Data Science/ML Full Stack

What we will do and gain?

- Build an in-depth understanding of all the data concepts.
- Create a strong social media profile on LinkedIn and GitHub.
- Build 15+ projects including 5+ Major Projects.
- Showcase your skills with a portfolio of real projects.
- Work on Live projects in parallel to understand how companies create end-to-end software solutions and apply ML models to real-life problems.

Technology Stack

- Python
- Data Structures
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Scikit-Learn
- Statsmodels
- Natural Language Toolkit (NLTK)
- Structure Query Language (SQL)
- HTML
- CSS
- JavaScript
- Bootstrap
- 5 Major Projects
- Git and GitHub
- NLTK
- SpaCy
- TensorFlow & Keras
- Sklearn
- Huggingface
- OpenCV
- Streamlit

1 | Python Programming and Logic Building

Learning Python programming and logic building enhances students' problem-solving skills, fosters logical thinking, and encourages creativity in designing solutions. It cultivates computational thinking, vital for diverse careers, offering a foundation for software development, data analysis, and automation, empowering students in an increasingly digital world.

- Python basics, Variables, Operators, Conditional Statements
- List and Strings
- While Loop, Nested Loops, Loop Else
- For Loop, Break, and Continue statements
- Functions, Return Statement, Recursion
- Dictionary, Tuple, Set
- File Handling, Exception Handling
- Object-Oriented Programming(Inheritance, Polymorphism, Encapsulation and Abstraction)
- Constructors, Decorators
- Modules and Packages

2 | Pandas, Numpy, Matplotlib

Learning Pandas enables efficient data manipulation, Numpy enhances numerical operations, and Matplotlib facilitates data visualization. Together, they empower students to handle data effectively, perform complex calculations, and create compelling visual representations, fostering strong analytical skills crucial across various disciplines and industries.

Numpy

- Vectors, Matrix
- Operations on Matrix
- Mean, Variance, and Standard Deviation
- Reshaping Arrays
- Transpose and Determinant of Matrix
- Diagonal Operations, Trace
- Add, Subtract, Multiply, Dot, and Cross Product.

Pandas

- Series and DataFrames
- Slicing, Rows, and Columns
- Operations on DataFrame
- Different ways to create DataFrame
- Read, Write Operations with CSV files
- Handling Missing values, replace values, and Regular Expression
- GroupBy and Concatenation

Matplotlib

- Graph Basics
- Format Strings in Plots
- Label Parameters, Legend
- Bar Chart, Pie Chart, Histogram, Scatter Plot

3 | Statistics

Statistics equips students with critical thinking skills to interpret data, make informed decisions, and solve real-world problems. It fosters analytical reasoning, enhances research abilities, and enables informed conclusions in diverse fields, empowering students to navigate and excel in an increasingly data-driven world.

Descriptive Statistics

- Measure of Frequency and Central Tendency
- Measure of Dispersion
- Probability Distribution
- Gaussian Normal Distribution
- Skewness and Kurtosis
- Regression Analysis
- Continuous and Discrete Functions
- Normality Test
- ANOVA
- Homoscedasticity
- Linear and Non-Linear Relationship with Regression

Inferential Statistics

- t-Test
- z-Test
- Hypothesis Testing
- Type I and Type II errors
- t-Test and its types
- One way ANOVA
- Two way ANOVA
- Chi-Square Test
- Implementation of continuous and categorical data

4 | Machine Learning

Understanding Machine Learning empowers students to grasp data patterns, make informed decisions, and innovate across disciplines. It fosters critical thinking, and problem-solving skills, and opens doors to careers in technology, research, and artificial intelligence-driven fields, preparing them for a future shaped by data-driven insights.

Supervised and Unsupervised Machine Learning:

- Linear Regression
- Logistic Regression
- Decision Tree
- Gradient Descent
- Random Forest (Bagging)
- Ridge and Lasso Regression
- Naive Bayes
- Support Vector Machine
- K-Means Clustering
- K-Nearest Neighbors (KNN)
- Principal Component Analysis (PCA)
- Dimension Reduction
- AdaBoost
- XGBoost
- Gradient Boosting
- DBSCAN & Hierarchical Clustering
- Linear Discriminant Analysis

Model Evaluation and Optimization:

1) Model Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC Curve
- AUC-ROC

2) Hyperparameter Tuning

- Random Search
- Grid Search

- Bayesian Optimization
- Randomized search CV
- K fold cross-validation

3) Overfitting and Regularization:

- Bias-Variance Trade-off
- Applying Regularization

4) EDA Techniques and Visualization

- Uni & multi Variate Analysis
- Data imputation
- Identifying and normalizing Outliers

5) Model Evaluation

- k-fold cross validation
- Stratified K-fold cross validation

Feature Engineering and Data Preprocessing

- Handling Class Imbalance
- Normalization and Standardisation
- Polynomial Feature Transformation
- One-hot Encoding
- Binning
- Feature Selection
- Feature Split
- Extracting Date
- Imputation
- Handling Outliers

5 | Computer Vision

Learning Computer Vision equips students to comprehend and create technologies that process visual information. It fosters problem-solving, innovation, and understanding of AI applications. Students gain skills in image recognition, robotics, healthcare, and more, empowering them for careers in tech, research, and diverse industries.

Introduction to OpenCV

- Basic Image read/write Operations
- Image Filters
- Edge detection
- Hough transform
- Corner detection
- Perspective Transformation
- HOG Filters

Basics and Fundamentals:

1) Neurons and Activation Functions:

- Biological Neurons
- Perceptrons
- Sigmoid, Tanh, ReLU, Leaky ReLU, and other activation functions

2) Feedforward Neural Networks (FNN):

- Architecture and layers
- Forward pass and backpropagation

3) Loss Functions:

- Mean Squared Error (MSE), Cross-Entropy, Huber loss
- Binary and multiclass classification losses

4) Hyper Parameter Tuning

5) Batch Normalization

Convolutional Neural Networks (CNN) & Neural Network Architectures:

1) Convolutional layers

- Introduction to Convolutional Neural Networks
- Convolution, Pooling, Padding & its mechanisms
- Forward Propagation & Backpropagation for CNNs
- CNN architectures
  - ✧ LeNet
  - ✧ AlexNet
  - ✧ VGGNet
  - ✧ InceptionNet
  - ✧ ResNet
  - ✧ EfficientNet
- Transfer Learning
- Data Augmentation

2) Recurrent Neural Networks (RNN):

- Introduction to Sequential data
- RNN and its mechanisms
- Vanishing & Exploding gradients in RNNs
- LSTMs - Long short-term memory

3) Auto encoders:

- Unsupervised learning for feature learning

Optimization and Regularization Techniques:

1) Gradient Descent and Backpropagation:

- Stochastic Gradient Descent (SGD)
- Mini-batch and batch gradient descent

2) Optimizers:

- Adam, RMSprop, Adagrad
- Learning rate schedules

3) **Weight Initialization:**

- Xavier/Glorot initialization
- Kaiming/He initialization

4) **Dropout and Batch Normalization:**

- Techniques to reduce overfitting

Hyperparameter Tuning:

1) **Learning Rate Tuning**
   - Grid search and random search

2) **Batch Size Optimization**
   - Influence on convergence and performance

Transfer Learning:

1) **Pre-trained Models:**
   Leveraging pre-trained models for new tasks
   Fine-tuning

6 | Natural Language Processing

Learning Natural Language Processing (NLP) equips students with tools to analyze, understand, and generate human language. It fosters critical thinking, improves problem-solving skills, and opens doors to fields like AI, data science, and linguistics, empowering students to innovate in communication, technology, and research.

Text Preprocessing

- Stop Words
- Tokenization
- Stemming and lemmatization
- Bag of Words Model
- N-grams
- Word Vectorizer
- TF-IDF

- POS Tagging
- Named Entity Recognition
- Word Embeddings

Natural Language Processing

- Text Classification
- Semantics and Sentiment Analysis

Sequence Models and Transformers

- RNN, LSTM, GRU
- Sequence to Sequence Models
- Attention mechanism
- Neural Machine Translation
- Transformers and Self-Attention
- Transformers, BERT, GPT-2

7 | Web Development

Learning web development with Streamlit and Flask equips students with practical skills to create interactive web apps (Flask) and data-driven, user-friendly interfaces (Streamlit). These tools foster coding proficiency, enhance problem-solving, and prepare students for tech careers by enabling them to build functional, modern applications.

Streamlit Flask

- Introduction to Flask
- Templates and Views
- Request Handling
- Flask Models
- Flask-RESTful
- Deployment
- CRUD web application with Flask

Git & Github

8 | Projects

1. **House Price Prediction (Regression):**

Importance : House Price Prediction through regression models like linear regression or decision trees utilizes property features to forecast house prices. This practice sharpens data analysis skills, aids in understanding real estate dynamics, and cultivates proficiency in predictive modeling—an essential skill in finance, real estate, and data-driven industries.

Description: Use a dataset containing features like the number of bedrooms, square footage, and location to predict the price of houses. Implement a regression model such as linear regression or decision tree regression.

2. **Iris Flower Classification (Classification)**:

Importance : The Iris Flower Classification task is crucial as it introduces beginners to machine learning by predicting iris species from specific features. Utilizing algorithms like logistic regression or decision trees on the famous Iris dataset aids in understanding classification models, feature importance, and model evaluation, forming a foundational step in ML learning.

Description: Use the famous Iris dataset to build a classification model that predicts the species of iris flowers based on features like sepal length, sepal width, petal length, and petal width. Try algorithms such as logistic regression or decision trees.

3. Diabetes Prediction (Classification):

Importance : Predicting diabetes using classification algorithms leverages health data to determine the likelihood of the disease. Employing tools like support vector machines, random forests, or gradient boosting aids in accurate predictions. This project not only hones data analysis skills but also contributes to proactive healthcare interventions, potentially improving patient outcomes.

Description: Utilize a dataset with various health indicators to predict whether a person is likely to have diabetes or not. Implement classification algorithms such as support vector machines, random forests, or gradient boosting.

4. Customer Churn Prediction (Classification):

Importance : Customer Churn Prediction is crucial for businesses. Analyzing customer data helps foresee potential departures, enabling proactive retention strategies. Using classification methods like logistic regression or neural networks, predicting churn becomes efficient, aiding companies in preserving customer loyalty and improving service quality based on user behavior patterns.

Description: Analyze a dataset containing customer information and usage patterns to predict whether a customer is likely to churn (leave) a service. Implement classification algorithms like logistic regression, random forests, or neural networks.

5. Blood Glucose Level Prediction (Regression):

Importance : Predicting blood glucose levels through machine learning regression aids in diabetes management. Using factors like diet, exercise, and health metrics, this predicts future levels. This project's significance lies in proactive health monitoring, offering insights crucial for individuals managing diabetes to make informed lifestyle adjustments for better health.

Description: Use machine learning regression techniques to predict blood glucose levels based on features such as diet, physical activity, and other health-related parameters. This project could be valuable for diabetes management.

6. Predicting Electrical Power Output from Solar Panels (Regression):

Importance : Predicting solar panel power output through regression models utilizes environmental data to forecast energy generation. It's crucial for optimizing solar energy harvesting, aiding in efficient resource utilization. This project not only enhances renewable energy utilization but also hones predictive modeling skills, vital in various industries.

Description: Develop a regression model to predict the electrical power output of solar panels based on environmental factors such as sunlight intensity, temperature, and time of day. This project is relevant for optimizing solar energy harvesting.

7. **Heart Disease Classification:**

Importance : Creating a Heart Disease Classification model is crucial for early detection and prevention. By analyzing patient data using machine learning, it predicts heart disease risk, aiding timely interventions. This proactive approach based on factors like age, cholesterol, and blood pressure improves patient care, potentially saving lives through early identification and treatment.

Description: Implement a machine learning classification model to predict the presence or absence of heart disease based on patient data, including features like age, cholesterol levels, and blood pressure.

8. **Tumor Detection in Medical Images using CNN:**

Importance: Tumor detection in medical images using CNN involves employing Convolutional Neural Networks to identify and classify tumors in MRI or CT scans. This project is crucial for accurate medical diagnostics, aiding in early detection and precise treatment, potentially saving lives through advanced technology in healthcare.

Description: Use a Convolutional Neural Network to classify and detect tumors in medical images, such as MRI or CT scans. This project is valuable for medical diagnostics.

9. **Biomedical Text Classification using NLP:**

Importance : Biomedical Text Classification with NLP categorizes scientific texts, aiding quick access to critical information in healthcare. By automating sorting into disease, treatment, or research categories, it accelerates data analysis, supports faster

research insights, and improves medical decision-making, enhancing overall healthcare efficiency and advancements.

Description: Apply NLP techniques to classify biomedical texts, such as scientific articles or clinical notes, into relevant categories, such as diseases, treatments, or research areas.

10. Chest X-ray Image Classification for Disease Diagnosis using CNN:

Importance : Chest X-ray image classification using CNN aids in swift and accurate diagnosis of respiratory illnesses like pneumonia or tuberculosis. Leveraging Convolutional Neural Networks (CNN) enables precise disease identification from images, expediting treatment and improving patient outcomes in healthcare.

Description: Develop a CNN to classify chest X-ray images for the diagnosis of respiratory diseases, such as pneumonia or tuberculosis.

11. Clinical Notes Summarization using NLP:

Importance : Clinical Notes Summarization using NLP condenses extensive medical records into succinct, informative summaries. This process enhances healthcare efficiency by offering quick access to vital patient information, aiding accurate decision-making for medical professionals. NLP-driven summarization saves time, ensures better communication, and facilitates streamlined patient care.

Description: Develop an NLP model for summarizing lengthy clinical notes, providing concise and informative summaries for healthcare professionals.

12. Voice Command Recognition using NLP:

Importance : Voice Command Recognition with NLP enables understanding and responding to spoken instructions, powering smart home controls or seamless human-computer interaction. It enhances user experience, revolutionizes accessibility, and drives innovation in voice-controlled applications, shaping a future where technology seamlessly integrates with daily life.

Description: Apply NLP techniques to recognize and interpret voice commands in the context of smart home automation or human-computer interaction, contributing to voice-controlled applications.

13. Automated Visual Inspection in Manufacturing using CNN:

Importance : Automated Visual Inspection using CNNs revolutionizes manufacturing by deploying neural networks to identify and categorize defects in electrical parts swiftly. This streamlines quality checks, reducing human error, enhancing productivity, and ensuring consistent, high-quality products, pivotal in today's precision-driven manufacturing landscape.

Description: Implement a CNN to inspect and classify defects in manufactured electrical components or circuits, enhancing the efficiency of quality control processes.

14. Predictive Maintenance for Electrical Grids (ANN):

Importance : Predictive Maintenance for Electrical Grids using Artificial Neural Networks (ANN) is crucial for preempting grid instability. Analyzing UCI data, this approach forecasts potential failures, aiding in timely interventions. By leveraging ANN, it enables real-time prediction, enhancing grid reliability and preventing potential disruptions, ensuring a more stable electrical infrastructure.

Description: Use the UCI Electrical Grid Stability Simulation Data to predict the stability of an electrical grid using an Artificial Neural Network. The model could forecast potential disruptions or failures based on real-time data.

15. Healthcare IoT Patient Classification (ANN or NLP):

Importance : Healthcare IoT Patient Classification using ANN/NLP leverages MIMIC-III for insights. ANN predicts patient outcomes, while NLP extracts valuable data from clinical notes. This project aids in personalized care, treatment optimization, and efficient healthcare delivery by harnessing advanced technology to analyze vast patient data.

Description: Use the MIMIC-III Clinical Database for a healthcare IoT project. Apply an Artificial Neural Network for patient classification based on electronic health records or explore Natural Language Processing for extracting insights from clinical notes.

16. Predictive Forecasting for Semiconductor Manufacturing (Forecasting):

Importance : Predictive forecasting in semiconductor manufacturing, using the SECOM dataset, is crucial for anticipating production hurdles and resource needs. By analyzing historical data, this model aids in predicting issues, optimizing processes, and planning resources efficiently. It enhances productivity and ensures proactive measures to tackle manufacturing challenges.

Description: Use historical data from the semiconductor manufacturing process (SECOM Manufacturing Dataset) to create a predictive forecasting model, helping in anticipating future production challenges or resource requirements.