# SEC595: Applied Data Science and Machine Learning for Cybersecurity Professionals

| 6 Day Program | 36 CPEs | Laptop Required |
|---|---|---|

## You Will Be Able To

- Apply statistical models to real world problems in meaningful ways
- Generate visualizations of your data
- Perform mathematics-based threat hunting on your network
- Understand and apply unsupervised learning/clustering methods
- Build Deep Learning Neural Networks
- Build and understand Convolutional Neural Networks
- Understand and build Genetic Search Algorithms
- Build AI anomaly detection tools
- Model information security problems in useful ways
- Build useful visualization dashboards
- Solve problems with Neural networks

Data Science, Artificial Intelligence, and Machine Learning arne't just the current buzzwords, they are fast becoming one of the primary tools in our information security arsenal. The problem is that, unless you have a degree in mathematics or data science, you're likely at the mercy of the vendors. This course completely demystifies machine learning and data science. More than 70% of the time in class is spent solving machine learning and data science problems hands-on rather than just talking about them.

Unlike other courses in this space, this course is squarely centered on solving information security problems. Where other courses tend to be at the extremes, teaching almost all theory or solving trivial problems that don't translate into the real world, this course strikes a balance. We cover only the theory and math fundamentals that you absolutely must know, and only in so far as they apply to the techniques that we then put into practice. The course progressively introduces and applies various statistic, probabilistic, or mathematic tools (in their applied form), allowing you to leave with the ability to use those tools. The hands-on projects covered were selected to provide you a broad base from which to build your own machine learning solutions.

Major topics covered include:

- Data acquisition from SQL, NoSQL document stores, web scraping, and other common sources
- Data exploration and visualization
- Descriptive statistics
- Inferential statistics and probability
- Bayesian inference
- Unsupervised learning and clustering
- Deep learning neural networks
- Autoencoders
- Loss functions
- Convolutional networks
- Embedding layers

### Author Statement

"AI and Machine Learning are everywhere. How do the vendor solutions work? Is this really black magic? I wrote this course to fill an enormous knowledge gap in our field. I believe that if you are going to use a tool, you should understand how that tool works. If you don't, you don't really know what the results mean or why you are getting them. This course provides you a crash-course in statistics, mathematics, Python, and machine learning, taking you from zero to...I'm reluctant to promise 'Hero...' Let's say competent-person-who-can-solve-real-problems-today!"

—David Hoelzer

# Section Descriptions

### SECTION 1: Data Acquisition, Cleaning, and Manipulation

This section introduces some of the terminology in the data science and machine learning fields, in addition to introducing a number of the technologies that are used as data sources. Since the first step in any data science or machine learning project is to acquire data, the balance of the day is focused on hands-on exercises to prepare the student for these tasks. The first necessary skill is the use of Python, our chosen language for this course. The only course prerequisite is a fundamental understanding of Python. If you've written even one line of Python, you are probably knowledgeable enough to get started! We will cover lists, arrays, tuples, dictionaries, comprehensions and then begin introducing the numpy variants. Following the Python refresher the course provides some theory followed immediately by hands-on exercises to give you just enough knowledge of SQL, MongoDB, and webscraping to get real work done.

**TOPICS:** Data Science; Python; SQL; NoSQL; Webscraping

### SECTION 2: Data Exploration and Statistics

This section begins with the fundamentals of statistics that matter for data science and machine learning. Following this introduction and hands-on exercises that provide practical uses for these techniques against real-world data, the course transitions to probability theory. Probability theory is an extensive field of its own. Following the introduction of some fundamentals, the course works directly toward deriving the Bayesian theorem. Building on this introduction, students then engage in a hands-on lab that builds a useful Bayesian analysis tool, upon which students will improve later in the course. The remainder of this section is translating the statistical knowledge gained into the field of signals analysis. After a discussion concerning the derivation and applications of the Fourier series, the Fast Fourier Transformation, and the Discrete Fourier Transformation, students use these tools in a real-world threat hunting activity.

**TOPICS:** Statistics; Robust Measures; Probability; Bayes Theorem and Inference; Fourier Series and Related Derivations

### Who Should Attend

- InfoSec professionals who want to understand machine learning
- Professionals desiring to apply data science principles to real-world problems
- Anyone who has tried to learn the basics but can 't figure out how to translate your problem into something that can be solved with machine learning
- Blue Team and SOC members looking to identify anomalies and perform custom threat hunting

### SECTION 3: Essentials of Machine Learning – Part I

The remaining 18+ contact hours of this course are spent learning about and immediately applying various machine learning models. After each topic is introduced and discussed, students engage in lengthy hands-on labs to develop an intuitive understanding and apply the technique to real problems. The section begins with various clustering approaches and unsupervised machine learning. The exploration begins with Support Vector Classifiers, kernel functions, and Support Vector Machines. Following this discussion and exercises, we continue the clustering theme by considering the K-Means and KNN approaches. After working through examples in just two or three dimensions, we turn our attention to methods for determining the ideal number of clusters. With this done, we finally explore high-dimensional applications and dimensionality reduction through Primary Component Analysis. The DBSCAN algorithm is covered in some depth, with application made to threat hunting and efficient SOC analysis of large scale data. The balance of this section is spent discussing Decision Trees. After a hands-on activity and discussion of the limitations of Decision Trees, we expand into Random Forests and explore hands-on how these provide better inferences in most cases. The section wraps up with a cluster-based approach to finding anomalies in user activity on a network.

**TOPICS:** Support Vector Classifiers; Support Vector Machines; Kernel Functions; Primary Component Analysis; DBSCAN; K-Means; KNN; Elbow Functions; Decision Trees; Random Forests; Anomaly Detection

### SECTION 4: Essentials of Machine Learning – Part II

The entire focus of this section is on the theory, development, and use of supervised learning approaches in the field of information security. Building on the mathematics and statistics covered in section 2, this section begins with linear regressions and ends with the application of deep learning neural networks to multi-class classification problems involving real-time network data. The material is focused on using supervised machine learning and mathematics to create predictive models. The initial discussion and exercises center around forecasting and trends analysis for anomaly detection. Following this, the majority of the material focuses on classification problems. Building on the Bayes approach used in Section 2, this section introduces deep learning neural networks and fully connected dense networks through the development of a far more accurate phishing detection network. Following this, the course explores visualization and measurement of neural network training performance, in addition to discussing overfitting, overtraining, and how to identify (and avoid!) them. The next portion of this section turns to categorical problems, during which students will build a real-time network protocol classification system. More importantly, students will implement anomaly detection in this classification system, a task typically reserved for unsupervised approaches.

**TOPICS:** Regression and Fitting; Loss and Error Functions; Vectors, Matrices, and Tensors; Fundamentals of the Perceptron; Dense Networks

### SECTION 5: Essentials of Machine Learning – Part III

This section of the course is dedicated to expanding students ' knowledge of deep learning solutions. The first half of the section is focused entirely on convolutional networks (CNNs). The class explores the application of CNNs to text classification problems, but also to predictive identification of zero-day malware. The second half of this section of the course focuses on autoencoders. The class examines what autoencoders do, why they work, how to select a latent representation, and how reconstruction loss functions work. This knowledge is then applied to creating an automatic log anomaly detection solution that does not use any signatures or human intervention to identify anomalies. Building on this, students work on the building blocks for a large-scale ensemble autoencoder for detecting network threats.

**TOPICS:** Convolutional Neural Networks; Embedding Layers; Applying CNNs to Text Problems; Autoencoders; Reconstruction Loss Measurements; Creating Ensemble Autoencoders

### SECTION 6: Essentials of Machine Learning – Part IV

The final section of this course continues discussing Convolutional Neural Networks and the application of CNNs and fully connected networks for solving regression problems. The major focus of this section is on the creation of a deep neural network using TensorFlow 's functional pattern for both testing the quality of and solving CAPTCHAs. Whether you are on a red, blue, or purple team, you will learn how to think through and use machine learning to solve what amounts to a computer vision problem and to solve it at greater than 95% accuracy! After this, we explore a different way to think about the problem that results in even greater accuracy with far less training time. The final portion of the section investigates Genetic Algorithms as they can be applied to machine learning problems.

**TOPICS:** Convolutional Neural Networks; Functional definition of Neural Networks; Deep Learning Networks with Multiple Outputs; Thinking about Machine Learning Problems; Genetic Algorithms