Course contents for AI and Machine Learning in Biology

1. **Department/Faculty:** FLSB
2. **Course Code:** LSB-603
3. **Course Title: AI and Machine Learning in Biology**
4. **Number of Credits:** Three
5. **Course Objectives:**

This syllabus introduces Biotechnology/Life Science students to AI/ML in a way that complements their background while preparing them for advanced research and industry applications. The syllabus can be adapted to emphasize high-level, human-readable tools like the Tidyverse and Tidymodels.

1. **Data Foundation and Data Processing and visualization**: To aware students about different types of biological data/datasets, variables, data structure etc. Know/how of processing biological data/datasets using R for implementing AI and ML algorithms. Lastly, how to visualize these results. (This Includes introduction to R Language).
2. **Understand AI, ML, and Deep Learning Foundations**: Introducing students with a foundational understanding of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL).
3. **Model Development and Optimization**: To train students in developing, evaluating, and optimizing machine learning and deep learning models for biotechnology-related tasks. This will done using specialized packages in R and unified interfaces for all algorithm (Decision Trees, SVM, Naive Bayes), to fit, evaluate, and optimize models consistently.
4. **Explore Practical Applications in Biotechnology**: To explore how AI, ML, and DL can be applied to solve real-world problems in biotechnology
5. **Problem Solving and Critical Thinking**: To foster problem-solving and critical thinking skills necessary for addressing complex challenges in biotechnology through AI, ML, and DL techniques.

6. Minimum prerequisites for taking this course, if any: Biostatistics

7. Course structure with units, if applicable:

Unit 1. Introduction to with R, Data Foundation, and Visualization (18 Hours)

- **The R Ecosystem for Biologists**: Introduction to R and RStudio as a platform for biological research.
- **Data Foundation (Tidyverse)**: Different Data types, Data Structures, variables. Use of dplyr and tidyr for data handling to manage large biological datasets effectively.
- **Data Processing**: Cleaning, normalizing, and transforming raw biological data into analysis-ready formats.
- **Visual Analytics**: Utilizing ggplot2 to create publication-quality heatmaps, scatter plots, and clustering visualizations for biological trend analysis.

Unit 2. Foundational Concepts of AI and Machine Learning (2 Hours)

- **Overview**: History and relevance of AI and ML in life sciences.
- **Learning Paradigms**: Introduction to supervised, unsupervised, and reinforcement learning within a biological context.
- **User-Centric Tools**: Exploring R-based platforms and Bioconductor tools that simplify ML for non-programmers

Unit 3. Supervised Learning for Biological Prediction (8 Hours)

- **The Tidymodels Framework**: Using a unified R interface for model development and optimization.
- **Regression & Classification**: Implementing Linear/Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Naive Bayes.
- **Life Science Applications**: Training models to predict disease markers or enzyme functions

Unit 4. Unsupervised Learning & Pattern Discovery (5 Hours)

- **Clustering Techniques**: Implementing Hierarchical and Partitioning Clustering to identify novel biological groupings.
- **Dimensionality Reduction**: Using PCA in R to simplify complex genomic or proteomic datasets.
- **Evaluation**: Assessing cluster validity and biological significance

Unit 5. Deep Learning for Sequences and Imaging (5 Hours)

- **Neural Network Architectures:** Introduction to the concepts of deep learning (DL).
- **Specialized Models:** Utilizing Convolutional Neural Networks (CNNs)/ and Recurrent Neural Networks (RNNs/LSTMs) for biological Application (examples/case study
- **R Interfaces:** Accessing Keras and Torch through R to implement DL models as a "user"

Unit 6. Bioinformatics Pipelines and Interactive Dashboards (Applied Unit) (4 Hours)

- **Neural Network Architectures:** Introduction to the concepts of deep learning (DL).
- **Specialized Models:** Utilizing Convolutional Neural Networks (CNNs)/ and Recurrent Neural Networks (RNNs/LSTMs) for biological Application (examples/case study
- **R Interfaces:** Accessing Keras and Torch through R to implement DL models as a "user"

8. a. Reference Books:

1. "Tidy Modeling with R: A Framework for Modeling in the Tidyverse" by Max Kuhn and Julia Silge.
2. "Data analysis for the life sciences with R" by Rafael A. Irizarry and Michael I. Love
3. "The New Statistics with R_ An Introduction for Biologists (Second Edition)" by Andy Hector
4. "A Primer in Biological Data Analysis and Visualization Using R" by Greg Hartivigsen
5. "A Primer in Biological Data Analysis and Visualization Using R (version 2)" by Greg Hartivigsen.
6. "A Guide to Applied Machine Learning for Biologists"

b. Other Resources:

R Resources

- **swirl**: learn R interactively from within the R console.
- R reference card (PDF) by Tom Short (more can be found under Short Documents and Reference Cards here)

- **Quick-R**: quick online reference for data input, basic statistics and plots
- Thomas Girke's R & Bioconductor manuals
- R programming class on Coursera, taught by Roger Peng, Jeff Leek and Brian Caffo
- The free "try R" class from Code School is also a good place to start: http://tryr.codeschool.com/
- Data structures summary by Hadley Whickham

Course Outcomes:

By the end of the course, students will be able to:

1. **Explain Core Concepts**: Able to demonstrate a clear understanding of AI, ML, and deep learning principles, algorithms, and their relevance to biotechnology.
2. **Process Biological Data:** Use R-based tools/packages to clean and visualize diverse life science datasets for implementation of Machine Learning or general analysis.
3. **Deploy ML Models: Able** Utilize the Tidymodels and R framework to implement and optimize predictive models.
4. **Build Pipelines:** Gain knowledge to create automated bioinformatics workflows using Bioconductor and R
5. **Solve Biotechnology Problems Using AI/ML/DL**: Helps them to address real-world biotechnology challenges using AI, ML, and DL.
6. **Create Dashboards:** Gain knowledge to create interactive Shiny applications to present AI-driven findings to the biotechnology industry.
7. **Apply Critical Thinking to Biotechnology Issues**: Approach complex biotechnology problems with critical thinking, creativity, and scientific rigor using AI, ML, and DL solutions.