AI/ML & Data Science Program

Goals:

- Transform the skills of the Engineering Organization through a learning program on AI/ML and Data Science skills
- Create a Community of Practice (and Expertise) in Juniper Engineering and across Juniper

Proposed Curriculum:

The level 1 curriculum focuses deeply on ensuring that the learner has a strong understanding of Data Science with some overview of machine learning application.

Name	Duration
Introduction This section provides an overview of Data Science and the main principles around AI and Machine Learning. It is a combination of video lecture and easy to follow slides.	
• Introduction to Data Science	1 hr, 48 mins
• Introduction to AI and Machine Learning	1 hr, 30 mins
Statistics and Probability These courses provide you with the basic knowledge of statistics and probability which you will need for future work in this area. The **Linear Algebra** course in the optional section is also highly recommended.	

• Statistics Foundations – 1	2 hrs, 6 mins
• Practical Statistics for Data Scientists (chapters 1-4)	1 hr, 30 mins
Python for DataScience Python has become a de facto standard for data science and should be considered as a prerequisite for any work in this area. This course goes into a lot of detail on how to use Python for Data Analysis. **Be sure to do all of the exercises associated with this course including installing Anaconda on your laptop and the associated exercises.** If you are not familiar with Python, please be sure to complete the introductory Python courses found in this curriculum - Link to Python Basics Curriculum	
• Python for Data Analysis	2 hrs, 30 mins
Data Engineering This section offers training on the basics of data engineering. It introduces concepts of the different types of data, data handling, ingestion, and transformation.	
Data Ingestion • Data Ingestion with Python (chapter 1-6) • Introduction to Kafka	 1 hr 1 hr
Data Types • Prometheus: Up and Running (chapter 1) • Time Series Data Analysis (chapters 1-4)	 1 hr 1 hr, 56 mins
Data Storage • Advanced NoSQL for Data Science	 1 hr, 56 mins

Data Transformation ● Big Data Analysis with Hadoop and Apache Spark ● Stream Processing Design Patterns with Spark	 1 hr, 1min 1 hr, 9 mins
Assignments/Practical Work This assignment needs to be complete before the learner comes to an in-person class. Students will need the Anaconda package installed that was covered in the Python for Data Science section. Be sure to complete the exercises in that class before attempting this project. Review the additional section on Pandas in the Optional courses.	
Project: ● Do a simple EDA - Exploratory Data Analysis on a given data set ● Load your answers into the teamroom	(2-3 hours)
Total Duration	**20 hrs**

Optional Training & Resources:

- Essence of Linear Algebra
- Python for Data Science – part 1
- Python for Data Science – part 2
- Machine Learning with Python Cookbook (chapters 2-5, 6-8)
- https://developers.google.com/machine-learning/data-prep
- Machine Learning with Python Cookbook (chapters 9-10)
- https://learning.oreilly.com/interactive/?classification=content-scenario - search for pandas - will provide interactive scenarios to practice
- https://www.linkedin.com/learning/amazon-web-services-data-services-2/scalable-data-solutions-on-aws?u=85418506 - Amazon Data Services - Vendor Specific

Working pages:

1. Introduction

This section provides an overview of Data Science and the main principles around AI and Machine Learning. It is a combination of video lecture and easy to follow slides.

- Introduction to Data Science (1hr 48min)
- Introduction to AI and Machine Learning (1 hr 30 mins)

2. Statistics and Probability

These courses provide you with the basic knowledge of statistics and probability which you will need for future work in this area. The **Linear Algebra** course in the optional section is also highly recommended.

- Statistics Foundations – 1 (2 hrs 6 mins)
- Practical Statistics for Data Scientists (1 hr 30 mins?) (chapter 1-4 are essential, 5-7 can be optional)

3. Python for DataScience

Python has become a de facto standard for data science and should be considered as a prerequisite for any work in this area. This course goes into a lot of detail on how to use Python for Data Analysis.

If you are not familiar with Python, please be sure to complete the introductory Python courses found in this curriculum - Link to Python Basics Curriculum

- Python for Data Analysis (2 hrs 30 mins)

4. Data Engineering

This section should cover the below:

- Data sources (Metrics, logs, events, records, graphs etc.) (prometheus)
- Data ingestion (APIs, Webhooks, databases etc. Introduction to Kafka.)
- Data storage (Relational/NoSQL data stores , Hadoop- options in public cloud - S3)
- Querying data (Athena, BigQuery etc.) BQ is Google, Athena AWS
- Transformation (Hadoop ETL, Spark etc.)
- Streaming vs Batch
- In-memory computing

This section offers training on the basics of data engineering. It introduces concepts of data handling, cleansing, modeling and visualizing data.

https://www.linkedin.com/learning/paths/become-a-data-engineer-mastering-the-concepts?u=85418506

Data Sources and Data Handling

- **Data Ingestion**
- https://www.linkedin.com/learning/data-ingestion-with-python

  [Chapters 1-6]

- Introduction to Kafka

  https://www.linkedin.com/learning/learn-apache-kafka-for-beginners

- **Data Types**
- Metrics - prometheus - https://learning.oreilly.com/library/view/prometheus-up/9781492034131/
- **Introduction to Time Series data**
  https://learning.oreilly.com/learning-paths/learning-path-machine/9781492025528/9781492025504-video318127 - Chapters 1-4

- **Data Storage**
- NO/SQL - https://www.linkedin.com/learning/advanced-nosql-for-data-science/welcome?u=85418506

- **Data Transformation**
- Introduction to Spark & Hadoop
  - https://www.linkedin.com/learning/big-data-analytics-with-hadoop-and-apache-spark/the-combined-power-of-spark-and-hadoop-distributed-file-system-hdfs?u=85418506
  - https://www.linkedin.com/learning/stream-processing-design-patterns-with-spark/streaming-with-spark?u=85418506

5. Assignments/Practical Work

(need to complete before they come to an in-person class) will need the Anaconda package installed

- Do a simple EDA - Exploratory Data Analysis on a given data set
  a. Loading dataset into pandas
  b. Categorical and Numerical data
  c. Univariate analysis and plots [Matplotlib/seaborn]
  d. Dealing with Null data
  e. Data transformations
  f. Multivariate analysis
  g. Correlations, heatmaps

h. Draw conclusions from EDA

Approach:
Option-1:

- Choice to pick a project from different data sources
    a. Example : https://www.kaggle.com/datasets
    b. Cancer mortality in US
    c. https://www.kaggle.com/sudalairajkumar/covid19-in-usa
    d.

Option-2:

- MIST throughput prediction dataset
    - Little complex for EDA. Not intuitive to understand the data
- Northstar? - Is it good for EDA??

- Business Analytics Foundations: Descriptive, Exploratory, and Explanatory Analytics
- Solution for the problem (make available when they complete step1)

6. Optional Training

Any external links are possible here
Paid content can also be linked here

- AI vs. Machine Learning vs. Deep Learning (vs. Data Science) (15 mins)
- Learning Data Science: Understanding the Basics
- Introduction to Machine Learning Problem Framing
- Stanford theoretical foundation course on ML
- Essence of Linear Algebra
- Python for Data Science – part 1
- Python for Data Science – part 2
- https://www.linkedin.com/learning/apache-flink-real-time-data-engineering/real-time-processing-and-analytics?u=85418506

- Machine Learning with Python Cookbook (chapters 2-5, 6-8)
- https://developers.google.com/machine-learning/data-prep
- Feature Engineering - https://www.linkedin.com/learning/applied-machine-learning-feature-engineering/the-secret-of-effective-machine-learning?u=85418506
- Machine Learning with Python Cookbook (chapters 9-10) https://www.linkedin.com/learning/machine-learning-with-scikit-learn/effective-machine-learning-with-scikit-learn?u=85418506
- https://www.linkedin.com/learning/amazon-web-services-data-services-2/scalable-data-solutions-on-aws?u=85418506 - Amazon Data Services

Data Visualization and Modeling

- Data Visualization: A Lesson and Listen Series
- Data Visualization: Storytelling

Curriculum Review

Reviewed by Ajit Patankar

Introduction to Data Science

Several good sections:

- Data Scientist Vs Data Engineers
-
- Good break-down between different sub-fields
-
- Data science life cycle
  - formulate a question
  - acquire clean data

Section on conducting hypothesis is not appropriate. One would typically use some library function. This material can be skipped.

Learning Data Science: Understanding the Basics
(Reviewer Ajit Patankar)

Makes it clear that is this course not for people who want to be data scientists but want to know little about it.

Little dated -- says Hadoop is most popular tool
MapReduce and Spark references-- obsolete
Includes quizzes but these are disappointing.

Drop this course from recommendation list. Mostly very high level and obsolete content. Not relevant for Juniper engineering.

Introduction to AI and Machine Learning Reviewer Ajit Patankar

First half of the course -- foundations, overview of concepts like different ML models etc. is good.
The second half is strictly Google world and can be avoided.

Introduction to Machine Learning Problem Framing Reviewer Ajit Patankar

This is a very good course. This is a series of web pages, no video or audio. It is up to point, concise and correct.

The best part is that there are short exercises in the course and we can have the participants try those exercises.

Adding the notes I discussed in the meeting on the next chapters(Saby). The ML algorithm one we can drop as we are not covering for this course.

Data Sources and Data Handling:

Introduction to Data Science:
https://www.linkedin.com/learning/introduction-to-data-science-2
Chapter (5 to 7)

Reference Text:
https://learning.oreilly.com/library/view/machine-learning-with/9781491989371/
Chapter (2-5) Optional (6-8)

Feature Engineering:

Course name: Applied Machine Learning- Feature Engineering.
https://www.linkedin.com/learning/applied-machine-learning-feature-engineering/
(Chapter 2-6) 2Hr 25Min

Reference Text:
https://learning.oreilly.com/library/view/machine-learning-with/9781491989371/
Chapter(9-10)

Machine Learning Modeling:

Course name: Python for data science (essential training part-2)
https://www.linkedin.com/learning/python-for-data-science-essential-training-part-2/

Chapter 2 to 6

Reference Text:

https://learning.oreilly.com/library/view/machine-learning-with/9781491989371/

(Chapter 11-20)

---------- End (Saby) ----------