KIET Group of Institutions, Ghaziabad

COMPUTER SCIENCE

Internship Report

on

Python Programming & Machine Learning

Summer Internship at YBI Foundation

(Industrial Internship)

Aug - September

(2022)

Submitted By: ARCHIT RAJESH SRIVASTAVA

Course: B.TECH
Branch: CS
Sem: 5
Sec: A
ClassRoll No: 2000290120039

Index

ACKNOWLEDGEMENT

I’ve got this golden opportunity to express my kind gratitude and sincere thanks to my Head of Institution, KIET Group of Institutions of Engineering and Technology, and Head of Department of “**Computer Science”** for their kind support and necessary counselling in the preparation of this project report. I’m also indebted to each and every person responsible for the making up of this project directly or indirectly.

I must also acknowledge or deep debt of gratitude each one of my colleague who led this project come out in the way it is. It’s my hard work and untiring sincere efforts and mutual cooperation to bring out the project work. Last but not the least, I would like to thank my parents for their sound counselling and cheerful support. They have always inspired us and kept our spirit up.

Archit Rajesh Srivastava
BTech
Computer Science
Semester- 5
University Roll No: 2000290120039

CERTIFICATE

This is to certify that the internship project report entitled **"Python Programming & Machine Learning"** submitted by **Mr.Archit Rajesh Srivastava** in the Department of **C.S** of KIET Group of Institutions, Ghaziabad, affiliated to Dr. A. P. J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh, India, is a record of candidate summer internship. He/She has successfully completeted his/her internship under my supervision and guidance and is worthy of consideration for the same.

Signature of Supervisor:
Supervisor's Name:
Date:

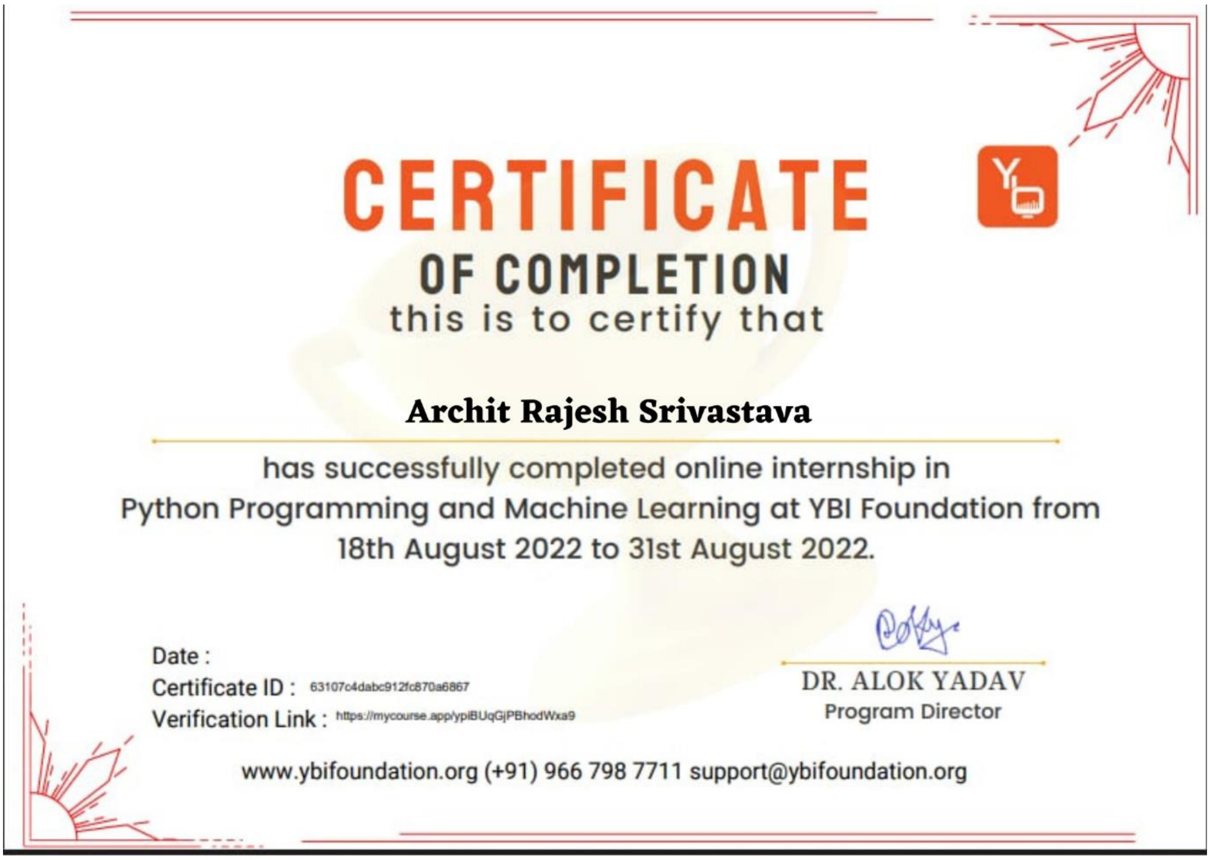
COMPANY PROFILE

YBI Foundation is a Private Company, who was incorporated 2 Year(s) 0 Month(s) 10 Day(s) ago on dated 22-Oct-2020 . YBI FOUNDATION is classified as Non-govt company and is registered at Registrar of Companies located in ROC-DELHI. As regarding the financial status on the time of registration of YBI FOUNDATION Company its authorized share capital is Rs. 1000000 and its paid up capital is Rs. 500000.

As Per Registration of Company, It involves under in Business Activity Class / Subclass Code 80903, Main Activity of the said Company YBI FOUNDATION is : , Activities of the individuals providing tuition, It Comes Under Division EDUCATION and this come under scetion EDUCATION.

YBI FOUNDATION's Corporate Identification Number is U80903DL2020NPL371984 and its registration number is 371984 .Its Email address is alokyadav@yantrabyte.com and its registered address is where Company is actual registered : C-176B, GALI NO. 38,MAHAVIR ENCLAVE PART 3,. For any Query You can reach this company by email address or Postal address.

Internship Certificate



CERTIFICATE

OF COMPLETION

this is to certify that

Archit Rajesh Srivastava

has successfully completed online internship in
Python Programming and Machine Learning at YBI Foundation from
18th August 2022 to 31st August 2022.

Date :
Certificate ID : 63107c4dabc912fc870a6867
Verification Link : https://mycourse.app/ypiBUqGjPBhodWxa9

DR. ALOK YADAV
Program Director

www.ybifoundation.org (+91) 966 798 7711 support@ybifoundation.org

Abstract

In this internship I made a project I made a diabetes prediction system.

Diabetes is an increasingly growing health issue due to our inactive lifestyle. If it is detected in time then through proper medical treatment, adverse effects can be prevented. To help in early detection, technology can be used very reliably and efficiently. Using machine learning we have built a predictive model that can predict whether the patient is diabetes positive or not.

In this project, the objective is to predict whether the person has Diabetes or not based on various parameters like Glucose level, Insulin, Age, BMI. We will use the PIMA India Databases from the UCI Machine learning repository. We will develop this project in six steps which follows data gathering to model deployment.

All the standard libraries like NumPy, pandas, matplotlib and seaborn were used. We use NumPy for linear algebra operations, pandas for using data frames, matplotlib and seaborn for plotting graphs. The dataset is imported using the pandas command read_csv().

Introduction of Project Internship

Project:

In this project, the objective is to predict whether the person has Diabetes Or not based on various features like Glucose level, Insulin, Age, and BMI.

Dataset used:

The dataset used in this project is **Pima Indians Diabetes Dataset** downloaded using Kaggle This original dataset has been provided by the **National Institute of Diabetes and Digestive and Kidney Diseases**. The link to the dataset and code is provided in the GitHub repository link below.

This dataset is used to predict whether a patient is likely to get diabetes based on the input parameters like Age, Glucose, Blood pressure, Insulin, BMI, etc. Each row in the data provides relevant information about the patient. It is to be noted that all patients here are females minimum 21 years old belonging to Pima Indian heritage.

Features of Dataset used:

The dataset contains 768 individuals' data with 9 features set. The detailed description of all the features are as follows:

- *Pregnancies:* indicates the number of pregnancies
- *Glucose:* indicates the plasma glucose concentration
- *Blood Pressure:* indicates diastolic blood pressure in mm/Hg
- Skin Thickness: indicates triceps skinfold thickness in mm
- Insulin*:* indicates insulin in U/mL
- *BMI:* indicates the body mass index in kg/m2
- *Diabetes Pedigree Function:* indicates the function which scores likelihood of diabetes based on family history
- *Age:* indicates the age of the person
- *Outcome:* indicates if the patient had a diabetes or not (1 = yes, 0 = no)

Link to GitHub repository:

https://github.com/archit1203/DiabetesPrediction

Details of task/s assigned

MOOC COURSE:

We were required to complete 30 hrs. of MOOC. The certificate was submitted titled

"**Introduction to Artificial Intelligence (AI)**". The course was instructed by Rav Ahuja.

Literature Review Report:

A summary of research papers was required to be submitted of 10 published research papers.

(They are attanched at the end)

MINI PROJECT:

I built a project on prediction of diabetes using various machine learning techniques.
The project was built on Jupyter notebook using Python language.

We were assigned a task to take the dataset of diabetes patient and analyze its various attributes and by selecting suitable attributes and with the help of various machine learning algorithms We had to develop a model and train it with this dataset so that it can predict whether a patient is diabetic or not on its own. Here, we used various algorithms like Logistic regression, Decision tree regression, Random Forest and KNN and ran the code on Jupyter Notebook and use various ML libraries like Pandas, seaborne, NumPy, matplotlib to generate better result and to achieve a good accuracy

Details of Technical learning during delivery of task

During the MOOC, I learned about what AI is, its applications and how it is transforming our lives. Basic concepts of AI and how it learns from the data. I also learnt about issues and concerns surrounding AI, including - ethical considerations, bias, jobs, etc. - their impact on society. At last, I learnt about the future with AI, as well as hear from experts about their advice to learn and start a career in AI. And also demonstrate AI in action by utilizing Computer Vision to classify images. Also, we were introduced to many IDEs like Jupyter notebook and Spyder, PyCharm, etc and multiple libraries also.

To sum up things, following were the things that We learned during this project:

- Basic knowledge of Python programming.
- Using different libraries on Python.
- Basics of Artificial intelligence.
- Different algorithms and their working principle.
- Working remotely on a Linux based computer.
- Searching internet and communities to find things and fix up errors in codes.
- Working with team.
- How to differentiate between different algorithms and where to use them.
- Accuracy, Precision and recall values along with confusion matrix.
- Reading and evaluating research papers and how to write them.

Conclusion

The internship was a good learning experience for me. I learned about various machine learning algorithms and classification, regression and its implementation. We were also taught about Python and using it via Anaconda and Jupyter notebook. Various datatypes like list, tuple and dictionary. We also learnt about Data Analysis & visualization –using NumPy, panda matplotlib, SciPy etc.

The project we built will give us confidence in our knowledge of various machine learning techniques, python IDE and whatever we learned will help us build more such AI projects.

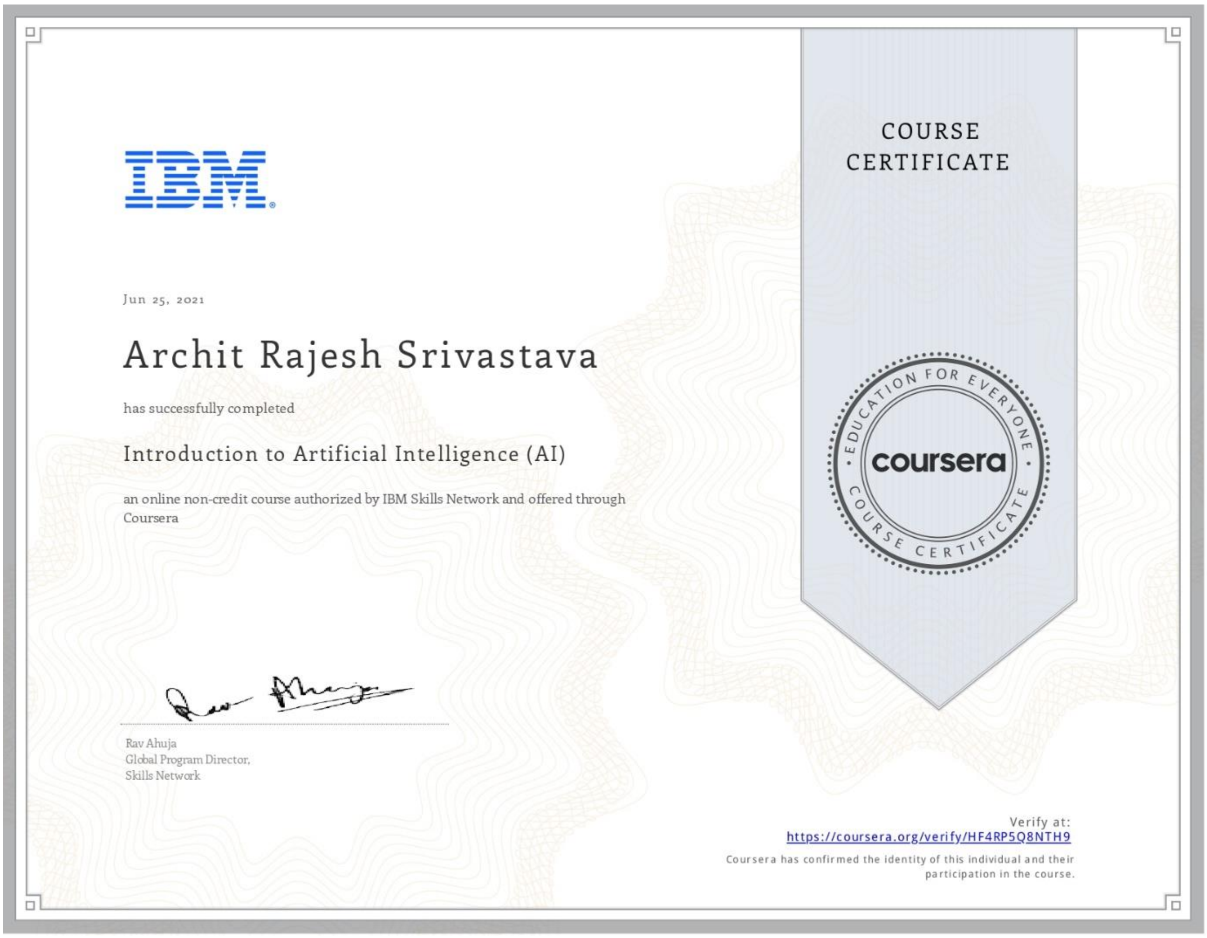
Some highlights of what we learned in the internship:

- Introduction to Artificial Intelligence
- Types of AI
- Ethical question about AI
- Python
  - Data Types
  - Functions in Python
  - Loops
- Data analysis and Exploration
  - Data Manipulation using NumPy, Pandas and Matplotlib.
  - Preprocessing of machine learning
- “Machine learning” a field of AI & its application
  - Implementation of various machine learning algorithms
  - Evaluation of various machine learning algorithm
- Project on Diabetes Prediction

Future scope of work

We build the model using some Machine Learning algorithms to predict the whether the person is diabetic or not, we can usze some other algorithms such as Naive Bayes, SVM (Support Vector Machine), ID3 ,etc to make our model more efficient. We can setup a module for visitor's query, where visitors can post their queries to doctors and doctors can send reply to those queries. We can also add a treatment module, where a patients can upload their diagnostic reports and doctors can see these reports and then they will suggest them proper treatment details.

MOOC Certificate(with verification link)



IBM

COURSE CERTIFICATE

Jun 25, 2021

Archit Rajesh Srivastava

has successfully completed

Introduction to Artificial Intelligence (AI)

an online non-credit course authorized by IBM Skills Network and offered through Coursera

EDUCATION FOR EVERYONE
coursera
COURSE CERTIFICATE

Rav Ahuja
Global Program Director,
Skills Network

Verify at:
https://coursera.org/verify/HF4RP5Q8NTH9
Coursera has confirmed the identity of this individual and their participation in the course.

https://coursera.org/share/f30872dcc85a3cc4b411a03393c32f58

Literature Review

1. An Introduction to Logistic Regression Analysis and Reporting

CHAO-YING JOANNE PENG KUK LIDA LEE GARY M. INGERSOLL

Indiana University-Bloomington

In this paper, we demonstrate that logistic regression can be a powerful analytical technique for use when the outcome variable is dichotomous. The effectiveness of the logistic model was shown to be supported by

(a) significance tests of the model against the null model,

(b) the significance test of each predictor,

(c) descriptive and inferential goodness-of-fit indices,

(d) and predicted probabilities.

During the last decade, logistic regression has been gaining popularity. The trend is evident in the JER and higher education journals. Such popularity can be attributed to researchers' easy access to sophisticated statistical software that performs comprehensive analyses of this technique. It is anticipated that the application of the logistic regression technique is likely to increase. This potential expanded usage demands that researchers, editors, and readers be coached in what to expect from an article that uses the logistic regression technique. What tables, charts, or figures should be included? What assumptions should be verified? And how comprehensive should the presentation of logistic regression results be? It is hoped that this article has answered these questions with an illustration of logistic regression applied to a data set and with guidelines and recommendations offered on a preferred pattern of application of logistic methods.

2. Supervised Machine Learning: A Review of Classification Techniques

S. B. Kotsiantis

Department of Computer Science and Technology University of Peloponnese, Greece

There are several applications for Machine Learning, the most significant of which is data mining. People are often prone to making mistakes during analyses or, possibly, when trying to establish relationships between multiple features. Every instance in any dataset used by machine learning algorithms is represented using the same set of features. The features may be continuous, categorical, or binary.

If instances are given with known labels, then the learning is called supervised, in contrast to unsupervised learning, where instances are unlabeled. By applying these unsupervised algorithms, researchers hope to discover unknown, but useful, classes of items. Numerous ML applications involve tasks that can be set up as supervised. In the present paper, we have concentrated on the techniques necessary to do this.

In this model, new features are computed as linear combinations of the previous ones. Decision trees can be significantly more complex representation for some concepts due to the replication problem. A solution is using an algorithm to implement complex features at nodes to avoid replication.

3. Selection of relevant features and examples in machine learning

Avrim L. Bluma, Pat Langley

School of Computer Science, Carnegie Mellon University Pittsburgh, PA 15213-3891, USA

More generally, weighting methods are often cast as ways of merging advice from different knowledge sources that may themselves be generated through learning. In this light, the weighting process plays an interesting dual role with respect to the filter methods discussed earlier. Filter approaches pass their output to a black-box learning algorithm, whereas weighting approaches can take as input the classifiers generated by black-box learning algorithms and determine the best way to combine their predictions. Each of these approaches shows improvement over use of all features, but only the latter reports comparisons with a simple selection of attributes.

This suggests a second broad type of relevance that concerns the examples themselves, and here we briefly consider techniques for their selection. Researchers have proposed at least three reasons for selecting examples used during learning. Another reason is if the cost of labelling is high, but many unlabeled examples are available or are easy to generate. Yet a third reason for example selection is to increase the rate of learning by focusing attention on informative examples, thus aiding search through the space of hypotheses.

The learner can also select data even before it has been labelled. This can be useful in scenarios where unlabeled data is plentiful, but where the labelling process is expensive. One generic approach to this problem, which can be embedded within an induction algo-rhythm that maintains a set of hypotheses consistent with the training data, is called query by committee [891. Given an unlabeled instance, the method selects two hypotheses at random from the consistent set and, if they make different predictions, requests the label for the instance.

4.Creativity and artificial intelligence

Margaret A. Boden

Creativity is a fundamental feature of human intelligence, and an inescapable challenge for AI. Creativity is not a special "faculty", nor a psychological property confined to a tiny elite. Rather, it is a feature of human intelligence in general. It is grounded in everyday capacities such as the association of ideas, reminding, perception, analogical thinking, searching a structured problem-space, and reflective self-criticism.

Current AI models of creativity focus primarily on the cognitive dimension. The ability to produce novelties of the former kind may be called P-creativity, the latter H-creativity. P-creativity is the more fundamental notion, of which H-creativity is a special case.

For its concepts evolve as processing proceeds. This research is guided by the theoretical assumption that seeing a new analogy is much the same as perceiving something in a new way. A part-built description that seems to be mapping well onto the nascent analogy is maintained and developed further. Whether the approach used in Copycat is preferable to the more usual forms of mapping is controversial.

5. Breast Cancer Classification Using Machine Learning Techniques: A Review

Article in Turkish Journal of
Computer and Mathematics Education (TURCOMAT) · September 2021

Artificial intelligence has been utilized for diagnosis early, rapidly, and accurately breast tumors. The objective of this paper is to review recent studies for classifying these tumors. The results showed that the SVM achieved high accuracy, about 97%, therefore, the researchers utilized various functions for this algorithm and added more features such as bagging and boosting to increase its efficacy. In addition, deep learning obtained high accuracy using CNN which is higher than 98%.

Recently, many scholars have mentioned that the mortality rate has raised in women due to breast cancer. Cancer is a creation of abnormal cells that come from a modification in these cells genetically and spreads into the body, a late in diagnosis and treatment leads to death. There are two types of breast cancer, invasive and non-invasive. The former is harmful, malignant, ability to infect other organs, and classified as cancerous.

Therefore, the research community proposed an automatic system called a CAD for better classification of tumors, accurate results, and rapid executing without needing for radiologists or experts. Machine learning algorithms are suggested as an alternative to human vision and experience for analyzing medical images and taking the final decisions with high accuracy.

6. Artificial Intelligence Contract as automation: representing a simple financial agreement in computational form

Mark D. Oliver R.

Financial legal structure of a well-written financial contract follows a state-transition logic that can be formalized mathematically as a finite-state machine (specifically, a deterministic finite automaton or DFAThe core of a contract describes the rules by which different sequences of events trigger particular sequences of state transitions in the relationship between the counterparties. By conceptualizing and representing the legal structure of a contract in this way, we expose it to a range of powerful tools and results from the theory of computation.

These allow, for example, automated reasoning to determine whether a contract is internally coherent and whether it is complete relative to a particular event alphabet. We illustrate the process by representing a simple loan agreement as an automaton. The key is that the state transition structure is sufficiently fundamental to a financial agreement that we can represent it using the standard computational formalism of a DFA without disrupting the contract's organizing principles.

Actual implementation of more complex financial contracts will use these shortcuts. Nonetheless, the DFA is a good starting point for highlighting the conceptual link between contracting and computation. There is a danger of the sorcerer's apprentice problem: that the unwise application of powerful computational tools could encourage inexperienced drafters with only limited understanding of the issues involved to create contracts of unmanageable complexity.

7. Car Price Prediction Using Machine Learning

Ketan Agrahari1, Ayush Chaubey2 , Mamoor Khan3 , Manas Srivastava

With the recent arrival of internet portals, buyers and sellers may obtain an appropriate status of the factors that ascertain the market price of a used automobile. Lasso Regression, Multiple Regression, and Regression Trees are examples of machine learning algorithms. We will try to develop a statistical model that can forecast the value of a pre-owned automobile based on prior customer details and different parameters of the vehicle.

This paper aims to compare the efficiency of different models' predictions to find the appropriate one. On the subject of used automobile price prediction, several previous studies have been conducted. Especially on higher-priced cars, the estimated value is not very close to the real price. In forecasting the price of a used car, they found that support vector machine regression outperformed neural networks and linear regression by a little margin.

With the rise in auto ownership, the used automobile market is ripe for growth. The healthy development of the used car market requires an accurate used car pricing evaluation. Since the developed system can be real-time and user friendly in terms of its handling. The authors of this study compared Linear Regression to Lasso Regression. The data for this study was gathered from Kaggle and then analysed using the Python programming language.

8. Deep Learning based single sample Face Recognition : A Survey

Fan Lie , Delong Chen , Fei Wang

Face recognition has long been an active research area in the field of artificial intelligence, particularly since the rise of deep learning in recent years. In some practical situations, each identity has only a single sample available for training. Face recognition under this situation is referred to as single sample face recognition and poses significant challenges to the effective training of deep models. Therefore, in recent years, researchers have attempted to unleash more potential of deep learning and improve the model recognition performance in the single sample situation. While several comprehensive surveys have been conducted on traditional single sample face recognition approaches, emerging deep learning based methods are rarely involved in these reviews. Accordingly, we focus on the deep learning-based methods in this paper, classifying them into virtual sample methods and generic learning methods. In the former category, virtual images or virtual features are generated to benefit the training of the deep model. In the latter one, additional multi-sample generic sets are used. There are three types of generic learning methods: combining traditional methods and deep features, improving the loss function, and improving network structure, all of which are covered in our analysis. Moreover, we review face datasets that have been commonly used for evaluating single sample face recognition models and go on to com-pare the results of different types of models.

9. Issue on Internet of Things (IoT) for in-vehicle systems

Duan et al The Internet of Things describes a network of smart devices in the physical world, endowed with embedded sensors and networking capabilities. The features of IoT that differ from traditional computing include the use of a multitude of devices with embedded sensors and connected to networks, introducing many benefits in terms of automation and optimization. For example, Google’s and GM’s self-driving car. Referring to the first paper «Data Privacy Protection for Edge Computing of Smart City in a DIKW Architecture», Duan et al. discuss the privacy content of multiple resources of types of the Data Information, Knowledge, and Wisdom architecture. The authors also categorize target privacy resources of data and information according to their presence in the model searching space in the DIKW architecture as explicit and implicit divisions with protection solutions. First, the cyclomatic complexity is computed according to the number of nodes and arcs in the workflow model by the Kuhn-Munkres algorithm which is used to map the independent paths from two workflow models and generate the optimal mapping that the overall distance is the minimum. Companies should select and implement an adequate set of cybersecurity solutions for both software and hardware of their vehicles.

10. Potential, challenges and future directions for deep learning in prognostics and health management applications

Olga Fink and Melanie Ducoffe

Deep learning applications have been thriving over the last decade in many different domains, including computer vision and natural language understanding. The drivers for the vibrant development of deep learning have been the availability of abundant data, breakthroughs of algorithms and the advancements in hardware. Despite the fact that complex industrial assets have been extensively monitored and large amounts of condition monitoring signals have been collected, the application of deep learning approaches for detecting, diagnosing and predicting faults of complex industrial assets has been limited. The goal of Prognostics and Health Management is to provide methods and tools to design optimal maintenance policies for a specific asset under its distinct operating and degradation conditions, achieving a high availability at minimal costs. PHM integrates the detection of an incipient fault, its isolation, the identification of its origin and the specific fault type and the prediction of the remaining useful life. The system health management goes beyond the predictions of failure times and supports optimal maintenance and logistics decisions by considering the available resources, the operating context and the economic consequences of different faults.