Intelligence Academy

The Center for Research and Development

Statistical Data Science

1. 👩‍🏫 Instructor & Course Logistics

- 👩‍🏫 **Instructor**: Mejbah Ahammad
- 🗓️ **Semester**: Spring Semester
- ⏰ **Class Times**: 8:00 PM – 10:00 PM
- 📆 **Class Days**: Tuesday and Friday
- 💻 **Mode**: Remote (Zoom)

📢 Contact Information

- 📧 **Email**: hello@softwareintelligence.ai
- ☎️ **Phone**: +8801874603631
- 🌐 **Website**: http://softwareintelligence.ai/

2. 📝 Course Description

Statistical Data Science merges:

- 🎓 **Foundational Statistics** (probability, distributions, hypothesis testing)
- 📂 **Data Wrangling & EDA** (cleaning, transformation, exploration)
- 💻 **Machine Learning** (regression, classification, ensemble methods, clustering)
- 🧐 **Advanced Topics** (dimensionality reduction, Bayesian methods, interpretability)
- 🗣️ **Professional Communication** (reports, dashboards, ethical & business considerations)

Students will develop an **end-to-end data science pipeline**, culminating in a **capstone project** that illustrates practical application and professional best practices.

3. 🎯 Learning Outcomes

By the end of this course, you will:

1. 🎓 **Beginner-Level Skills**
    - 👉 **Understand** fundamental probability and descriptive statistics.
    - 👉 **Perform** basic data loading, cleaning, and visualization in Python.
2. 📈 **Intermediate-Level Skills**
    - 👉 **Apply** hypothesis testing, regression, classification, and clustering.
    - 👉 **Employ** feature engineering, dimensionality reduction, and ensemble methods.
3. 💡 **Advanced-Level Skills**
    - 👉 **Integrate** Bayesian methods, neural networks, or other specialized ML techniques.
    - 👉 **Assess** and mitigate model bias, interpret black-box models, and use fairness frameworks.
4. 🗣️ **Communication & Collaboration**
    - 👉 **Create** professional-quality visualizations and summaries for stakeholders.
    - 👉 **Collaborate** effectively in teams, giving and receiving structured feedback.

4. 🏷️ Prerequisites

1. 🎓 **Mathematics & Statistics**
    - Basic algebra, probability, and inferential statistics (e.g., normal distribution, p-values).
2. 💻 **Programming**
    - Proficiency in Python (data structures, basic scripting).
    - Familiarity with `NumPy`, `pandas`, `matplotlib`, `scikit-learn`.
3. 💼 **Logistics & Tools**
    - Reliable internet connection for Zoom.
    - Ability to install and manage Python environments (Anaconda recommended).

5. 📚 Course Materials

A. Required Texts/Readings

1. 📗 ***Practical Statistics for Data Scientists*** by Peter Bruce & Andrew Bruce (O'Reilly).
2. 📕 ***An Introduction to Statistical Learning (ISL)*** by James, Witten, Hastie, Tibshirani (Springer).

B. Recommended & Advanced

- 📒 ***The Elements of Statistical Learning (ESL)*** by Hastie, Tibshirani, Friedman (Springer).
- 📘 ***Python for Data Analysis*** by Wes McKinney (O'Reilly).
- 📕 ***Bayesian Data Analysis*** by Gelman et al. (CRC Press).

C. Software & Tools

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

6. 🗓️ 10-Week Schedule & Format

- **10 Weeks** total, **20 classes** (two per week).
- Each class is **2 hours**: typically **theory** + **hands-on** coding/discussion.
- **Participation** is integral to mastering the material.

Week	Class	Level	Topic	Key Highlights
1	Class 1	Beginner	📍 **Course Intro & Probability Basics**	Syllabus overview, environment setup, discrete/continuous distributions
	Class 2	Beginner	📍 **Data Wrangling & EDA Fundamentals**	Missing values, outliers, summary stats, basic plots (pandas/seaborn)
2	Class 3	Beginner → Intermediate	📍 **Statistical Inference & Hypothesis Testing**	t-tests, p-values, confidence intervals, real vs. simulated data
	Class 4	Intermediate	📍 **ANOVA & Experimental Design**	One-way ANOVA, assumptions, multiple comparisons, A/B testing
3	Class 5	Intermediate	📍 **Linear Regression (Simple & Multiple)**	OLS derivation, assumptions, R-squared, residuals, coding with `sklearn`

Week	Class	Level	Topic	Key Highlights
	Class 6	Intermediate	📍 **Logistic Regression & Classification Metrics**	Confusion matrix, precision/recall, F1-score, ROC-AUC
4	Class 7	Intermediate	📍 **Feature Engineering & Selection**	Encoding (categorical, one-hot), polynomial features, feature importance
	Class 8	Intermediate	📍 **Regularization (Ridge, Lasso) & Bias-Variance**	Cross-validation, hyperparameter tuning, bias-variance trade-off
5	Class 9	Intermediate	📍 **Dimensionality Reduction (PCA, LDA)**	Eigen-decomposition, variance explained, optional t-SNE/UMAP for visualization
	Class 10	Intermediate	📍 **Clustering (K-means, Hierarchical, DBSCAN)**	Cluster metrics (silhouette), dendrograms, density-based approaches
6	Class 11	Intermediate	📍 **Ensemble Methods (Bagging, Random Forest, Boosting)**	Decision trees, random forests, AdaBoost/Gradient Boosting
	Class 12	Intermediate → Advanced	📍 **Time Series or Advanced Classifier**	Stationarity, ARIMA basics **OR** advanced algorithms (SVM, multi-class)

7. 📝 Assessment & Grading

1. 📄 **Weekly Assignments (40%)**
    - 👉 Coding tasks, problem sets, short reflections.
    - Reinforces both **conceptual** and **practical** skills.
2. 📝 **Quizzes (10%)**
    - 👉 Periodic checks (announced or pop).
    - Covers fundamental stats, ML, and Python usage.
3. 💼 **Capstone Project (40%)**

- 👉 Real-world data pipeline: wrangling → EDA → modeling → evaluation → presentation.
- Teams or individuals; final presentation + written report.

4. 🤝 **Participation (10%)**
   - 👉 Active Zoom attendance, Q&A, breakout discussions.
   - Peer reviews and constructive feedback are essential.

🏷️ **Grade Scale**

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

8. ⚖️ Course Policies

1. 🏷️ **Attendance & Engagement**
   - 👉 Timely Zoom attendance, camera encouraged. Notify absences in advance.
2. 📢 **Communication**
   - 👉 Important announcements via email. Check daily.
   - For help or clarifications, email hello@softwareintelligence.ai.
3. 🕐 **Late Submissions**
   - 👉 Potential penalties unless previously arranged.
   - Extensions granted for valid reasons (health, emergencies).
4. ⚠️ **Academic Integrity**
   - 👉 Plagiarism or unauthorized collaboration is prohibited.
   - Violations follow institutional policy.
5. 💻 **Technical Setup**
   - 👉 Ensure Python (Anaconda) is installed, Zoom stable.
   - Familiarity with version control (Git) is recommended for project work.

9. 🆘 Additional Support & Office Hours

- ⏰ **Office Hours**: By appointment (Zoom).
- 🎓 **Extra Help**: Instructor can provide supplementary resources or 1-on-1 guidance.

10. 📑 Detailed Weekly Highlights with Professional Focus

Below, each class has **extra bullet points** under 💼 **Professional/Industry Focus** to show how these concepts apply in real-world settings and build your professional toolkit.

Week 1

Class 1

- 📌 **Topics**: Syllabus Overview, Probability (Discrete/Continuous), Environment Setup
- 👉 **Assignment**:
  - Install Python libraries (NumPy, pandas, etc.).
  - Short probability exercise (theoretical + coding).
- 💼 **Professional/Industry Focus**:
  - Understanding basic distributions is crucial for **risk assessment** (finance, insurance).
  - Proper environment setup mirrors **DevOps best practices** in real companies.

Class 2

- 📌 **Topics**: Data Wrangling & EDA (Missing Values, Outliers, Basic Plots)
- 👉 **Assignment**:
  - Clean a small dataset; produce summary statistics and quick visualizations.
- 💼 **Professional/Industry Focus**:
  - Data cleaning is ~80% of real data science work: **verifying data integrity** is key.
  - EDA presentations often inform **stakeholders** about potential business decisions.

Week 2

Class 3

- 📌 **Topics**: Inferential Statistics (t-tests, Confidence Intervals, p-values)
- 👉 **Assignment**:
  - Conduct hypothesis tests on real or simulated data.
  - Present a short report on findings.
- 💼 **Professional/Industry Focus**:
  - Hypothesis testing underpins **A/B testing** in product optimization, marketing campaigns.
  - Communicating p-values/conclusions to non-technical **business leaders** is a vital skill.

Class 4

- 📌 **Topics**: ANOVA & Experimental Design (One-way ANOVA, A/B Testing)
- 👉 **Assignment**:
  - Compare multiple group means, interpret significance.
- 💼 **Professional/Industry Focus**:
  - A/B or multi-variant tests are standard in **e-commerce** (website design changes, user experience).
  - Solid experimental design prevents costly **misinterpretations** in real projects.

Week 3

Class 5

- 📌 **Topics**: Linear Regression (Simple & Multiple), OLS, Assumptions
- 👉 **Assignment**:
  - Apply multiple regression on a real dataset (e.g., housing prices).
  - Evaluate residuals, R-squared.
- 💼 **Professional/Industry Focus**:
  - Linear regression is the backbone for **forecasting** sales, **pricing** strategies, and resource planning.
  - Understanding assumptions is essential to avoid **legal/ethical pitfalls** (e.g., biased predictions in finance).

Class 6

- 📌 **Topics**: Logistic Regression & Classification Metrics (Precision, Recall, F1, ROC-AUC)
- 👉 **Assignment**:
  - Classification on Titanic-like dataset, interpret confusion matrix.
- 💼 **Professional/Industry Focus**:
  - Logistic regression is widely used in **credit risk** modeling, **customer churn** prediction.
  - Choosing the right metric (precision vs. recall) matters for applications like **medical diagnostics** vs. **spam detection**.

Week 4

Class 7

- 📌 **Topics**: Feature Engineering & Selection (Encoding, Polynomial Features, Feature Importance)
- 👉 **Assignment**:
  - Transform features, compare model performance with/without these transformations.
- 💼 **Professional/Industry Focus**:
  - Good feature engineering can drastically **reduce model complexity** and cost in production.
  - Feature selection helps in **compliance** scenarios (regulatory audits on used data fields).

Class 8

- 📌 **Topics**: Regularization (Ridge, Lasso) & Bias-Variance
- 👉 **Assignment**:
  - Tune alpha in Ridge/Lasso; compare error rates.
- 💼 **Professional/Industry Focus**:
  - Regularization is crucial for **financial forecasting** or **marketing analytics** where overfitting can be expensive.
  - Cross-validation is an industry standard for **robust model validation** before deployment.

Week 5

Class 9

- 📌 **Topics**: Dimensionality Reduction (PCA, LDA, Optional t-SNE)
- 👉 **Assignment**:
  - PCA on a high-dimensional dataset; interpret principal components.
- 💼 **Professional/Industry Focus**:
  - PCA is essential in **high-dimensional** scenarios (e.g., genetics data, sensor data).
  - Reducing features can improve **processing speed** and help in real-time applications.

Class 10

- 📌 **Topics**: Clustering (K-means, Hierarchical, DBSCAN)
- 👉 **Assignment**:
  - Apply at least two clustering methods; evaluate with silhouette score.
- 💼 **Professional/Industry Focus**:
  - Clustering is pivotal for **customer segmentation** and **market research**.
  - Hierarchical clustering often used in **gene expression** analysis or **text analytics**.

Week 6

Class 11

- 📌 **Topics**: Ensemble Methods (Bagging, Random Forest, Boosting)
- 👉 **Assignment**:
  - Compare random forest & gradient boosting on a classification or regression dataset.
- 💼 **Professional/Industry Focus**:
  - Ensemble methods dominate **Kaggle** competitions and are widely used in **finance** (fraud detection) and **healthcare** (diagnostics).
  - Random forests offer interpretability advantages in **regulatory** contexts compared to black-box models.

Class 12

- 📌 **Topics**: Time Series or Advanced Classifier (Choose Focus)
  - **Option A**: Time Series – Stationarity, ARIMA, seasonal patterns

- **Option B**: Advanced Classification – SVM, multi-class strategies

- 👉 **Assignment**:
  - Forecast a simple time series **OR** tune an SVM for a multi-class dataset.
- 💼 **Professional/Industry Focus**:
  - Time series forecasting is critical in **inventory management**, **financial trading**.
  - Advanced classifiers (SVM) are used for **image classification**, **bioinformatics**.

Week 7

Class 13

- 📌 **Topics**: Bayesian Methods & Probabilistic Modeling (Priors, Posterior, MCMC Intro)
- 👉 **Assignment**:
  - Implement Bayesian updates on a small dataset; compare to frequentist approach.
- 💼 **Professional/Industry Focus**:
  - Bayesian inference is key in **medical trials**, **market research** (incorporating prior knowledge).
  - MCMC methods are used in **complex risk modeling** (e.g., insurance, actuarial science).

Class 14

- 📌 **Topics**: Neural Networks (MLP) – Activation Functions, Feedforward Architecture
- 👉 **Assignment**:
  - Train a small MLP on a classification dataset (e.g., MNIST or tabular).
- 💼 **Professional/Industry Focus**:
  - Neural nets power **computer vision** (e-commerce product tagging) and **NLP** (chatbots, sentiment).
  - Balancing data requirements vs. model complexity is crucial for cost and performance in production.

Week 8

Class 15

- 📌 **Topics**: Model Evaluation & Interpretability (CV pitfalls, LIME/SHAP, Fairness)
- 👉 **Assignment**:
  - Apply an interpretability tool to a trained model; analyze bias or feature impact.
- 💼 **Professional/Industry Focus**:
  - Many industries (finance, healthcare) **require** interpretability to comply with regulations.
  - Tools like SHAP help **build trust** with clients and executives.

Class 16

- 📌 **Topics**: MLOps & Model Deployment (Flask/FastAPI, Docker, CI/CD)
- 👉 **Assignment**:
  - Containerize a model and deploy a simple API locally or on a cloud platform.
- 💼 **Professional/Industry Focus**:
  - Productionizing models is a **core skill** for data scientists in tech companies.
  - Docker/CI-CD ensures **reproducibility** and quick iteration in enterprise solutions.

Week 9

Class 17

- 📌 **Topics**: Capstone Project Workshop (Data Debugging, Methodology Refinement)
- 👉 **Assignment**:
  - Submit capstone progress outline or preliminary code.
- 💼 **Professional/Industry Focus**:
  - Project management (timeline, scope) aligns with **agile methodologies** used in industry.
  - Peer feedback mimics **code reviews** or **project stand-ups** in real teams.

Class 18

- 📌 **Topics**: Capstone Presentations (Part 1)
- 👉 **Deliverable**:
  - Live demos, peer Q&A, instructor critique.

- 💼 **Professional/Industry Focus**:
  - Presentation skills are essential when **pitching data insights** to C-level executives or non-tech stakeholders.
  - Showcasing end-to-end solutions fosters a **consultative approach** to data problems.

Week 10

Class 19

- 📌 **Topics**: Capstone Presentations (Part 2)
- 👉 **Deliverable**:
  - Remaining presentations, advanced discussion of methodology.
- 💼 **Professional/Industry Focus**:
  - Final demos reflect **client-facing** scenarios in consulting or internal data science teams.
  - Handling tough Q&A showcases **confidence** and readiness for industry interviews or stakeholder sessions.

Class 20

- 📌 **Topics**: Course Wrap-Up & Future Directions (Big Data, Deep Learning, Specialized Domains)
- 👉 **Assignment**:
  - Submit final capstone code/report.
  - Complete course evaluation survey.
- 💼 **Professional/Industry Focus**:
  - Understanding next steps (Spark/big data, advanced deep learning) is essential for **scaling** solutions.
  - Networking, continuous learning, and professional development keep **data scientists** at the cutting edge.

✅ Final Note

Welcome to *Statistical Data Science*! Over the next **10 weeks**, we will bridge **fundamental statistics** and **modern data science practices**, with each class enriched by **professional insights**. Keep these key points in mind:

- **Practice** regularly and **experiment** with different datasets.
- **Communicate** your work effectively—technical mastery + clarity = real-world impact.
- **Collaborate** and **ask questions**—learning from peers is invaluable.

We look forward to a **dynamic and career-focused** semester together!

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