Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Information Technologies

Bachelor Thesis

Machine Learning and Applications of AI

Nnaka Emmanuel Chinonso

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

BACHELOR THESIS ASSIGNMENT

Chinonso Emmanuel Nnaka

Systems Engineering and Informatics
Informatics

Thesis title

Machine Learning and Applications of AI

Objectives of thesis

The main goal of the thesis is to create predictive model utilizing machine learning and compare its accuracy to a classical predictive model based on statistical analysis.
Side objectives of the thesis include:
- conduct literature reviews of selected scientific materials regarding the incorporation of machine learning into predictive modelling
- collect a prepare the testing data
- propose two similar baseline models, one with machine learning and the other without it
- conduct experimental testing of both models and compare their predictive accuracy

Methodology

The theoretical part of the thesis is based on literature review of available literature, scientific journals and online sources regarding predictive modelling, machine learning and other topics relevant for the thesis. The practical part consists of collection and preparation of testing data, proposing two baseline models, and developing them using Python and/or R programming languages. Both models will be deployed using the testing data and their predictive accuracy will be measured and compared. Final conclusions will be based on the results of both thesis parts.

The proposed extent of the thesis

40-50

Keywords

AI, Machine learning, Predictive models, Python, R

Recommended information sources

David L. Poole and Alan K. Mackworth. Python Code For Artificial Intelligence: Foundations of Computational Agents. November 2017. University of British Columbia, Vancouver.isbn: 9781107195394.

Dr. Anasse Bari, Mohamed Chaouchi and Tommy Jung. Predictive Analytics For Dummies. December 2016. John Wiley & Sons. ISBN: 978-1-119-26700-3.

John D. Kelleher, Brian Mac Namee and Aoife D'Arcy. Fundamentals of Machine Learning for Predictive Data Analytics. July 24th 2015 MIT Press.ISBN-13: 978-0262029445 ISBN-10:0262029448.

John Paul Mueller, Luca Massaron. Machine Learning For Dummies. July 2016 For Dummies. ISBN: 978-1-119-24551-3.

Michael Bowles.Machine Learning in Python: Essential Techniques for Predictive Analysis.May 2015 Wiley .ISBN: 978-1-118-96174-2

Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning:From Theory to Algorithms.2014 Cambridge University Press. ISBN 978-1-107-05713-5.

Expected date of thesis defence

2020/21 SS – FEM

The Bachelor Thesis Supervisor

Ing. Jan Pavlík

Supervising department

Department of Information Technologies

Electronic approval: 29. 7. 2020

Ing. Jiří Vaněk, Ph.D.

Head of department

Electronic approval: 19. 10. 2020

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 13. 03. 2021

Declaration

I declare that I have worked on my bachelor thesis titled "Machine Learning and Applications of AI" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on 15.03.2021 ____________________________

Acknowledgement

I would like to thank my supervisor Ing. Jan Pavlík and my parents for their advice and support during my work on this thesis and overall studies, respectively.

Machine Learning and Applications of AI

Abstract

This thesis provides a general outlook on artificial intelligence and machine learning. The main part focuses on a very useful application of machine learning known as predictive analytics. Data were first collected from the online source Kaggle, which were then analyzed and manipulated to obtain specific details using summary statistics. Predictive models were created utilizing machine learning with the use of the python programming language, and a second model was made using SAS studio, a typical statistical tool. A simple regression method was applied where the results for the weight could be predicted from the height and gender. In addition to the creation of the models, the accuracy of both models was measured using the coefficient of determination, r-square and compared to know what model predicts the weight of the dataset used more accurately.

Keywords: AI, Machine learning, Predictive models, Python, R

Table of Contents

List of Figures

List of Tables

1 Introduction

The rapid growth of AI and machine learning technologies in just about every business sector is one that cannot be denied. This growth due to the importance of artificial intelligence and machine learning has led to a lot of hype and fantasies (as portrayed in movies such as Star Wars and Star Trek) about what they are, their current as well as possible future uses. In addition, these two technologies are not only being wrongly defined but are also used interchangeably. Unmasking both technologies by proper definitions, stating the differences, similarities, and what roles they play is therefore of outmost importance.

The content begins with a general theoretical overview of AI and machine learning where related topics such as data and big data are also looked at.

The main aim of this work is seen with the use of predictive analytics in the practical section where analysis was made and compared for a dataset using a statistical approach and a Machine learning algorithm.

2 Objectives and Methodology

2.1 Objectives

The main goal of the thesis is to create predictive model utilizing machine learning and compare its accuracy to a classical predictive model based on statistical analysis.

Side objectives of the thesis include:

- conduct literature reviews of selected scientific materials regarding the incorporation of machine learning into predictive modelling.
- collect a prepare the testing data.
- propose two similar baseline models, one with machine learning and the other without it.
- conduct experimental testing of both models and compare their predictive accuracy.

2.2 Methodology

The theoretical part of the thesis is based on literature review of available literatures, scientific journals and online sources regarding predictive modelling, machine learning and other topics relevant for the thesis. The practical part consists of collection and preparation of testing data, proposing two baseline models, and developing them using Python and/or R programming languages. Both models will be deployed using the testing data and their predictive accuracy will be measured and compared. Final conclusions will be based on the results of both thesis parts.

3 Literature Review

Before diving into the development of the predictive models described earlier, it is important to look at the theoretical overview needed for better understanding of machine learning and AI.

3.1 Artificial Intelligence

Today, we have the definition of artificial intelligence (AI) as the study and design of intelligent agents which are systems capable of perceiving their environment and taking actions based on the outcome of the processing undergone. According to John McCarthy (1956) who is widely regarded as the father of artificial intelligence "AI is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable". (1)

It is as equally important to define intelligence when AI is discussed. Also, according to John McCarthy, "Intelligence is the computational part of the ability to achieve goals in the world. Varying kinds and degrees of intelligence occur in people, many animals, and some machines". (1)

As far back as ancient Greeks, there has always been the idea of a possibility of placing a mind inside a mechanical body. This is evident by the Greek mythology first mentioned at about 700 B.C about a mechanical man named Talos, constructed of bronze who acted as a guardian for the island of Crete.
Works on artificial intelligence heavily rely on mathematics and logical reasoning. Early discoveries and publications that has assisted in creating better understanding and broadening our horizon on the concept of artificial intelligence include:

- *Dissertatio de arte combinatoria* (On the Combinatorial Art) by the mathematician and philosopher Gottfried Wilhelm Leibniz (in 1666), following the works of earlier philosophers especially Ramon Llull, Leibniz proposed that all human thoughts are a combination of way simpler concepts just as words are a combination of letters.

- Thomas Bayes an 18th century mathematician developed a framework in 1763 which was to serve the purpose of reasoning about the probability of events. Bayesian inference has become a useful approach in today's world of machine learning.
- George Boole in 1854 as his publications suggests, proposed that just as solving a system of equations, logical reasoning could be performed in a similar systematic way as well. Part of his work which we know today as Boolean logical algebra plays an indispensable role in not just artificial intelligence but also other spheres of computer science.
- Leonardo Torres y Quevedo a Spanish engineer besides his automatic calculating machine, was popular in 1914 with his chess-playing automata which became the first fully automatic chess playing machine. Although not perfect, it was designed for the end game of king and rook against king.
- Francis P. Houdina released a radio-controlled driverless car in New York in the year 1925.
- The 1929 development of the first Japan built robot by Makoto Nishimura which goes by the name Gakutensoku (meaning "*learning from the laws of nature*"). With the aid of an air pressure mechanism, Gakutensoku could change its facial expression and move its head and hand as well.
- Warren S. McCulloch and Walter Pitts publication of 1943 called *A Logical Calculus of the Ideas Immanent in Nervous Activity*, where they discussed simplified artificial "neurons" and their role in simple logical functions. The content of this publication as time went on, became useful in modern day computer-based neural networks.
- Claude Shannon an American Mathematician and Cryptographer published the first released article on a chess playing computer program in 1950.

However, the widely accepted true birth of AI as we know it today began with Alan Turing's publication "*Computing Machinery and Intelligence*" of 1950. In his paper, Turin explored the idea of how to determine whether machines can think. This was done by the imitation game consisting of three players A, B and C. Where player A is a computer, B is a human and an interrogator C. The interrogator C must be convinced by A and B that they are humans to win. If C cannot distinguish the human from the computer on a consistent basis, the computer B wins.

3.1.1 Types of Artificial Intelligence

Despite the mix-up in terms that arises whenever the type of AI is involved, the most general form of grouping of the types of AI are usually based on its capabilities (type-1) and its functionality (type-2).

Based on Capabilities (AI Type-1)

This grouping of AI types focuses on what that specific AI can do. It is divided into three:

- **Weak (Narrow) AI** which is the simplest and commonly available type of AI. It basically represents a huge chunk of all available AI today. This are AI systems that can only perform a specific task which they have been trained for. Good examples include AI in self-driven cars, AI for chess playing, AI for image and speech recognition.
- **General AI** which can perform intellectual tasks as efficient as humans can carry them out. Such AI agents can learn, perceive, understand and function completely like a human being. Currently, there are no such systems that exist. Although research involving this type of AI are ongoing, it will take a lot of time and effort before such systems are developed.
- **Strong AI** which are currently completely hypothetical concepts of a system capable of surpassing human intelligence and can perform any task better than humans. Such a system will have other added abilities to that of general AI such as making faster judgements, plans and communicates by its own.

Based on Functionality (AI Type-2)

This grouping of AI types is based on the system's likeness to the human mind. They are divided into four:

- **Reactive Machines** which are the oldest types of AI systems based on functionality. They have very limited capability and do not have a memory-based functionality. Therefore, they do not store past experiences for future actions and as such focus on only current scenarios. A well-known example fitting this description would be the IBM’s Deep Blue system which was able to beat Russian chess Grandmaster Garry Kasparov in 1997.

- **Limited Memory** which in addition to having the capabilities of reactive machines, are also capable of learning from past data before making decisions. The downside of such AI systems is that they store data for a short time only. A major example of this type of system are self-driving cars as they can store recent information about the environment when on the road needed to properly navigate successfully.
- **Theory of Mind** though still categorized as a hypothetical AI system, progress has been made with robots such as the latest humanoids. Such AI systems should be capable of understanding more complex things such as human emotions, beliefs and be able to interact socially like humans. Basically, what constitutes this type of AI is a decision-making ability equal to that of a human mind. Although, certain AI systems already possess such human capabilities in a way, such as voice assistants, they still are not capable of showing complete human emotional capacity or sounding and behaving like a human being completely.
- **Self-awareness** which is at the pinnacle in the classification of AI systems based on functionality. These systems will be capable of possessing their own consciousness, be self-aware and understand its own feelings. Such AI systems are expected to be smarter than humans. This is a futuristic idea of producing systems that not only possess the characteristics of Theory of Mind systems but takes it further with the possession of self-guided thoughts and reactions to these thoughts.

3.1.2 Applications and Uses of AI

As earlier stated, AI has encroached into just about every sector one can think of. AI at some level is being used in just about any system and application today. These technologies seamlessly carry out algorithm commands and work so well that one might not be able to know it exists. Here are fields in which the applications of AI are very evident:

AI in Healthcare

Applications of AI in medicine has expanded over the years, especially the past few decades. This has created an evolution into much more personalized medicine that tailors specifically the needs of individual patients. A very common usage of AI in medicine is the use of predictive models (which we will take a deeper look at later) for the diagnoses of diseases and making prediction of therapeutic response which can potentially be used in having such illnesses prevented later in the future. AI generally serves to improve the accuracy of these diagnosis leading to improved efficiency in the workflow and operations of hospitals.

In 2017, a huge impact of AI in medicine was made for a neurodegenerative disease research in which IBM Watson was used to identify additional RNA-binding proteins altered in amyotrophic lateral sclerosis. Another area of huge impact is in gastroenterology where AI-assisted endoscopy with computer-aided diagnosis (CAD) can be used for the detection, differentiation, and characterization of neoplastic and non-neoplastic colon polyps.

AI in Gaming

AI can be used for creating spectacular games that require game characters and functions with some level of intelligence. The most common ones are in strategic games like chess where the system needs to think of large number of possibilities and make the best decision. Examples of this includes IBM's Deep blue and Google's AlphaGo. Today, advancements have been made in AI in the gaming area as it plays important roles in creating responsive and adaptive video game experience which are generated via non-player characters (NPC), that thinks and acts intelligently as though they were controlled by a human.

AI in Banking and Finance

In the past years, the finance industry has been applying automation, chat boxes and algorithm trading amongst other implementations. The benefits of having AI in finance is immense as it makes work very easy in areas such as risk assessment, management, decisions for trading and general finance management. These implementations have overall changed the banking and finance field over the last decade. A main and obvious impact of AI in banking and finance is in modernized online banking and the security that accompanies it such as for fraud detection when and if needed.

AI in Complex Analysis and Machine Efficiency

AI has a role in controlling a machine in such a way as to obtain maximum efficiency. AI does this by controlling the use of resources so that the system does not overshoot.

For complex analysis which cannot be done by humans or done less efficiently, AI helps in analyzing such data and arriving at even more accurate results and conclusions.

AI in Automation

"A problem with some types of automation today is that an unexpected event, such as an object in the wrong place can actually cause the automation to stop" (2)

Just about all forms of automation is improved by the addition of AI to handle such unexpected changes that could potentially have had it stopped. Such an addition allows the automation to continue after encountering an unexpected event as though nothing happened which would increase its efficiency.

AI in Robotics

Robots being programmable machines that carry out physical processes therefore will be requiring a touch of AI since they are not human operated. This is probably the application with the most exciting and remarkable role that AI has played. With the help of AI, we can create intelligent robots that can perform tasks on their own experiences without been extensively pre-programmed to do so. They therefore are capable of learning from their own experience. Today's humanoids such as Sophia, the world's first humanoid robot is a great example of such. Modern-day humanoids are developed to carry out different human tasks and occupy different roles in employment sectors such as receptionists, front desk officers and personal assistants.

3.2 Machine Learning

Machine learning is an enticing and alluring field of study. It is important pointing out that machine learning is a subset of AI just as another term named deep learning (which would not be spoken of extensively) is a subset of machine learning. The field of machine learning has its algorithms revolving a lot around logic and mathematics.

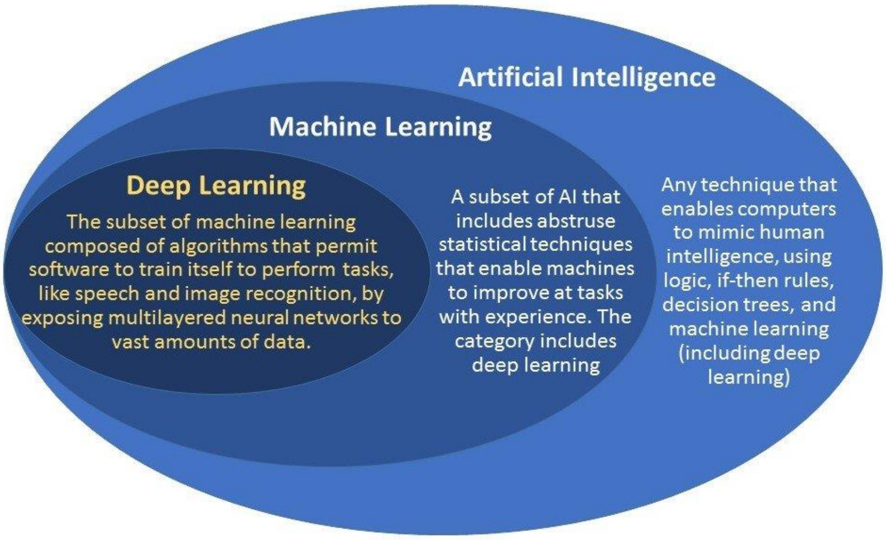

Figure 1 - Differentiating AI and Machine Learning

Source: https://geospatialmedia.s3.amazonaws.com

Machine learning is generally defined as an automated process that extracts patterns from data. Tom Mitchell (1997) describes machine learning as “the study of computer algorithms that allow computers to automatically improve through experience” (3). Basically, when we mention machine learning, we are speaking about the extraction of knowledge from data thereby enabling such systems to learn and function by themselves from their own past activities.

The term machine learning was first coined by Arthur Samuel in 1952. According to Arthur Samuel (1959), machine learning is a “field of study that gives computers the ability to learn without being explicitly programmed” (4).

3.2.1 How Machine Learning Works

Machine learning works to optimize models by summarizing a huge chunk of data in a meaningful way such that it can predict or determine what the appropriate response should be even when it receives an input that it has not seen before. The more accurately the predictive model can predict the response for new inputs, we say the better the model has learned from the inputs of the dataset provided. Therefore, an obvious criterion for making a machine learning search work is to look for models which are consistent with the data used. The fitting process of the predictive model to the data is done by algorithms and it is called training of the model.

“Machine learning algorithms work by searching through a set of possible predictive models for the best which captures the relationship between the descriptive features and the target feature in a dataset” (5).

The main central idea behind machine learning is that reality and its complexity can be expressed in terms of mathematical functions that the created algorithm does not know in advance but can guess after processing relevant data (in the form of input/output pairs). Basically, every machine learning algorithm is built upon a modifiable mathematics function.

“A successful learner should be able to progress from individual examples to broader generalization. This is also referred to as inductive reasoning or inductive inference” (6).

3.2.2 Types / Techniques of Machine Learning

Learning for machine learning systems comes in many ways, depending on the algorithm and its objective. Algorithms can be categorized by the learning style used. Here are the three main machine learning types:

- Supervised machine learning
- Unsupervised machine learning
- Reinforced machine learning

Supervised machine learning

This is the most popular used type of machine learning. Supervised learning is exciting because it works in a much similar way as how humans learn under the supervision of a teacher where the teacher provides examples for the student to memorize and learn by deriving general rules from examples provided.

“In this type of machine-learning system, the data that you feed into the algorithm, with the desired solution, are referred to as labels” (7). These labels are needed to predict the appropriate response correctly when new examples are inquired of later after being fully trained. Correct predictions occur over time as the algorithm learns to approximate the nature of the relationship between examples and their target labels. Since we deal with quantitative data, it is important to distinguish between regression problems having numeric value target and classification problems whose target is a qualitative variable. Such qualitative data are converted into quantitative data.

Supervised learning is generally described as task-oriented for the above reasons. It is highly focused on a singular task, feeding more and more examples to the algorithm until it can accurately perform the task. Typically, supervised learning is used in areas such as advertisement popularity/selection, face recognition, spam classification and numeric value prediction of a specific type of product or item after given a set of features called predictors.

Two important techniques are used in supervised learning: Linear regression and classification techniques.

- **Linear regression (Ordinary least squares):** Besides been the simplest, this is amongst the earliest learning techniques that is still been widely used. It is typical for predicting, forecasting, and finding relationships between quantitative data. In linear regression, the number of independent variables is one and there is a linear relationship between an independent (x) variable and the dependent (y) variable. In healthcare for example, this type of supervised learning can be used for determining if there is a linear relationship between a particular radiation therapy and tumor sizes.
- **Classification techniques:** This approach is used to forecast group membership for data instances. Classification technique is a predictive model that approximates a mapping function from input variables to identify discrete output variables (labels or variables), which can be performed on both structured and unstructured data. The widely used classification techniques include:
    - K-nearest neighbors
    - Logistic regression
    - Neural networks
    - Logistic regression
    - Neural networks
    - Support vector machines
    - Decision trees and random forests

It should be stated that one of the most significant difference between these two techniques is that linear regression can predict a continuous quantity while classification predicts discrete labels.

Unsupervised machine learning

As the name suggests, this is the very opposite of supervised machine learning as there is no supervision because of the absence of a given desired output. In unsupervised learning, conclusions are drawn from datasets consisting of input data without labelled responses. From there, the algorithm can learn to group, cluster, and organize the data in such a way that is meaningful. Since unsupervised learning is completely based upon the data and its properties, we say it is data driven. The outcomes from an unsupervised learning task are controlled by the data and the way it is formatted.

A good example of unsupervised learning can be a recommender system which functions by grouping a large database of every research paper ever published in such a way that the writer is always aware of the current progression within a particular domain of research. As the writer writes the research work and make notes, the algorithm makes suggestions to the writer about works which he or she might wish to cite. This is also used by entertainment media such as YouTube or Spotify for their video recommendations.

The most important unsupervised algorithms are further grouped into mainly clustering and association method as well as visualization and dimensionality reduction.

- **Clustering:** This is a common technique for statistical data analysis in many fields as well as been an important concept in unsupervised learning. Basically, we are dividing the set of observations into subsets called clusters using appropriate clustering algorithms. Applying clustering algorithms, you can find and modify your clusters into however number of groups desired. This technique is used extensively in field image processing and vector quantization for data reduction, compression, and summarization. It also generally serves as an intermediate step in data mining due to its summarization capability.

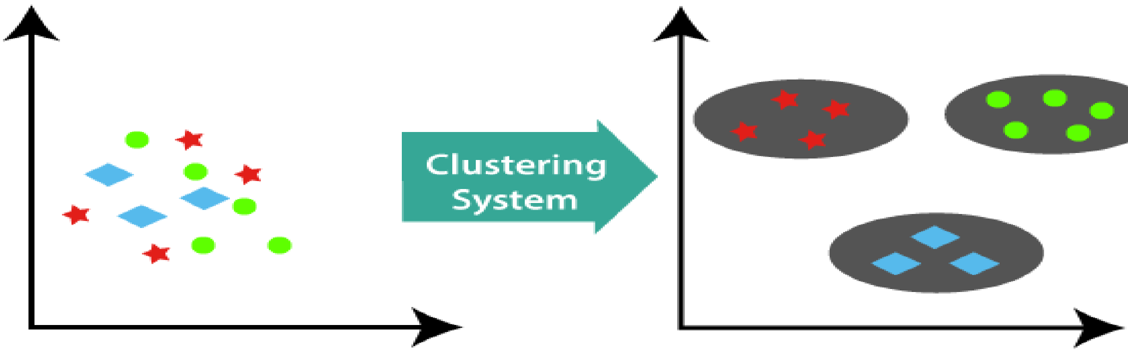

Figure 2 - Clustering Model

Source: https://www.tutorialandexample.com/ml-clustering-algorithm/

There are different types of clustering methods which can be used, such as the K-means clustering algorithm which helps to find local maxima in each iteration, the Hierarchical clustering which builds a hierarchy of clusters beginning with all the data, which is assigned to a cluster of their own where the two closest clusters are merged into the same cluster and the algorithm ends when there is only one cluster left. Other clustering types include the K nearest neighbors (K-NN), the Singular value decomposition and the Independent component analysis.

- **Association rule learning:** This type of unsupervised machine learning checks for the dependency of one data item on another data item. It is all about discovering interesting relationships between variables in large databases. Generally, this method is employed in market basket analysis and web usage mining. Examples of where this technique can be used include a subgroup of genetic diseased patients grouped by their gene expression results by a hospital as well as groups of online shoppers based on their purchasing history. The various algorithms used to implement association rule learning include the Apriori algorithm used for mining familiar item sets and relevant association rules for effective market basket analysis, the Equivalence Class Clustering & bottom-up Lattice Traversal (ECLAT) algorithm which is applied to achieve itemset mining. The ECLAT algorithm is more efficient and scalable when compared to the Apriori algorithm as it imitates the vertical manner of the Depth-First Search of a graph as opposed to the horizontal Breadth-First approach of the Apriori algorithm.

Finally, Visualization and dimensionality reduction techniques which are often not included as an unsupervised algorithm involves the task of deriving new artificial features that is smaller than the original set while still retaining most of the variance of the original data. It comprises of techniques that reduce the number of input variables in the dataset such as the Principal Component Analysis (PCA) which reduces the dimension of the data without much loss of information by finding a few orthogonal linear combinations (the principal components) of the original variables with the largest variance; the Kernel PCA which applies a kernel function to project dataset into a higher; and the t-Distributed Stochastic Neighbor Embedding(t-SNE) which is particularly suited for very high dimensional data is a non-linear dimensionality reduction technique that functions by compressing small distances, therefore bringing close neighbors together.

Reinforcement learning

This form of reinforced learning is different from supervised and unsupervised learning as it neither deals with the presence of a label nor its absence. “Reinforced learning is connected to applications for which the algorithm must make decisions (so the product is prescriptive, not just descriptive, as in unsupervised learning), and the decisions bear consequences. In the human world, it is just like learning by trial and error” (2). For this learning type, an agent (AI system) observes the environment, take actions (usually containing mistakes) and then get a reward for that action. We must provide a positive signal which associates with a good or expected action and a negative signal associating with a bad action. Over time, the algorithm learns to make less mistakes due to the rewards it receives. With this type of learning, the AI system learns by itself.

Reinforcement learning is used in many sectors such as video games where such games are made richer and challenging by applying variety of ways ranging from non-player character (NPC) control to procedural content generation (PCG). Google’s Alpha Zero and AlphaGo fall into this category. Another major application example of reinforcement learning is in robotics where robots learn to perform a certain activity by continuously learning from their mistakes.

Semi-supervised learning

The major basic disadvantage of supervised learning algorithm is that datasets must be hand-labelled while that of unsupervised learning is that there are limitations to the applications it can be used for. Although not usually listed as a machine learning type, semi-supervised learning is a sort of hybrid of the earlier mentioned learning types as it typically comprises of a very small amount of labelled data and a very large amount of unlabeled data. Basically, this involves clustering similar data using an unsupervised learning algorithm and then use the existing labelled data to label the rest of the unlabeled data.

Applications of this learning type algorithm is found in speech analysis and all forms of classifications of very large data sizes such as protein sequence classification.

3.3 Data and Big Data

Since datasets will be analyzed and manipulated, it makes absolute sense having a brief look at certain key points associated with data as well as big data.

3.3.1 Data vs Big Data

Data is simply a collection of processed information which are managed by the computer through applications that perform tasks using various types of algorithms. These algorithms carry out a systematic set of operations on the data of which we have four basic types which are to Create, Read, Update and Delete (CRUD).

As datasets become larger, the computer can do more work using the algorithms found in such applications. This is where big data comes in. According to Gartner, "Big data is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making" (8). Although big data implies a lot of data, it is not only defined by this as it includes complexity and depth. When we speak about big data, we are talking about a big data source describing something in enough details that one can begin working with that data to solve problems. Big data can the seen as far back as in 1663 when John Graunt who is credited with being the first person to use statistical data analysis for his statistical study of the bubonic plague which was ravaging Europe at that time. Also, during world war II, the British invented a machine for recognizing patterns that scanned 5000 characters per second of messages intercepted from the Germans. An example of a more modern usage of big data is in self-driving cars as such car systems must take into consideration a lot of information such as the car's hardware, position in space, the decisions of humans on the road, other road and environmental conditions as well as other vehicles on the road. All of which must be taken into consideration in real time, which is difficult. This example of big data showing how much of an enormous amount of data needs to be processed in split seconds is an evidence of how challenging it could be controlling and managing big data. Part of the problem faced now is determining how to control big data.

3.3.2 Data Mining

Without a doubt, big data is one of the most valuable assets of any organization especially data-driven companies such as Facebook as it is full of many hidden precious patterns that needs to be dug out. The process of having such patterns dug out is what we refer to as data mining. Data mining is the discovery of hidden and potentially useful patterns of data from big data by the usage of machine learning and sophisticated algorithms as mining tools.

3.3.3 Data Mining Techniques and Tools

Generally, data mining and business knowledge merges together to produce what is known as predictive analytics (which in most cases employ machine learning algorithms) and as such possesses similarities with machine learning concepts.

Techniques in mining data includes Classification used to retrieve important and relevant information about data and metadata by first classifying them into different classes; the Clustering technique which identifies data like each other, thereby helping to understand the differences and similarities between the data; Regression which mines data by identifying and analyzing relationships between variables. It is used in identifying the likelihood of a specific variable, given the presence of other variable(s); Association Rules that help find the association between two or more items, thereby discovering hidden patterns; an Outer detection technique which is employed for observing data items in the dataset which do not match an expected pattern. It plays an important role in areas such as intrusion and fraud detection; Sequential Patterns which helps to discover similar patterns in transaction data for certain period; the Prediction technique which uses a combination of other data mining techniques like sequential patterns, clustering, classification, and others to analyze past events or instances in a right sequence for predicting future events.

The tools employed for data mining are also like those used in predictive analysis with machine learning as they include programming languages such as R, Python, SQL, JULIA, Java, and MATLAB as well as application software such as Oracle Data Mining, IBM SPSS Modeler, Orange, and Tableau.

3.3.4 Applications of Data Mining

- Data mining help organizations get concrete knowledge-based information.
- They help organizations make profitable adjustments in operation and production.
- It serves as a cost-effective and efficient solution compared to other statistical data applications.
- Assists with decision-making processes.
- It is a speedy process which makes it easy for users to analyze huge amounts of data in less time.
- It is applied in areas such as communication and insurance for predicting customer behavior and to promote sellable offers to new and existing customers.

3.3.5 Types of Big Data

Structured Big Data

As the name does suggest, this data type deals with big data that is already well stored in the database in an ordered manner. It accounts for about 20% of the total existing data and used mostly in programming activities. These are highly organized information that can be readily and seamlessly stored and accessed from a database by search engine algorithms. Structured data can be obtained via two sources- machines and humans.

Unstructured Big Data

While structured data are organized in the database, unstructured data is the exact opposite as they have no clear format in storage. Therefore, making it very difficult and time-consuming to process and analyze such data. Statistics show that most of the data encountered (80%) belong to this category.

Unstructured data is further divided into – Captured data and User-generated data. Where captured data Is data based on user's behavior and user-generated data is created when the user inputs data into the internet such as tweets, like, comments on other social media platforms.

Semi-Structured Big Data

As the name clearly suggests, this is a big data type containing features of the above-mentioned types. To be specific, they are not classified under a particular repository like in the structured data, but they contain organizational properties such as vital information or tags that separates individual elements within the data making it easier to process.

3.4 Predictive Analytics

Predictive analytics, a very important subfield of data analytics is the art of building and using models that make predictions by patterns extracted from historical data. This involves the use of mathematical processes to predict future events and outcomes by analyzing patterns and correlations (found in relevant historical data) that are likely to forecast future results. According to Kimberly Nevala, Predictive analytics are " 'Simple forecasting models in which past performance' has been used to determine future outcomes" (9).

It is important to point out that predictive analytics is a complex process embodying other linked processes such as Data collection which has to do with preparing relevant data from multiple sources; Data analysis entailing the process of inspecting, cleaning and transforming data with the objective of discovering useful information; Statistical analysis which enables the validation of hypothesis and testing them using statistical predictive models; Modelling and the deployment of the models as well as its monitoring to review the model's performance to ensure it provides the reasonable results.

3.4.1 Machine Learning for Predictive Analytics

Machine learning is used to enable a program accurately analyze data, point out correlations and make use of these correlations as an insight to solve problems.

"Machine learning can categorize new and upcoming unknowns, learn from them based on its previous processing of the data, and get better at incorporating them into the known data" (10). Therefore, we can say with an enormous amount of certainty that machine learning is the core of modern-day predictive analytics.

Machine learning is considered as a superset of predictive analytics.

3.4.2 Applications of Predictive Analytics

Predictive analytics plays a vital role in making predictions about future events in just about every sector thought of. Popular applications of predictive analytics include customer relationship management (CRM) where objectives such as marketing campaigns, sales and profitable customer services throughout the customer’s life cycle are achieved; Risk assessment in health care, where patients who are at risk of developing certain illnesses such as diabetes can be determined. It can also be applied to other areas where risks are involved such as in banking for issuing of loans; Fraud detection; Price prediction for businesses as prices need to be adjusted to maximize returns when faced with changes such as seasonal change and shifting customer demands; Direct marketing where the most effective combination of product versions, marketing material and timing that should be used to target a larger group of customers.

3.4.3 Predictive Analytics Tools

It is very important differentiating the different roles different software tools play for predictive analytics. Generally, the two programming platforms mostly used and regarded as data analytics languages are Python and R. Both having tremendous advantages when it comes to predictive analytics (or data analytics at large) with python providing access to a huge array of libraries that can do just about anything with data and R for its ease of use as well as its heavy leaning into statistical computation.

There are also three commonly used precursor statistical software that perform statistical analysis as it is a process needed for predictive analytics. The three commonly used ones include the relatively more difficult to use SAS studio, the easier to use IBM’s SPSS and Stata.

Other frameworks and tools include the deep learning framework Apache Singa, Apache Spark MLlib which is a machine learning library, the framework Google TensorFlow, the library Oxdata H2O which can directly access any Hadoop Distributed File Store (HDFS) using Java, Python and R. The very old library Shogun can be used as well as it covers a broader range of programming languages.

3.4.4 Building Machine Learning Predictive Models

To build a machine learning predictive application, it is important to note that supervised machine learning is used. As already mentioned earlier, this machine learning type automatically learn of the relationship between a set of descriptive and target features based on examples received during its training.

Generally, creating a successful machine learning predictive model depends heavily on selecting the ideal machine learning algorithm as well as using a standard process to manage the entire project. The Cross Industrial Standard Process for Data Mining (CRISP-DM) which was conceived in 1996 is amongst such standard process.

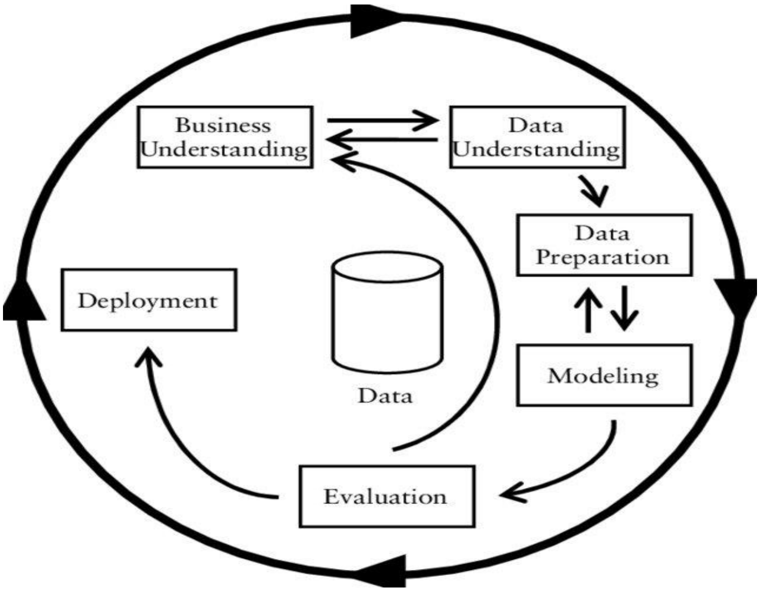

Figure 3 - CRISP-DM Process model

Source: https://www.researchgate.net

The above diagram of the CRISP-DM process shows the six key phases and indicates the important relationships between them. Predictive analytics projects must start out with a well-structured purpose which is to be understood by the data analyst. Once this purpose has been understood, the understanding of the different data sources and the kinds of data it contains must be understood as well. This is then followed by the data preparation which entails organizing the available data in a special structured form called analytics base table (ABT). The modelling process follows as different machine learning algorithms are used to build various predictive models from which the best models for that specific task are selected and evaluated. The evaluation of models to know if they will be able to make accurate predictions are done just before they are deployed for use which is the final step in the CRISP-DM process. The deployment phase entails all work done for the successful integration of the machine learning predictive model into the processes within an organization.

4 Practical Part

4.1 General Overview

For the practical part, a very important application of AI in data science in today's business world for making vital business decisions known as predictive analytics will be conducted. This chapter gives a detailed overview of the vital steps taken to obtain results for our predictive models. The experimental results obtained as well as the overall calculated accuracy of selected results will be shown and discussed in chapter 5.

Two predictive models will be created. One with a machine learning algorithm which we will call model A and the other with a more traditional statistical analysis tool which we will refer to as model B. For the analysis with both models, we will be making use of a dataset from Kaggle (https://www.kaggle.com/mustafaali96/weight-height) consisting of 3 columns (variables) of weight in pounds (lbs.) as the dependent variable, height in inches & gender (m/f) as independent variables and 10,000 rows (entries). All values will be rounded up to at least 4-decimal places for manually calculated results.

For both predictive models, we shall be applying linear regression which is a supervised learning algorithm under machine learning classifications. Generally, in this analysis we will be having the predicted dependent variable weight on the y-axis and the independent variable height on the x-axis with dots to represent the values for the height and corresponding weight for each entry and a regression line drawn across which tells us the expected value for every height-weight combination from which we can get how far off they are from the actual values.

The purpose of our predictive models would be finding the best fit line for our available data to minimize the total sum of errors.

4.2 Model A: Machine Learning Predictive Model

The machine learning model was created with the python programming language using the IDE PyCharm. Several libraries and methods were applied to make the analysis work easier and faster. Libraries including pandas, NumPy and Scikit-learn as well as functions and methods such as dataset.info(), LabelEncoder() and LinearRegression() amongst other methods and functions were used.

Steps taken to have this machine learning predictive model executed are shown as follows:

- First and foremost, the python programming language package called Pandas had to be imported. This library with the help of one of its functions ".read_csv()", was used to read the csv file which contain the data to be worked on. Besides the importation of various file types, this library possesses various methods which generally makes data filtering and a host of other data manipulations possible and convenient. Some of such methods and functions are shown in subsequent steps.

- This was then followed by a general analysis of the read data by the usage of the methods ".info()" and ".describe()" which gave out results such as the names of the variables in the dataset, range and count as well as statistical information such as the mean, standard deviation, various percentiles, maximum and minimum values…in addition to the methods already mentioned, the ".isnull().sum()" method was also used to confirm that there are no null or empty entries where the outputs of 0 in the 3 variables for all data entries was obtained.

- Since we are dealing with quantitative data, a slight challenge was faced as the gender variable is in qualitative data format (Male or Female). This was solved by replacing females and males for each data entry with 0 and 1 using the ". replace ()" method.

  Making a recall from chapter 3, when creating a ML algorithm, the dataset to be used is split into a training set which is the data from which our predictive model learns from and a smaller test set which as the name suggests is the dataset in which the algorithm is tested with to measure its performance. The Sklearn function "train_test_split" which splits datasets randomly, was used to split the dataset into two parts with the test size made 20% of the entire dataset by assigning it the value 0.2.

- "LinearRegression" was imported and another Sklearn method ".fit" was used which is equivalent to training. These were used to train our model in which when called on the rest of the training dataset (80% of entire data), the best representative function for the data points is estimated and the linear regression line created.

- The already created test dataset (20% of entire data) was used in making our prediction based on the trained data model. The prediction of the dependent variable weight for the corresponding independent variable height considering gender as well was achieved with the help of the ".predict" method.
- The weight of any random value can now be tested. The gender is also included by the usage of 0 or 1 depending on what gender the tested value is categorized under. This was done by assigning a name to the now created ".predict" method and inputting the appropriate values to be queried for.
- Finally, we have a very important final step which is the model accuracy of our result which tells us how closely the data are to the regression line created by our ML model. Although there are a couple of other statistical tests such as the mean squared error that could be used, I found the r-squared test most appropriate to be applied here. This statistical test used as well as the result produced will be further looked at in the subsequent chapter. For this ML model, metrics was imported from sklearn and the method ".r2_score" was used.

4.3 Model B: Statistical Predictive Model

Applying the classical statistical analysis tool SAS studio software, the regression formula to be used was created along with useful values to be inputted into the formula such as population slope coefficient and the y-intercept were obtained.

It is important to state that this predictive model can be created by either clicking on already made SAS studio tools for whatever processes/analysis is to be executed (where the generated code created can be viewed) or can be made by writing the appropriate statements and syntax for making the execution possible. For both approaches, similar results were obtained.

Below is an overview of the steps taken, accompanied with figures for emphasis and clarification:

- The FILENAME function was used to assign a fileref to the external csv file to be made use of for the data analysis and manipulation.

- For the now assigned external csv file, the PROC IMPORT statement was then applied to import this external file to a SAS data set. The syntax of the PROC IMPORT statement includes functions such as the DATAFILE which contains the complete file path and filename which was simply done by assigning it the name of the already assigned file path, DBMS which is needed for specifying the file type where in our case it is the CSV identifier, the OUT and GETNAMES functions were also included.

- The PROC CONTENT statement was also used for listing the contents of the SAS data sets. It is important to state that the statements, the “RUN;” command is required to execute these statements.

```
FILENAME REFFILE '/folders/myfolders/ML_PROJECT/ml_weight_height_sex.csv';

PROC IMPORT DATAFILE=REFFILE
    DBMS=CSV
    OUT=REFFILE.ml_whg_dataset;
    GETNAMES=YES;
RUN;

PROC CONTENTS DATA=REFFILE.ml_whg_dataset; RUN;


%web_open_table(REFFILE.ml_whg_dataset);
```

Figure 4 - Importing dataset in SAS studio

Source: SAS studio

- After the successful importation of the dataset for the analysis, general analysis of the dataset is done by the usage of summary statistics under tasks of tasks and utilities. This is done to get the basic statistical details of the dataset to be sure there are no errors in the dataset. The dataset can be further grouped or classified by the gender to get the values of the summary statis for males and females separately. This can be done by selecting gender for the “group analysis by” option.

- The statistical model was then created by selecting “Linear Regression” from the Linear models’ option which is still under tasks of the tasks and utilities section. In this stage, there was a selection of appropriate variables under the various options such as the dependent variable as weight, the classification variable as gender, the continuous variable as height, creating the model effects with the intercept and the appropriate variables with every other factor left in default. Upon clicking on the run button and selecting the code option, the code for this procedure was generated which could be used directly as well.

It is important to note that the model could be further separated for both gender by selecting gender under the "Group analysis by" option.

```
proc glmselect data=REFFILE.ML_WHG_DATASET
        outdesign(addinputvars)=Work.reg_design;
    class Gender / param=glm;
    model Weight=Height Gender / showpvalues selection=none;
run;

proc reg data=Work.reg_design alpha=0.05 plots(only)=(diagnostics residuals
        observedbypredicted);
    where Gender is not missing;
    ods select DiagnosticsPanel ResidualPlot ObservedByPredicted;
    model Weight=&_GLSMOD /;
    run;
quit;

proc delete data=Work.reg_design;
run;
```

Figure 5 - Regression model in SAS studio

Source: SAS studio

5 Results and Discussion

5.1 General Results

The results obtained from both predictive models are shown, computed, and analyzed. This includes not just the results for the expected weights of selected heights and gender but also the degree by which the model is accurate which will be given by the r square (R^2) test as it tells us how close the data are to the fitted regression line of our models. In addition, these randomly selected result's accuracy will be manually checked also by using r-square. Since r-square gives the percentage variation in the dependent y variable explained by the dependent x variable, it serves as an indicator for how good a model fits for the observed parameters. Generally, higher R^2 percentages indicate a better goodness of fit.

For the ML predictive model using Python, we obtained our results directly after inputting the gender using 0 or 1 and the height which we intend predicting the weight for whereas in that of the statistical model, only the result of values is obtained in which we must put into the regression formula to get the result of the predicted weight.

Apart from the actual results, a couple of descriptive statistical measures such as the mean, median, standard deviation and percentiles were also derived to provide more information about our data as well as its accuracy when the results of both models are compared.

Measures of descriptive statistics (shown in table 1) are the same for both models.

	Height (inches)	**Weight (lbs)**
Number of observations	10000	10000
Mean	66.3676	161.4404
Standard deviation	3.8475	32.1084
Minimum	54.2631	64.7001
Lower Quartile (25%)	63.5053	135.8170
Median (50%)	66.3181	161.2129
Upper Quartile (75%)	69.1749	187.1754
Maximum	78.9987	269.9897

Table 1 - Descriptive statistics for ML and SAS predictive models

Source: Own work

This similarity in descriptive statistics is a good indication that both the ML and statistical models do correlate.

Weight results obtained for both models for ten (10) randomly selected height values from both genders as well as the actual results from the dataset are shown:

Key

Gender- (0 = Female; 1 = Male)

X- Height (inches)

Y- Actual weight from original dataset (lbs)

$Y^{\wedge}$- Predicted weight from ML model (lbs)

Y”- Predicted weight from Statistical model (lbs)

Gender	**X(inches)**	**Y(lbs)**	**$Y^{\wedge}$(lbs)**	**Y”(lbs)**
0	61.9152	123.9587	125.1912	125.1126
0	64.1073	138.7307	138.2619	138.2523
0	65.5448	147.9932	146.8331	146.8688
0	67.1198	156.5384	156.2242	156.3095
0	69.8941	172.1216	172.7663	172.9390
1	63.1786	153.6068	152.1482	152.1594
1	64.4579	158.9167	159.7762	159.7863
1	66.2522	170.1095	170.4749	170.4836
1	71.1630	199.4440	199.7561	199.7608
1	74.1010	217.1876	217.2742	217.2765

Table 2 – Sample datapoints – height, weight, and both predictions

Source: Own work

5.2 Model A Results

Apart from similar results with the SAS model already mentioned above, other results not mentioned for the ML model includes the outcome of the "**.isnull().sum()**" python method which resulted in 0 for all three variables of gender, height and weight. This implies that there are no null values in any of the entries of the dataset.

Also, using the python method "**.r2_score**" resulted in an r-square value of 0.9053 which represent 90.53% for the coefficient of determinant.

5.3 Model B Results

Unlike the ML model, results for the predicted weights are not obtained directly but relevant values obtained from SAS studio will have to be inserted into the linear regression formula to have the needed results of the weight obtained. These relevant values obtained from SAS studio are the values for intercept and height parameters which were obtained separately for both genders.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-246.01327	3.35616	-73.30	<.0001
Height	1	5.99405	0.05263	113.88	<.0001

Figure 6 - Values of parameter estimates for females in SAS studio

Source: SAS studio

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-224.49884	3.41085	-65.82	<.0001
Height	1	5.96177	0.04937	120.75	<.0001

Figure 7 - Values of parameter estimates for males in SAS studio

Source: SAS studio

Results obtained are summarized in the table below for better clarification.

GENDER	Y-INTERCEPT	HEIGHT
MALES	-224.4988	5.9618
FEMALES	-246.0133	5.9941

Table 3 - Values of parameter estimates for both genders obtained in SAS studio

Source: Own work

The results obtained from SAS studio for the statistical model as shown in table 5-4 is inputted into the equation for linear regression. This equation is given as;

$Y''_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Where:

Y''_i - Dependent Variable (Weight)

β_0 - Population Y-intercept

β_1 - Population Slope Coefficient

X_i - Independent Variable (Height)

ε_i - Random Error (which will be ignored for its relative insignificance as well as for the sake of easier computation)

This resulted in obtaining the weight (Y”) variable shown in table 2

Results of the **r-square** value obtained were also shown by SAS studio for both genders.

Root MSE	10.01409
Dependent Mean	161.44036
R-Square	0.9027
Adj R-Sq	0.9027
AIC	56085
AICC	56085
SBC	46104

Figure 8 - R-square result SAS studio

Source: SAS studio

The r-square value of 0.9027 representing a 90.27% coefficient of determination is slightly lower than the 90.53% value obtained from the ML predictive model.

6 Conclusion

This thesis was carried out to gather more information about AI and machine learning in general. This included information such as the proper definitions of the terms, their uses, history as well as the differences between them due to their often misuse and misinterpretation. In addition, useful modern areas such as big data, data mining and predictive analytics were glanced at.

The practical part focused on the accuracy of predictions made by the predictive models created which were built upon the linear regression method. The two models made were used in finding the relationship between the independent predictor (height) variable and the dependent (weight) variable. The models made were the ML predictive model and the statistical model which used the python programming language and SAS studio software, respectively.

A total population size of 10,000 entries were used for our dataset. These 10,000 entries consisted of variables height, weight and age which were analyzed, and the regression line formed for both models. Analysis and results obtained for both models included descriptive statistics measures such as the mean, median, standard deviation, and percentiles whose results were compared for both models.

Besides the ML predictive model being easier to use in terms of the ease with which obtaining the final predicted weight results for corresponding height and gender values can be done, the measured accuracy is also a factor to be considered when stating its pros.

After the required calculations and computation of the results, both models were compared using the coefficient of determination, R^2, where higher R^2 values signify better accuracy. The R^2 value was obtained for the entire dataset as well as for a randomly selected sample of the dataset to further test the accuracy obtained for our models.

With an R^2 value of 90.53% and 90.27% for the ML and SAS studio models respectively for the entire dataset, we can say for certain that in terms of the accuracy of obtaining the values of the dependent weight values from the independent height values and their corresponding gender that the ML model created is slightly more accurate.

From the above results and comparisons made, it can be deduced and concluded that for this experiment, the ML predictive model with Python is a slightly more accurate predictor than the statistical predictive model with SAS studio.

7 References

1. What is AI? / Basic Questions. jmc.stanford.edu [online] http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html Accessed 11 January 2021.
2. John Paul Mueller, Luca Massaron. Machine Learning for Dummies. July 2016 For Dummies. ISBN: 978-1-119-24551-3.
3. What is machine learning? https://pub.towardsai.net/differences-between-ai-and-machine-learning-and-why-it-matters-1255b182fc6 Accessed 11 January 2021.
4. Bhavsar, Parth & Safro, Ilya & Bouaynaya, Nidhal & Polikar, Robi & Dera, Dimah. (2017). Machine Learning in Transportation Data Analytics. 10.1016/B978-0-12-809715-1.00012-2.
5. John D. Kelleher, Brian Mac Namee and Aoife D'Arcy. Fundamentals of Machine Learning for Predictive Data Analytics. July 24th, 2015 MIT Press.ISBN-13: 978-0262029445 ISBN-10:0262029448.
6. Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning:From Theory to Algorithms.2014 Cambridge University Press. ISBN 978-1-107-05713-5.
7. Rudolph Russell. Machine Learning: Step-by-Step Guild to Implement Machine Learning Algorithms with Python. May 22, 2018 CreateSpace Independent Publishing Platform. ISBN-10: 1719528403 ISBN-13: 978-1719528405.
8. Big data https://www.gartner.com/en/information-technology/glossary/big-data Accessed 11 January 2021
9. What is predictive analytics? https://www.dataversity.net/what-is-predictive-analytics/#:~:text=(Kimberly%20Nevala),technology.%E2%80%9D%20(Georgetown%20University) Accessed 11 January 2021
10. Dr. Anasse Bari, Mohamed Chaouchi and Tommy Jung. Predicive Analyics For Dummies. December 2016. John Wiley & Sons. ISBN: 978-1-119-26700-3

8 Appendix

Additional references (not directly cited)

1. https://www.forbes.com/sites/gilpress/2016/12/30/a-very-short-history-of-artificial-intelligence-ai/
2. https://www.sciencedaily.com/terms/artificial_intelligence.htm
3. http://www-formal.stanford.edu/jmc/whatisai.pdf
4. https://www.javatpoint.com/types-of-artificial-intelligence
5. https://learn.g2.com/types-of-artificial-intelligence
6. https://www.forbes.com/sites/cognitiveworld/2019/06/19/7-types-of-artificial-intelligence/
7. https://www.javatpoint.com/application-of-ai
8. https://bernardmarr.com/default.asp?contentID=1908
9. https://www.sciencedirect.com/science/article/pii/S0016510720344667
10. https://link.springer.com/article/10.1007/s00401-017-1785-8
11. https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours
12. https://www.arm.com/glossary/ai-in-gaming
13. https://www.javatpoint.com/application-of-ai
14. https://marutitech.com/ways-ai-transforming-finance/
15. https://policyatmanchester.shorthandstories.com/on_ai_and_robotics/index.html
16. https://www.ieee-ras.org/humanoid-robotics
17. https://sciencefordemocracy.org/artificial-intelligence-in-humanoid-robots/
18. https://medium.com/towards-artificial-intelligence/what-is-machine-learning-ml-b58162f97ec7
19. https://towardsdatascience.com/the-basic-underlying-concepts-of-machine-learning-28282c8421f4
20. https://www.dummies.com/programming/big-data/data-science/how-does-machine-learning-work/
21. https://algorithmia.com/blog/how-machine-learning-works
22. https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f
23. https://www.sciencedirect.com/topics/computer-science/supervised-learning

24. https://social.technet.microsoft.com/wiki/contents/articles/53298.introduction-to-machine-learning-and-ml-net-part-1.aspx
25. https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a
26. https://www.researchgate.net/figure/Supervised-learning-classification-techniques_fig1_319370844
27. https://www.edureka.co/blog/classification-in-machine-learning/#classification
28. https://in.springboard.com/blog/regression-vs-classification-in-machine-learning/
29. https://www.guru99.com/unsupervised-machine-learning.html
30. https://www.tutorialandexample.com/ml-clustering-algorithm/