

Introduction to Machine Learning in Pharmaceutical Sciences (Theory)

BP301T

A Textbook for B. Pharm Semester III

As per the new PCI syllabus 2026 (NEP 2020)

Kunal Sharad Surwade

Senior Lecturer

SNJB's Shri Deepchand Fakirchand
Lodha Pharmacy College, Chandwad,
Nashik, India.

Dr. Yogesh B. Zambare

Professor and HOD of Department of
Pharmaceutical Chemistry

IVM's Krishnarao Bhegade Institute of
Pharmaceutical Sciences & Research,
Talegaon Dabhade, Maval, Pune, India

Mahadevan R. Swamy

Assistant Professor

Department of Pharmaceutical Chemistry,
PCET'S School of Pharmacy, Sate, Maval,
Pune, India



Introduction to Machine Learning in Pharmaceutical Sciences (Theory)

Course code: BP301T

A Textbook for B. Pharm Semester III

As per the new PCI syllabus 2026 (NEP 2020)

Authors: Kunal Sharad Surwade, Dr. Yogesh B. Zambare, Mahadevan R. Swamy

Edition: First Edition, September 2026

Language: English

Copyright © 2026 Wiseleaf Scientific Ventures Pvt. Ltd.

All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, except for brief quotations in reviews or scholarly works permitted by copyright law.

ISBN: 978-81-180383-6-8



MRP: 375/-

Published by: Wiseleaf Scientific Ventures Pvt. Ltd.

Address: Wework - Berger Delhi One Floor 19, C-001/A2, Sector 16b, Noida, Uttar Pradesh, 201301, India

Printed by: Patraloka | Prints and More

Immadihalli Main Rd, Immadihalli, Whitefield, Bengaluru, Karnataka 560066

Printed in India

Preface

The pharmaceutical sciences are undergoing a remarkable transformation driven by advances in digital technologies, data science, artificial intelligence, and computational methods. Among these emerging technologies, **Machine Learning (ML)** has gained considerable importance for its ability to identify patterns, learn from complex datasets, and support informed decision-making. Its applications are increasingly visible across drug discovery, formulation development, pharmacology, pharmacovigilance, clinical research, healthcare analytics, and pharmaceutical manufacturing.

The book “**Introduction to Machine Learning in Pharmaceutical Sciences (Theory)**” has been developed to provide undergraduate pharmacy students with a clear and structured introduction to the fundamental concepts of machine learning and their relevance to pharmaceutical sciences. The objective is not merely to introduce computational terminology, but to help students understand how data-driven approaches can be applied to solve real-world pharmaceutical problems.

The book begins with the basic concepts of artificial intelligence, machine learning, datasets, variables, and data preprocessing before progressing towards important learning approaches and commonly used machine-learning techniques. Concepts such as **supervised and unsupervised learning, classification, regression, clustering, model evaluation, feature selection, and basic predictive modelling** are presented in a systematic manner. Wherever appropriate, pharmaceutical examples and applications are incorporated to connect theoretical concepts with practical areas of pharmacy.

Particular emphasis has been placed on explaining concepts in a **simple, logical, and student-friendly manner**. Tables, illustrations, workflows, examples, and concise explanations have been included to facilitate conceptual understanding and revision. The book also introduces students to emerging applications of machine learning in areas such as **drug discovery and development, pharmacokinetics, pharmacodynamics, pharmaceutical formulation, medical data analysis, personalized medicine, pharmacovigilance, and pharmaceutical manufacturing**.

As machine learning is fundamentally dependent on data, the book also highlights the importance of **data quality, appropriate model selection, interpretation of results, validation, ethics, privacy, and responsible use of artificial intelligence** in pharmaceutical and healthcare settings. Understanding these aspects is essential for future pharmacy professionals who will increasingly work alongside digital and intelligent systems.

This book has been prepared with the learning requirements of **B.Pharm students** in mind and is intended to serve as a foundation for further study in artificial intelligence, data analytics, computational pharmacy, and related interdisciplinary fields. It may also be useful to teachers, researchers, and students seeking an introductory understanding of the intersection between machine learning and pharmaceutical sciences.

It is hoped that this book will encourage students to look beyond conventional approaches and appreciate the growing role of **data-driven intelligence in modern pharmacy**. The authors sincerely hope that the knowledge presented here will help learners develop curiosity, analytical thinking, and confidence to explore the rapidly evolving landscape of machine learning and its applications in pharmaceutical sciences.

Authors

Kunal Sharad Surwade

Dr. Yogesh B. Zambare

Mahadevan R. Swamy

Course Objectives

The objectives of this course are to:

1. Introduce the fundamental concepts of Artificial Intelligence, Machine Learning, and Data Science and their relevance to pharmaceutical sciences.
2. Develop conceptual understanding of machine learning workflows, including model building, training–testing strategies, and evaluation.
3. Explain the principles and applications of regression models for predicting continuous outcomes in pharmaceutical and healthcare data.
4. Familiarize students with classification and tree-based models used for clinical prediction and decision-making.
5. Introduce unsupervised learning techniques and demonstrate their applications in pattern recognition and healthcare data analysis.

Course Outcomes

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

1. Explain the basic concepts of artificial intelligence and machine learning, including types of learning and model development.
2. Apply regression models to analyze pharmaceutical data and interpret model outputs and performance metrics.
3. Analyze classification models for predicting categorical clinical outcomes using evaluation measures such as accuracy, sensitivity, and ROC–AUC.
4. Interpret tree-based and ensemble learning models for decision-making in pharmaceutical and healthcare applications.
5. Utilize unsupervised learning techniques such as clustering to identify patterns in healthcare and pharmaceutical datasets.

Syllabus

Unit	Topics	No. of lectures
I	Foundations of Machine Learning Definition and scope of Artificial Intelligence, Machine Learning, and Data Science • Types of machine learning: supervised, unsupervised, and reinforcement learning • Features and labels, training data and testing data • Concept of model building and prediction • Train-test split, overfitting and underfitting • Conceptual explanation of bias-variance trade-off • Overview of the basic ML workflow	6 Hours
II	Regression Models in Healthcare Overview of the concepts in predictive modeling for continuous outcomes, with emphasis on interpretation of outputs, and applications in pharmaceutical sciences. • Linear regression and multiple linear regression • Model coefficients and their interpretation • Residuals and goodness-of-fit • Performance metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2 score • Implementation of regression models on pharmaceutical data (e.g. dose–response relationships, predicting drug dissolution rates, estimating PK parameters), and demonstration of predictive modeling using python libraries such as sklearn	6 Hours
III	Classification Models in Clinical Applications Overview of the following ML models used for categorical prediction, with emphasis on conceptual understanding, interpretation of outputs and applications in pharmaceutical sciences. • Logistic regression • Probability output and threshold selection • Confusion matrix • Accuracy, sensitivity, specificity, precision, recall • Intuitive understanding of ROC curve and AUC, and k-Nearest Neighbors • Demonstration of predictive modeling using pharmaceutical and clinical examples such as predicting adverse drug reactions, disease risk prediction, binary therapeutic outcome modeling.	6 Hours
IV	Tree-Based Models & Ensemble Learning Overview of the following ML models with emphasis on intuitive and conceptual understanding. • Decision Trees (structure and splitting criteria), interpretation of	6 hours

	decision paths and feature importance • Random Forest, overview of ensemble concept • Advantages and limitations of tree-based models • Demonstration of these models for pharmaceutical and clinical applications such as ADR risk stratification, Patient classification, Predicting treatment outcomes etc.	
V	Unsupervised Learning & Practical Case Studies Overview of the unsupervised learning with emphasis on intuitive and conceptual understanding of pattern recognition and clustering, and their applications in pharmaceutical sciences. • Concept of unsupervised learning • Clustering overview, K-means clustering, choosing number of clusters, Interpreting cluster outputs • Mini-case study integrating regression, classification, or clustering on healthcare datasets to demonstrate Patient segmentation, and drug grouping based on properties.	6 Hours

Contents

Preface.....	iii
Course Objectives	iv
Course Outcomes	iv
Contents	vii
Abbreviations.....	xii
Chapter 1: Artificial Intelligence, Machine Learning and Data Science: Scope and Types of Learning	1
Learning outcomes	1
1.1 Introduction	1
1.2 Artificial intelligence, machine learning and data science.....	1
1.3 Types of machine learning	4
1.4 Features and labels	7
1.5 Training data and testing data	8
1.6 Model building and prediction	9
1.7 Common misconceptions	10
Quick recall	10
Chapter summary	11
Self-assessment	11
References	14
Chapter 2: Building a Model: Data, Training, Testing and the Bias-Variance Trade-off	15
Learning outcomes	15
2.1 Introduction	15
2.2 The basic machine learning workflow	15
2.3 Train-test split, validation and cross-validation	18
2.4 Underfitting and overfitting	20
2.5 The bias-variance trade-off	22
2.6 Reducing overfitting in practice.....	23
2.7 The workflow applied: a formulation example	24
2.8 Common misconceptions	25
Quick recall	25
Chapter summary	25
Self-assessment	26

References	28
Chapter 3: Linear Regression: Fitting a Line to Pharmaceutical Data	29
Learning outcomes	29
3.1 Introduction	29
3.2 The straight-line model	29
3.3 Fitting the line: residuals and least squares	30
3.4 Interpreting the coefficients	32
3.5 Residuals and goodness-of-fit	33
3.6 Assumptions and limitations of the straight line	34
3.7 Correlation and its relation to R^2	35
3.8 Worked example: a dose-response line	36
3.9 Worked example: first-order elimination and half-life	37
3.10 Linear regression in scikit-learn	37
3.11 Common misconceptions	38
Quick recall	38
Chapter summary	38
Self-assessment	39
References	41
Chapter 4: Multiple Regression, Goodness-of-Fit and Regression in Practice	42
Learning outcomes	42
4.1 Introduction	42
4.2 Multiple linear regression	42
4.3 Performance metrics for regression	44
4.4 The regression workflow in scikit-learn	46
4.5 Applications in pharmaceutical sciences	48
4.6 Limitations of regression models	49
4.7 Common misconceptions	49
Quick recall	49
Chapter summary	50
Self-assessment	50
References	53
Chapter 5: Logistic Regression and the Confusion Matrix	54
Learning outcomes	54

5.1 Introduction	54
5.2 From the straight line to the sigmoid	54
5.3 From probability to decision: the threshold	57
5.4 The confusion matrix	57
5.5 Metrics derived from the confusion matrix.....	59
5.6 The effect of the threshold	60
5.7 Logistic regression in scikit-learn	60
5.8 Applications in clinical prediction	61
5.9 Common misconceptions	62
Quick recall	62
Chapter summary	62
Self-assessment	63
References	65
Chapter 6: ROC Curves, AUC and k-Nearest Neighbours in Clinical Prediction.....	66
Learning outcomes	66
6.1 Introduction	66
6.2 The ROC curve	66
6.3 The area under the ROC curve (AUC).....	68
6.4 k-Nearest Neighbours.....	69
6.5 Both classifiers in scikit-learn	72
6.6 Applications: three clinical demonstrations	73
6.7 Common misconceptions	74
Quick recall	74
Chapter summary	74
Self-assessment	75
References	77
Chapter 7: Decision Trees: Structure, Splitting and Interpretation.....	78
Learning outcomes	78
7.1 Introduction	78
7.2 Structure of a decision tree.....	78
7.3 Measuring impurity: Gini and entropy.....	79
7.4 Worked example: an ADR risk tree grown by hand	80
7.5 Growing, stopping and pruning.....	82

7.6 Decision paths and feature importance	82
7.7 Decision trees in scikit-learn	83
7.8 Trees for numeric labels	84
7.9 Advantages and limitations of single trees	84
7.10 Applications	85
7.11 Common misconceptions	85
Quick recall	85
Chapter summary	86
Self-assessment	86
References	88
Chapter 8: Random Forest and Ensemble Learning	89
Learning outcomes	89
8.1 Introduction	89
8.2 The ensemble concept	89
8.3 Bootstrap sampling and bagging	90
8.4 The random forest	91
8.5 Single tree and random forest compared	93
8.6 Advantages and limitations of tree-based models	94
8.7 Random forest in scikit-learn	95
8.8 Applications	95
8.9 Common misconceptions	96
Quick recall	96
Chapter summary	96
Self-assessment	97
References	99
Chapter 9: Unsupervised Learning and K-means Clustering	100
Learning outcomes	100
9.1 Introduction	100
9.2 Supervised and unsupervised learning compared	100
9.3 Clustering and the K-means algorithm	101
9.4 Practical matters	103
9.5 Choosing the number of clusters	104
9.6 Interpreting the clusters	106