BP703T · B.Pharm Semester VII · Theory Notes

AI in Clinical Applications

Unit V — Guided Supervised Learning Project (6 hours)

Unit V (6 hours)

Pills & Python Notes

From-zero teaching standard: every technical term is defined before use.

Syllabus coverage: Unit V (6 hours).

PILLS & PYTHON

Pharmacy • Data • AI • Healthcare

Instagram @pillsandpython · pillsandpython.com · YouTube @PillsAndPython

CHAPTER 11: Guided Supervised Learning Project — Pharmaceutical and Clinical Applications

Learning Objectives

After studying this chapter, you will be able to:

1 List and explain the nine required steps of the guided supervised learning project.

2 Define every project term (dataset, predictor, outcome, preprocessing, metric) in simple words.

3 Work through a complete linear regression project (dissolution rate prediction) with all arithmetic shown.

4 Work through a complete logistic regression project (ADR prediction) with all steps shown.

5 Work through a third project (diabetes risk prediction) including data preprocessing calculations.

6 Evaluate a model with the correct metric for the problem type.

7 Discuss limitations, biases, and ethical issues of a student project honestly.

8 Plan and present project findings using the given outline and rubric.

Definitions

- **Project:** A planned piece of work with a question, a method, results, and conclusions.
- **Dataset:** A table of data (rows = patients/batches, columns = measurements) used for the project.
- **Publicly available dataset:** A dataset anyone can download and use, usually shared for learning or research.
- **Predictor variable:** A column used to make the guess (e.g., age). Also called a feature or input.
- **Outcome variable:** The column we want to guess (e.g., ADR yes/no). Also called the target or label.
- **Data preprocessing:** Cleaning and preparing the dataset before modeling — handling missing values, fixing errors, scaling numbers.
- **Missing value:** An empty cell where a measurement was not recorded.
- **Normalization (scaling):** Converting numbers to a common scale (e.g., 0 to 1) so that large numbers do not dominate the model.
- **Evaluation metric:** A number that tells us how good the model is (e.g., accuracy, RMSE).
- **Baseline:** A very simple reference model (e.g., "always predict the average") used to check whether our model is actually useful.
- **Presentation:** Explaining the project clearly to teachers or classmates with slides or a poster.

Terminology Table

Term	Simple Meaning	Example
Dataset	Table of rows and columns	10 batches × (hardness, dissolution %)
Predictor (feature)	Column used for guessing	Tablet hardness in kg/cm²
Outcome (label)	Column to be guessed	Dissolution % at 30 min
Training set	Rows used to build the model	7 of 10 batches
Testing set	Rows kept aside to check the model	3 of 10 batches
Preprocessing	Cleaning before modeling	Filling a missing hardness value
Normalization	Rescaling numbers to 0–1	Converting glucose 70–200 mg/dL to 0–1
RMSE	Average prediction error size	RMSE 4% means predictions miss by ~4%
R^2	Fraction of variation explained	R^2 0.94 = model explains 94%
Accuracy	Fraction of correct yes/no guesses	8 correct out of 10
Confusion matrix	Table of the four guess outcomes	TP, FP, TN, FN
Bias (project)	Unfairness from unrepresentative data	Toy data from one lab only
Ethics	Doing the work honestly and safely	Not presenting toy results as real findings
Overfitting	Memorizing the small dataset	Perfect on training, poor on new batches
Reproducibility	Another person can repeat your steps	All steps and numbers written down

11.1 The Nine Required Project Steps

Simple Explanation

Every project in this course follows the same nine steps, like a recipe. The steps turn a vague idea ("I want to predict something") into a finished, honest piece of work you can present.

The Nine Steps

Step	What You Do	Output
1. Problem definition	Write the exact question in one or two sentences	"Can disintegrant % predict dissolution % at 30 min?"
2. Dataset selection	Choose a suitable publicly available dataset (or a teacher-approved toy dataset)	Dataset name, source, size
3. Predictor/outcome identification	Name which columns are predictors and which is the outcome	Predictors: X1, X2; Outcome: Y
4. Data preprocessing	Clean: missing values, errors, scaling	Clean dataset ready for modeling
5. Model application	Apply linear regression (numbers) or logistic regression (yes/no)	Model equation
6. Performance evaluation	Calculate the right metrics	RMSE/R^2 or accuracy/sensitivity/specificity
7. Interpretation	Explain what the results mean in pharmacy/clinical words	"Each 1% more disintegrant adds ~8% dissolution"
8. Limitations, bias, ethics	Honestly list weaknesses	Small data, one lab, toy dataset
9. Presentation	Present findings to mentors/peers	Slides or poster + short talk

Flowchart: The Project Journey

```
1. Define the problem (one clear question)
       |
       v
2. Select dataset (public / teacher-approved toy)
       |
       v
3. Identify predictors (X) and outcome (Y)
       |
       v
4. Preprocess (clean, fill gaps, scale)
       |
       v
5. Apply model: linear regression (number?) or
   logistic regression (yes/no?)
       |
       v
6. Evaluate with correct metrics
       |
       v
7. Interpret in pharmacy/clinical language
       |
       v
8. Discuss limitations, bias, ethics honestly
       |
       v
9. Present findings
```

INFOGRAPHIC CONCEPT: "The project pipeline as a factory line" — Picture a conveyor belt with nine stations. Station 1: a question mark on a card (problem). Station 2: a filing cabinet (dataset). Station 3: two labeled jars, "X = clues" and "Y = answer" (predictors/outcome). Station 4: a washing machine (preprocessing/cleaning). Station 5: a calculator pressing out an equation (model). Station 6: a report card with grades (metrics). Station 7: a pharmacist explaining to a patient (interpretation). Station 8: a magnifying glass finding cracks (limitations). Station 9: a stage with a presenter (presentation). The belt only moves forward when each station's output is written down — reproducibility.

Key Takeaway

Nine steps, in order: problem → dataset → predictors/outcome → preprocessing → model → metrics → interpretation → limitations/ethics → presentation. Write everything down so anyone can repeat your work.

11.2 Worked Project A — Dissolution Rate Prediction by Linear Regression

Step 1 — Problem definition

"Can the percentage of disintegrant in a tablet formulation predict the percentage of drug dissolved at 30 minutes?" A higher dissolution rate usually means faster drug release.

Step 2 — Dataset selection

Illustrative toy dataset (invented for learning only; not real laboratory data). Six tablet batches:

Batch	Disintegrant x (%)	Dissolution y (% at 30 min)
1	1	50
2	2	59
3	3	65
4	4	75
5	5	81
6	6	90

Step 3 — Predictor and outcome

- Predictor (X): disintegrant % — known before testing.
- Outcome (Y): dissolution % at 30 min — the number we want to predict (regression problem, because it is a number).

Step 4 — Preprocessing

Check the table: no missing values, no impossible values (all between 0 and 100%), no duplicates. Data is clean; no changes needed. (In a real project, this is where you would fill gaps and fix errors.)

Step 5 — Apply linear regression (all arithmetic shown)

Mean of x = (1+2+3+4+5+6)/6 = 21/6 = 3.5. Mean of y = (50+59+65+75+81+90)/6 = 420/6 = 70.

x	y	x − 3.5	y − 70	$(x-3.5)^2$	$(x-3.5)(y-70)$
1	50	−2.5	−20	6.25	50
2	59	−1.5	−11	2.25	16.5
3	65	−0.5	−5	0.25	2.5
4	75	0.5	5	0.25	2.5
5	81	1.5	11	2.25	16.5
6	90	2.5	20	6.25	50
Total				**17.5**	**138**

- Slope b = 138 / 17.5 = 7.89 (rounded to two decimals: 7.89).
- Intercept a = 70 − 7.89 × 3.5 = 70 − 27.61 = 42.39 ≈ 42.4.

Model: predicted dissolution % = 42.4 + 7.89 × (disintegrant %).

Prediction example: for 2.5% disintegrant → 42.4 + 7.89 × 2.5 = 42.4 + 19.73 = 62.13 ≈ 62%.

Step 6 — Evaluate performance

Predictions for x = 1..6: 50.29, 58.18, 66.07, 73.96, 81.85, 89.74.

Actual	Predicted	Error	$Error^2$
50	50.29	−0.29	0.08
59	58.18	0.82	0.67

Actual	Predicted	Error	Error²
65	66.07	−1.07	1.14
75	73.96	1.04	1.08
81	81.85	−0.85	0.72
90	89.74	0.26	0.07
		Total	**3.76**

- MSE = 3.76 / 6 = 0.63 → **RMSE = √0.63 ≈ 0.79%** (predictions miss by less than 1 percentage point on average).
- Total y-variation = $(-20)^2 + (-11)^2 + (-5)^2 + 5^2 + 11^2 + 20^2$ = 400+121+25+25+121+400 = 1092.
- **$R^2 = 138^2 / (17.5 \times 1092) = 19044 / 19110 \approx 0.997$** — the model explains 99.7% of the variation.

Step 7 — Interpretation (pharmacy language)

Each additional 1% of disintegrant raises predicted dissolution by about 7.9 percentage points. The fit is excellent on this toy data ($R^2 \approx 0.997$, RMSE ≈ 0.8%). A formulator could use this relationship to target a desired dissolution rate — but must confirm with real experiments, because toy data cannot replace laboratory work.

Step 8 — Limitations, bias, ethics

- **Tiny dataset:** only 6 batches — real models need many more.
- **One factor only:** real dissolution also depends on compression force, particle size, and coating; ignoring them is a limitation.
- **Toy data:** these numbers were invented for learning; they prove the method, not the formulation.
- **Ethics:** presenting toy results as real lab findings would be dishonest — always label toy data clearly.

Step 9 — Presentation notes

Show the scatter plot with the best-fit line, the equation, RMSE/R^2, the plain-language interpretation, and one slide honestly listing limitations.

Key Takeaway

Project A shows the full regression path: clean toy data → line y = 42.4 + 7.89x → RMSE ≈ 0.8%, $R^2 \approx 0.997$ → interpretation in formulation language → honest limitations.

11.3 Worked Project B — ADR Prediction by Logistic Regression

Step 1 — Problem definition

"Can patient age group (above 60 or not) predict whether a patient will experience an adverse drug reaction (ADR)?" This is a yes/no question → classification.

Step 2 — Dataset selection

Illustrative toy dataset (invented for learning only; not real patient data). Eight patients on the same medicine:

Patient	Age > 60? (X)	ADR? (Y)
1	No (0)	No
2	No (0)	No
3	No (0)	No
4	No (0)	Yes
5	Yes (1)	Yes
6	Yes (1)	Yes
7	Yes (1)	Yes
8	Yes (1)	No

Step 3 — Predictor and outcome

- Predictor (X): age above 60 (0 = no, 1 = yes).
- Outcome (Y): ADR occurred (0 = no, 1 = yes).

Step 4 — Preprocessing

No missing values. Both columns are already 0/1 coded — no scaling needed for a single binary predictor. In a real project, age would be kept as a number and checked for missing entries.

Step 5 — Apply logistic regression (all arithmetic shown)

Logistic regression works with **odds** and **log-odds**:

- Odds = P(ADR) / P(no ADR). Log-odds = ln(odds). Probability = odds / (1 + odds).

Group 1 — age ≤ 60 (X = 0): 1 ADR out of 4 → P = 1/4 = 0.25; odds = 0.25/0.75 = 0.333; log-odds = ln(0.333) = −1.10.

Group 2 — age > 60 (X = 1): 3 ADR out of 4 → P = 3/4 = 0.75; odds = 0.75/0.25 = 3; log-odds = ln(3) = 1.10.

Model (log-odds form): log-odds of ADR = −1.10 + 2.20 × (age>60).

Check: for X = 0 → −1.10 → odds = e^(−1.10) = 0.333 → P = 0.333/1.333 = 0.25 ✓. For X = 1 → 1.10 → odds = 3 → P = 3/4 = 0.75 ✓.

Prediction rule: predict "ADR likely" if P ≥ 0.5, i.e., for age > 60 group.

Odds ratio: 3 / 0.333 = 9 — patients above 60 have 9 times the odds of ADR in this toy data.

Step 6 — Evaluate performance

Confusion matrix (predicted "ADR likely" for the 4 older patients):

	True ADR: Yes	True ADR: No
Predicted: ADR likely	TP = 3 (patients 5, 6, 7)	FP = 1 (patient 8)
Predicted: ADR unlikely	FN = 1 (patient 4)	TN = 3 (patients 1, 2, 3)

- Accuracy = 6/8 = 0.75 (75%)
- Sensitivity = 3/4 = 0.75 — catches 75% of true ADR cases

- Specificity = 3/4 = 0.75 — correctly clears 75% of non-ADR patients
- Precision = 3/4 = 0.75 — 75% of "ADR likely" flags are correct

Step 7 — Interpretation (clinical language)

In this toy data, being above 60 multiplies the odds of ADR by 9. The model correctly classifies 75% of patients. The one false negative (patient 4, younger but had an ADR) reminds us that age alone cannot explain all ADRs — medicines, dose, kidney function, and genetics also matter.

Step 8 — Limitations, bias, ethics

- **Tiny, invented data:** 8 patients cannot support real conclusions.
- **Single predictor:** real ADR risk needs many factors; this model is deliberately oversimplified.
- **Ethics:** a real ADR model that misses cases (false negatives) could harm patients — toy results must never guide real prescribing.
- **Bias risk:** if training data under-represents women or the elderly, the model will fail them (see Chapter 10).

Step 9 — Presentation notes

Present the 2×2 table, the odds ratio (9×), the confusion matrix with all four metrics, and emphasize: "toy demonstration of the method, not a clinical tool."

Key Takeaway

Project B shows the full classification path: binary toy data → log-odds model → odds ratio 9 → confusion matrix (accuracy/sensitivity/specificity/precision all 75%) → clinical interpretation → honest ethics warning.

11.4 Worked Project C — Diabetes Risk Prediction from Health Markers

Step 1 — Problem definition

"Can fasting blood glucose and BMI together predict whether a person has diabetes (yes/no)?"

Step 2 — Dataset selection

Illustrative toy dataset (invented for learning only; not real patient data). Six persons:

Person	Fasting glucose (mg/dL)	BMI (kg/m²)	Diabetes?
1	90	22	No
2	100	24	No
3	110	26	No
4	126	29	Yes
5	140	31	Yes
6	155	33	Yes

Step 3 — Predictor and outcome

- Predictors: fasting glucose (X1), BMI (X2).
- Outcome: diabetes yes/no (Y) — classification.

Step 4 — Preprocessing (with calculations)

Suppose person 3's BMI was missing. **Mean imputation:** mean of the other BMIs = (22+24+29+31+33)/5 = 139/5 = 27.8 → fill 27.8. (Our toy table is complete, so this is shown as the method you would use.)

Normalization (min-max to 0–1): formula = (value − min) / (max − min).

- Glucose min = 90, max = 155. Person 4: (126 − 90)/(155 − 90) = 36/65 = 0.554.
- BMI min = 22, max = 33. Person 4: (29 − 22)/(33 − 22) = 7/11 = 0.636.

Normalized table (glucose, BMI):

Person	Glucose (0–1)	BMI (0–1)	Diabetes?
1	0.000	0.000	No
2	0.154	0.182	No
3	0.308	0.364	No
4	0.554	0.636	Yes
5	0.769	0.818	Yes
6	1.000	1.000	Yes

Step 5 — Apply logistic regression (simplified hand-worked scoring)

For teaching, we use the trained toy model: log-odds = −4.0 + 5.0 × glucose_norm + 3.0 × BMI_norm. (In a real project, software finds these numbers from the training data; here they are given so you can practice the scoring steps by hand.)

Person 4: log-odds = −4.0 + 5.0(0.554) + 3.0(0.636) = −4.0 + 2.77 + 1.91 = 0.68 → odds = e^0.68 = 1.97 → P = 1.97/2.97 = 0.66 ≥ 0.5 → predict **Yes** ✓ (correct).

Person 3: log-odds = −4.0 + 5.0(0.308) + 3.0(0.364) = −4.0 + 1.54 + 1.09 = −1.37 → odds = e^(−1.37) = 0.254 → P = 0.254/1.254 = 0.20 < 0.5 → predict **No** ✓ (correct).

Person 2: log-odds = −4.0 + 5.0(0.154) + 3.0(0.182) = −4.0 + 0.77 + 0.55 = −2.68 → P = 0.064 → predict **No** ✓.

Person 5: log-odds = −4.0 + 5.0(0.769) + 3.0(0.818) = −4.0 + 3.85 + 2.45 = 2.30 → P = 0.91 → predict **Yes** ✓.

Person 1: log-odds = −4.0 → P = 0.018 → **No** ✓. Person 6: log-odds = −4.0 + 5.0 + 3.0 = 4.0 → P = 0.982 → **Yes** ✓.

Step 6 — Evaluate performance

All 6 predictions correct: accuracy = 6/6 = 1.00 (100%); sensitivity = 3/3 = 1.00; specificity = 3/3 = 1.00. (Perfect scores here only because the toy data is cleanly separated — real data is never this tidy.)

Step 7 — Interpretation

Higher glucose and higher BMI both push the log-odds up (positive coefficients 5.0 and 3.0), matching medical knowledge. The model separates the toy groups perfectly, but this reflects the toy's simplicity, not real-world performance.

Step 8 — Limitations, bias, ethics

- Only 6 persons; real diabetes models use thousands.
- Ignores family history, age, blood pressure, and other key factors.
- **Ethics:** a diabetes label affects insurance, employment, and mental health — a student model must never be used to label real people.
- Toy data labeled clearly; no real patient data was used.

Step 9 — Presentation notes

Show the preprocessing steps (imputation, normalization with one worked calculation), the scoring arithmetic for two persons, the perfect toy metrics with the warning that real data will score lower.

Key Takeaway

Project C adds preprocessing (missing-value filling, normalization) to the classification workflow and shows how two predictors combine in a log-odds equation — with all arithmetic visible.

11.5 Choosing Between Regression and Classification

Question type	Example	Model	Main metrics
"How much / how many?" (number)	Dissolution %, ICU stay days, tablet hardness	Linear regression	RMSE, R^2
"Yes or no / which group?" (category)	ADR yes/no, diabetes yes/no, batch pass/fail	Logistic regression	Accuracy, sensitivity, specificity, precision

Rule of thumb: look at your outcome column. Numbers → regression. Categories → classification.

11.6 Suggested Project Topics Mapped to the Syllabus

#	Suggested Topic (from syllabus)	Problem Type	Suggested Predictors → Outcome
1	Dissolution rate prediction	Regression	Disintegrant %, compression force → dissolution %
2	Tablet hardness prediction	Regression	Compression force, binder % → hardness
3	Stability degradation modelling	Regression	Temperature, humidity, time → % drug remaining
4	QC batch failure prediction	Classification	Process parameters → pass/fail
5	Assay variability modelling	Regression	Analyst, instrument, day → assay %
6	Impurity prediction	Regression	Storage conditions → impurity %

#	Suggested Topic (from syllabus)	Problem Type	Suggested Predictors → Outcome
7	Moisture impact on formulation stability	Regression	Moisture %, time → stability measure
8	Coating thickness prediction	Regression	Spray rate, pan speed → thickness
9	Solubility enhancement modelling	Regression	Carrier ratio, method → solubility
10	Tablet defect classification	Classification	Process readings → defect/no defect
11	ADR prediction	Classification	Age, dose, kidney function → ADR yes/no
12	Therapeutic response prediction	Classification	Patient markers → responder/non-responder
13	Hospital readmission risk prediction	Classification	Age, medicines, prior admissions → readmitted yes/no
14	Dose requirement prediction (PK data)	Regression	Weight, age, kidney function → dose
15	Diabetes risk prediction	Classification	Glucose, BMI, age → diabetes yes/no
16	Antibiotic treatment success prediction	Classification	Patient factors → success/failure
17	ICU stay duration prediction	Regression	Severity scores, age → days
18	Medication adherence analysis	Classification	Demographics, regimen complexity → adherent yes/no
19	Disease severity classification	Classification	Symptoms, labs → mild/moderate/severe

11.7 Presentation Tips

1 **One question per slide** — start with the problem in a single sentence.

2 **Show the data first** — a small table or plot before any equations.

3 **Show one worked calculation** — walk the audience through one prediction by hand.

4 **Metrics with meaning** — never say just "accuracy 75%"; add "so 3 of 4 true cases were caught."

5 **Limitations slide is mandatory** — honest weaknesses earn more marks than hidden ones.

6 **Label toy data every time** — say "illustrative toy dataset" on each data slide.

7 **End with the pharmacy message** — one sentence on why a pharmacist should care.

8 **Keep it short** — 8–10 slides for a 10-minute talk; rehearse once.

11.8 Sample Project Report Outline and Rubric

Report Outline (use these headings)

1 Title and student/group name

2 Problem definition (1–2 sentences)

3 Dataset description (source, size, rows × columns)

4 Predictors and outcome (named and justified)

5 Preprocessing steps (what was cleaned and how)

6 Model applied (regression or classification, with equation)

7 Worked calculations (at least one prediction by hand)

8 Performance metrics (correct metrics for the problem type)

9 Interpretation in pharmaceutical/clinical context

10 Limitations, potential biases, ethical considerations

11 Conclusion and presentation summary

12 References

Sample Marking Rubric (100 marks)

Component	Marks	What earns full marks
Problem definition	10	Clear, relevant, answerable question
Dataset selection & description	10	Suitable dataset, fully described, toy data labeled
Predictor/outcome identification	10	Correct variables, justified choice
Preprocessing	10	Missing values/errors handled, steps shown
Model application with workings	20	Correct model type, all arithmetic visible
Metrics & evaluation	15	Correct metrics, correctly calculated
Interpretation	10	Results explained in pharmacy/clinical language
Limitations, bias, ethics	10	Honest, specific, complete
Presentation & report quality	5	Clear, well-organized, correctly labeled

11.9 Pharmacy Relevance

- **Formulation development:** Regression projects (dissolution, hardness, stability) mirror real industrial R&D problems.
- **Clinical pharmacy:** Classification projects (ADR, readmission, adherence) mirror hospital decision-support problems.
- **Skill building:** The project teaches the full data-to-decision pipeline employers expect from modern pharmacists.
- **No coding needed:** Every step in this chapter is done with hand arithmetic, so any student can complete it.

11.10 India Context

- Students can use publicly shared teaching datasets or teacher-approved toy datasets; real hospital data needs ethics approval and patient consent, which is beyond a classroom project.
- Project topics like ADR prediction and adherence analysis connect directly to India's pharmacovigilance programme and public-health needs.
- Presenting in simple English with hand-worked steps makes the project accessible to students from diverse schooling backgrounds across India.

Key Points

1 The project has nine steps: problem → dataset → predictors/outcome → preprocessing → model → metrics → interpretation → limitations/ethics → presentation.

2 Regression = number outcomes (RMSE, R^2); classification = category outcomes (accuracy, sensitivity, specificity, precision).

3 Project A (dissolution): $y = 42.4 + 7.89x$, RMSE ≈ 0.8%, $R^2 \approx 0.997$ on toy data.

4 Project B (ADR): log-odds model, odds ratio 9 for age > 60, all metrics 75% on toy data.

5 Project C (diabetes): preprocessing (imputation, normalization) + two-predictor log-odds scoring, all hand-calculated.

6 Always label toy datasets as illustrative; never present toy results as real findings.

7 The limitations/ethics discussion is a scored, mandatory part of the project.

8 Present with 8–10 slides: data, one worked calculation, metrics with meaning, limitations, pharmacy message.

Common Confusions

Confusion	Clarification
"I must write computer code for this project"	No — this course's project is concepts plus hand-worked arithmetic; no programming is required.
"Toy data results prove the medicine works"	No — toy data demonstrates the method only. Real conclusions need real data and ethics approval.
"More predictors always make a better model"	No — too many predictors on tiny data cause overfitting.
"I should hide my model's mistakes"	No — limitations and errors are scored sections; honesty earns marks.
"Regression and classification metrics are interchangeable"	No — RMSE/R^2 for numbers; accuracy/sensitivity/specificity/precision for categories.

Common Mistakes

1 Choosing the wrong model type (regression for a yes/no outcome).

2 Evaluating on the same data used for training without saying so.

3 Forgetting to label the dataset as toy/illustrative.

4 Skipping preprocessing documentation ("the data was already clean" with no check shown).

5 Presenting metrics without interpretation in pharmacy language.

6 Omitting the limitations, bias, and ethics section.

Mnemonics

- **Nine steps — "Please Do Pick Pretty Models, Measure, Interpret, Limit, Present"**: Problem, Dataset, Pick (predictors/outcome), Preprocess, Model, Metrics, Interpret, Limitations, Present.
- **Regression vs classification — "Numbers Need Regression, Groups Get Classification."**
- **Metrics — "All Sick People Should Take medicine"**: Accuracy, Sensitivity, Precision, Specificity.

Exam Summary

Chapter 11 is the complete guide to the guided supervised learning project. The nine steps are: define the problem, select a dataset, identify predictors and outcome, preprocess, apply linear regression (number outcomes) or logistic regression (yes/no outcomes), evaluate with the correct metrics, interpret in pharmacy/clinical language, discuss limitations/bias/ethics, and present findings. Three fully worked toy projects are given: (A) dissolution rate prediction — line $y = 42.4 + 7.89x$, RMSE ≈ 0.8%, $R^2 \approx 0.997$; (B) ADR prediction — log-odds model with odds ratio 9 for age above 60, accuracy/sensitivity/specificity/precision all 75%; (C) diabetes risk prediction — mean imputation, min-max normalization, and two-predictor log-odds scoring, all by hand. Nineteen syllabus-suggested topics are mapped to problem types, with presentation tips and a 100-mark rubric. Every toy dataset is labeled illustrative; toy results demonstrate method, never real findings.

One-Minute Revision

Nine steps: problem, dataset, predictors/outcome, preprocessing, model, metrics, interpretation, limitations/ethics, presentation; numbers → linear regression (RMSE, R^2); yes/no → logistic regression (confusion matrix metrics); three worked toy projects (dissolution, ADR, diabetes) with all arithmetic shown; label toy data always; limitations slide mandatory; no coding required.

Short-Answer Questions

1 List the nine steps of the guided supervised learning project in order.

2 How do you decide between linear regression and logistic regression for a project?

3 What is data preprocessing? Name two preprocessing tasks with examples.

4 Show how to normalize the value 126 when min = 90 and max = 155.

5 In Project A, what does the slope 7.89 mean in pharmacy words?

6 In Project B, what does an odds ratio of 9 mean?

7 Why must toy datasets be labeled as illustrative?

8 Name three metrics for classification and three for regression.

9 What is mean imputation?

10 List any four headings of the sample project report outline.

Long-Answer Questions

1 Describe the nine steps of the guided supervised learning project, explaining what is produced at each step.

2 Work through Project A (dissolution rate prediction): build the regression line from the toy dataset, calculate RMSE and R^2, and interpret the results in formulation language.

3 Work through Project B (ADR prediction): derive the log-odds model, compute the odds ratio, build the confusion matrix, calculate all four metrics, and interpret each.

4 Explain data preprocessing with the examples from Project C: missing-value imputation and min-max normalization, showing all calculations.

5 Discuss limitations, potential biases, and ethical considerations that must appear in a student project report, using examples from the three worked projects.

MCQs

1 The first step of the guided project is:

a) Applying the model b) Defining the problem c) Calculating metrics d) Presenting findings

1 Predicting tablet hardness (a number) from compression force needs:

a) Logistic regression b) Linear regression c) A confusion matrix d) Normalization only

1 In Project A, RMSE ≈ 0.8% means:

a) The model explains 0.8% of variation b) Predictions miss by about 0.8 percentage points on average c) Accuracy is 0.8% d) 0.8% of batches failed

1 Mean imputation means:

a) Deleting the whole dataset b) Filling a missing value with the column mean c) Taking the average of all metrics d) Normalizing to 0–1

1 Which section is mandatory in the project report?

a) Only the metrics b) Limitations, bias, and ethical considerations c) Only the prediction equation d) None of these

Answer Key: 1-b, 2-b, 3-b, 4-b, 5-b

Viva Questions

1 What are the nine project steps?

2 When do you use linear regression vs logistic regression?

3 What does R^2 tell you?

4 How do you handle a missing value in your dataset?

5 What is an odds ratio?

6 Why is the limitations section important?

7 How would you present your project in 10 minutes?

Application-Based Questions

1 You want to predict whether a tablet batch will pass QC (pass/fail) from hardness and friability readings. Identify the problem type, model, predictors, outcome, and suitable metrics.

2 Given a toy dataset of 5 patients (age vs drug clearance), build a linear regression model by hand and predict clearance for a 45-year-old.

3 Your diabetes toy model scores 100% accuracy. Your teacher asks: "Is this model ready for a real clinic?" Write a five-point answer covering limitations, bias, and ethics.

References

- Aggarwal & Reddy, Healthcare Data Analytics
- Bonate, Pharmacokinetic–Pharmacodynamic Modeling and Simulation
- Campbell, Machin & Walters, Medical Statistics: A Textbook for the Health Sciences
- Schmidt & Derendorf, Applied Pharmacometrics
- Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating
- Strom, Kimmel & Hennessy, Pharmacoepidemiology