **Assignment**:
  - Implement SVM on a 2D classification problem, visualize decision boundaries.
- 💼 **Professional Insight**:
  - SVMs excel in **high-dimensional data** (text classification, genetics).
  - Understanding kernel selection is vital for certain **image** or **speech** tasks.

Lesson 11 (2 Hours)

📌 **Topic**: SVM in Practice & Tuning

- 🔎 **Focus**: RBF, polynomial kernels, grid/random search, practical pitfalls

- 👉 **Assignment**:
  - Use `GridSearchCV` to tune hyperparameters (C, gamma) on a real dataset.
- 💼 **Professional Insight**:
  - Proper parameter tuning can drastically shift **model accuracy**.
  - SVM remains a strong baseline in many **industrial** AI solutions.

Lesson 12 (2 Hours)

📌 **Topic**: Dimensionality Reduction (PCA, LDA)

- 🔎 **Focus**: Covariance, eigen-decomposition, supervised vs. unsupervised dimension reduction
- 👉 **Assignment**:
  - Apply PCA to a high-dimensional dataset; optionally compare LDA for classification.
- 💼 **Professional Insight**:
  - Reducing dimensionality helps in **visualization** and **speed** for real-time applications (IoT, sensor data).
  - LDA commonly used in **face recognition**, **medical classification** tasks.

Lesson 13 (2 Hours)

📌 **Topic**: Clustering (K-means, Hierarchical, DBSCAN)

- 🔎 **Focus**: Unsupervised basics, cluster validation, dendrograms, density-based algorithms
- 👉 **Assignment**:
  - Compare at least two clustering methods, interpret results with silhouette scores.
- 💼 **Professional Insight**:
  - Clustering widely used in **customer segmentation**, **market research**, and **anomaly detection**.
  - DBSCAN or hierarchical clustering can reveal **irregular** cluster structures in real data.

Lesson 14 (2 Hours)

📌 **Topic**: Probabilistic Graphical Models (Bayesian Networks)

- 🔎 **Focus**: Graph structures, conditional independence, small examples
- 👉 **Assignment**:
  - Construct or analyze a simple Bayesian network; perform basic inference.
- 💼 **Professional Insight**:
  - Graphical models appear in **medical diagnosis** (causal inference), **sensor fusion**, and **complex decision-making**.
  - Visualizing dependencies helps **stakeholders** grasp complicated relationships.

Lesson 15 (2 Hours)

📌 **Topic**: Hidden Markov Models & Sequential Data

- 🔎 **Focus**: Markov chains, forward-backward, Viterbi decoding, possible link to time series
- 👉 **Assignment**:
  - Implement an HMM for a toy sequence (e.g., weather states or text).
- 💼 **Professional Insight**:
  - HMMs are cornerstones in **speech recognition**, **bioinformatics** (DNA sequences), and **POS tagging** in NLP.
  - Sequential modeling is crucial in many real-time or streaming applications.

Lesson 16 (2 Hours)

📌 **Topic**: Neural Networks (MLP Intro)

- 🔎 **Focus**: Perceptron/MLP basics, activation functions, high-level backprop
- 👉 **Assignment**:
  - Train a small MLP on a classification dataset, discuss overfitting.
- 💼 **Professional Insight**:
  - Neural nets power advanced **computer vision** and **NLP** tasks in top tech companies.
  - Balancing model complexity vs. data is critical in **production** cost management.

Lesson 17 (2 Hours)

📌 **Topic**: Overfitting & Robust Validation

- 🔎 **Focus**: Dropout (in neural nets), early stopping, nested CV, data augmentation
- 👉 **Assignment**:
  - Show how advanced validation and regularization reduce overfitting in a prior model.
- 💼 **Professional Insight**:
  - Overfitting leads to **financial losses** (poor predictions) or **misdiagnoses** (healthcare).
  - Rigorous validation fosters **trust** and **reliability** in deployed ML systems.

Lesson 18 (2 Hours)

📌 **Topic**: Interpretability & Fairness in ML

- 🔎 **Focus**: Tools (LIME, SHAP), fairness and bias, ethical frameworks (GDPR)
- 👉 **Assignment**:
  - Analyze model outputs with SHAP on a selected dataset; discuss potential biases.

- 💼 **Professional Insight**:
  - Explainable AI is increasingly **required** in regulated sectors (finance, healthcare).
  - Addressing bias fosters **equitable** and **responsible** AI solutions.

Lesson 19 (2 Hours)

📌 **Topic**: Capstone Project Workshop

- 🔎 **Focus**: Data selection, model design, refining scope, peer/instructor Q&A
- 👉 **Assignment**:
  - Prepare a prototype or outline for your final project.
- 💼 **Professional Insight**:
  - Mimics **team stand-ups** or **project reviews** in corporate data science.
  - Early feedback loop ensures **agile** methodology and timely pivots.

Lesson 20 (2 Hours + 20 mins)

📌 **Topic**: Capstone Presentations & Course Wrap-Up

- 🔎 **Focus**: Student/Team presentations, Q&A, advanced resources, next steps in ML
- 👉 **Deliverable**:
  - Final code/report + demonstration.
  - Course feedback or survey.
- 💼 **Professional Insight**:
  - Polished presentations simulate **pitching** to executives or clients.
  - Reflecting on advanced directions (deep learning frameworks, big data) fosters **continual growth**.

8. Assessment & Grading

1. 📄 **Weekly/Regular Assignments (40%)**
   - 👉 Coding tasks, problem-solving, reflection papers.
   - Reinforces theory with hands-on practice.
2. 📝 **Quizzes (10%)**
   - 👉 Short checks on stats & ML fundamentals (announced or pop).
   - Encourages consistent revision.
3. 💼 **Capstone Project (40%)**

- 👉 End-to-end ML pipeline: data prep → modeling → validation → interpretability → presentation.
- Demonstrates integrated skills from the entire course.

4. 🤝 **Participation (10%)**
   - 👉 Active engagement, Q&A, breakout rooms, peer feedback.
   - Collaboration skill is essential in real-world DS teams.

🏷️ **Grade Scale**

- **A** = 90–100%
- **B** = 80–89%
- **C** = 70–79%
- **D** = 60–69%
- **F** = < 60%

9. Course Policies

1. 🏷️ **Attendance & Engagement**
   - Mandatory Zoom attendance (camera on recommended).
   - Inform absences in advance if possible.
2. 📢 **Communication**
   - Check email regularly for announcements.
   - Email hello@softwareintelligence.ai for questions or clarifications.
3. 🕛 **Late Submissions**
   - May incur penalties unless pre-approved.
   - Discuss extensions for valid reasons (health, emergencies).
4. ⚠️ **Academic Integrity**
   - No plagiarism or unapproved collaboration.
   - Violations follow institutional guidelines.
5. 💻 **Technical Setup**
   - Install/maintain Python environment (Anaconda).
   - Ensure stable internet, Zoom readiness.

10. Final Note

Welcome to *Statistical Machine Learning*! Over **20 Lessons** (total **40 hours 20 minutes**), expect an interactive deep dive into **stats** + **ML**. Keep in mind:

- **Practice** consistently with real datasets.
- **Engage** with peers for feedback and troubleshooting.
- **Document** your processes—transparency is key to professional data science.

We look forward to a **dynamic**, **hands-on** semester together!

👨‍🏫 **Instructor**: Mejbah Ahammad
☎️ **Phone**: +8801874603631
🌐 **Website**: http://softwareintelligence.ai/
📧 **Email**: hello@softwareintelligence.ai