وزارة التخطيط

دائرة التعاون الدولي

نموذج تقرير مشاركة في البرامج التدريبية

اولا : معلومات المشترك

اسم المشترك	سرى عبد اللطيف هادي
التحصيل الدراسي والاختصاص	ماجستير علوم حاسبات / ذكاء اصطناعي
العنوان الوظيفي	م.رئيس مبرمجين
اسم الجهه الحكومية	وزارة التعليم العالي والبحث العلمي / هيئة البحث العلمي
البريد الالكتروني	Sura٢٩١٠٧٩١@yahoo.com
رقم الهاتف	٠٧٧٠٩٢٦٠٨٣٧

ثانيا : معلومات البرنامج التدريبي

عنوان البرنامج	برنامج متخصص في تعليم الاليات
طبيعة البرنامج التدريبي	برمجي
البلد	الهند
الجهه الراعية	ITEC
الجهه المنظمه	C-DAC, Noida
مدة البرنامج	اسبوعين
التاريخ	من ٢٠٢٥/١/٢٢ الي ٢٠٢٥/٢/٤
الجهات الحكومية المشاركة في البرنامج	لايوجد

البلدان المشاركة الاخرى	ماليزيا ,اثيوبيا , استراليا , امريكا الجنوبية ,غينا

ثالثا : محاور ومواضيع البرنامج التدريبي

١: مقدمة في الذكاء الاصطناعي وتعلم الآلة

- ما هو الذكاء الاصطناعي (AI) ؟
- الفرق بين الذكاء الاصطناعي وتعلم الآلة والتعلم العميق .
- أنواع الذكاء الاصطناعي:

(Super AI) الخارق (General AI) العام (Narrow AI) الضيق.

- تطبيقات الذكاء الاصطناعي في مختلف الصناعات .

٢: أنواع تعلم الآلة:

١ - التعلم المراقب (Supervised Learning):

(Classification) التصنيف و(Prediction) التنبؤ.

٢- التعلم غير المراقب (Unsupervised Learning):

(Clustering) تحليل الأنماط والتجميع.

٣- التعلم المعزز و مفاهيمه الأساسية (Reinforcement Learning) .

٣: بيئة العمل والأدوات البرمجية:

- مقدمة في لغة Python لتعلم الآلة .
- المكتبات الأساسية : Scikit-learn، Seaborn، Matplotlib، Pandas، NumPy .
- بيئات التطوير: Google Colab و Jupyter Notebook .
- تمرين عملي على تحليل بيانات بسيطة باستخدام Pandas .

٤: معالجة البيانات وتحضيرها للنماذج:

- أهمية البيانات وجودتها في تدريب النماذج .
- تقنيات تنظيف البيانات (Data Cleaning).

- تحويل الميزات (Feature Engineering)
- تطبيع البيانات (Standardization) وتقييسها (Normalization) .
- تمرين عملي على تنظيف وتحليل مجموعة بيانات واقعية.

٥: بناء وتقييم نماذج تعلم الآلة

- تقسيم البيانات إلى مجموعات تدريب واختبارها.
- اختيار النموذج المناسب للمشكلة .
- تقييم أداء النموذج باستخدام مقاييس مثل الدقة، مصفوفة الالتباس، وخطأ الانحدار.

الأسبوع الثاني : التطبيقات المتقدمة والمشاريع العملية

٦: التعلم العميق والشبكات العصبية (Deep Learning).

- مقدمة في الشبكات العصبية الاصطناعية(ANN) .
- أستخدام لإنشاء نماذج التعلم العميق Keras و TensorFlow.
- شرح خوارزمية الانحدار الخطي Linear Regression.
- شرح خوارزمية الانحدار اللوجستي Logistic Regression.
- شرح خوارزمية الجيران الاقرب k-nearest Neighbors KNN .
- شرح خوارزمية الة المتجهات الداعمة Support vector machine SVM.

٧: رؤية الحاسوب (Computer Vision) ومعالجة الصور.

- مقدمة في رؤية الحاسوب وتطبيقاتها .
- استخدام TensorFlow/Keras و OpenCV لمعالجة الصور والتعرف على الأنماط .

تمرين عملي: بناء نموذج تصنيف صور باستخدامCNN.

٨: معالجة اللغة الطبيعية (NLP)

- ما هوNLP وماهي أهميته في تحليل النصوص.
- مقدمة في نماذج تحليل النصوص مثل BERT و Word٢Vec.

٩: تعلم الآلة التطبيقي نشر النماذج (MLOps)

- أهمية تشغيل نماذج تعلم الآلة في البيئات الإنتاجية MLOps.

١٠: مشروع عملي شامل

❖ فهم المفاهيم الأساسية للذكاء الاصطناعي وتعلم الآلة .

- استخدام أدوات Python لتحليل البيانات وبناء النماذج.
- تطوير نماذج تعلم الآلة والتعلم العميق لمهام مختلفة.
- تطبيق تعلم الآلة في معالجة الصور والنصوص.
- نشر النماذج في بيئات حقيقية واستخدامها في التطبيقات.

رابعا : المنهاج التدريبي والمواصفات التخصصية

١. Introduction to Machine Learning.
٢. What is Machine Learning ؟.
٣. Types of Machine Learning.
٤. Machine Learning Workflow.
٥. Popular ML Algorithms.
٦. Applications of Machine Learning.
٧. Challenges in Machine Learning.
٨. Future Trends in Machine Learning.

أ- **المنهاج التدريبي** : شمل التدريب كافة الادوات الرصينة بما يشمل من شرح مفاهيم تعلم الالة وجميع الخوارزميات المستخدمة وتنفيذ الكود البرمجي الخاص بكل خوارزمية ومعرفه مدى اهمية تلك الحوارزمية وسبب اختيارها دون غيرها من خوارزميات الذكاء الاصطناعي وتعالم الالة وتم االاستفادة من بناء كود برمجي لجميع الخوارزميات حيث يمكن تطبيق ذلك في الحياة العملية في العراق وخصوصا في المجال البحثي لجميع ابحاث الذكاء الاصطناعي .

ب- المواصفات التخصصية :

أشهر المصادر التخصصية المعتمدة عالميًا

❖ الجامعات:

- Stanford University (CS٢٢٩) ➢
- MIT (٦,٠٣٦ Introduction to ML) ➢
- Carnegie Mellon (١٠-٦٠١) ➢

❖ المنصات التعليمية العالمية:

- Coursera: Andrew Ng's ML course (Stanford) ➢
- edX: Columbia, Harvard AI/ML courses ➢
- Udacity: Machine Learning Engineer Nanodegree ➢
- Fast.ai: دورات عملية علىPyTorch ➢

❖ ٣. المعايير التقنية للمحتوى

❖ لغات البرمجة) Python: مع مكتبات Scikit-Learn, TensorFlow, PyTorch)

❖ أدوات البياناتPandas, NumPy, Matplotlib :

❖ تطبيقات: التصنيف، التنبؤ، التجميع، التعلم العميق، النماذج التوليدية

❖ ثانياً: ما هو معتمد أو مطبق في العراق

➢ الجامعات العراقية

❖ معظم الجامعات العراقية تُدرّس مفاهيم تعلم الآلة ضمن تخصصات:

❖ علوم الحاسوب(Computer Science)

❖ الذكاء الاصطناعي (في بعض الجامعات مثل جامعة النهرين والجامعة التكنولوجية)

❖ هندسة البرمجيات

❖ لكن غالبًا ما تكون:

❖ **المواد نظرية بشكل أكبر.**

❖ هناك نقص في الموارد العملية والمختبرات.

❖ قلة استخدام المنصات التعليمية الحديثة.

➢ **مراكز التدريب الأهلية**

❖ تُقدم بعض المراكز في بغداد وأربيل ودهوك برامج تدريبية في ML باستخدام كورسات مثل كورس Andrew Ng أو محتوى مترجم من Coursera و Udemy.

❖ لا توجد شهادات وطنية معترف بها رسميًا بهذا التخصص، ويعتمد الخريجون غالبًا على الشهادات العالمية.(Coursera, Google ML certification)

خامسا : النشاطات الصفية والميدانية

أ- **الصفيه او النظريه**

تطبيق جميع خوارزميات تعلم الالة خوارزمية وملاحظة الفرق بين كل واحدة ولماذا تم استخدام هذه الخوارزمية لتخصص كل خوارزمية بمجال محدد ونتائج مختلفة حيث تم عرض مشكلة قمت بااعدادها وهيه كيفية التبوء المبكر للامراض القلب بدمج خوارزميات تعلم الالة والتعلم العميق التي كانت تجربتي في الماستر بجامعة العلوم الماليزية USM حيث تناقشت مع الاستاذ شيفام بما حصلت عليهه من نتائج والخوارزميات التي قمت بااستخدامها .

ب- **الميدانية**

لم يتم اجراء اي نشاط ميداني سواء التجارب المختبرية التي قمنا بها داخل المختبر بااستخدام اجراء الكود البرمجي لجميع الخوارزميات ومقارنه بين كل واحدة منهم .

سادسا : التقارير والعروض التقديمية

أ- **التقارير:** تقرير بعنوان اكتشاف مرض التوحد باستخدام تقنيات الذكاء الاصطناعي وخوارزميات تعلم الالة والتعلم العميق CNN Convolutional Neural Network , ANN Artificial Neural Network

ب- العروض التقديمية :عرض تقديمي

عن استخدام تقنيات الذكاء الاصطناعي في اعادة تاهيل مرض السكتة الدماغية عن طريق خوارميات التصنيف واستخدام تقنيات الانتباه .

سابعا : البرامجيات والتقنيات التكنولوجيه الحديثة

أ- **البرمجيات:** برنامج بايثون البرامج التي تناولها البرنامج وتم التدريب عليه **لغة Python** هي لغة برمجة عالية المستوى وسهلة التعلم، تُستخدم في العديد من المجالات مثل تطوير البرمجيات، تحليل البيانات، الذكاء الاصطناعي، أتمتة المهام، تطوير الواجهات، وإنشاء تطبيقات الويب.

ب- **لماذا Python للمؤسسات الحكومية؟**

ت-مفتوحة المصدر ومجانية بالكامل.

ث-سهلة القراءة والتطوير، مما يسهل على الفرق غير المتخصصة التقنية فهم الشيفرات.

ج- مكتباتها قوية وتدعم معظم التطبيقات الحكومية مثل الأرشفة، التحليل، الأتمتة، والذكاء الاصطناعي.

ح- **ثانيًا: تطبيقات Python المفيدة للمؤسسات الحكومية**

خ- سأعرض الآن أهم البرامج أو الحلول التي يمكن تطويرها باستخدام Python، مع شرح لفائدتها وطريقة عملها:

د- **١نظام أرشفة إلكترونية**

ذ- الفكرة :بناء نظام لحفظ الملفات إلكترونيًا وربطها ببيانات وصفية (رقم القرار، التاريخ، القسم).

ر- **الفائدة:**

- سهولة الوصول للملفات.
- تقليل الورق وتقليل الوقت الضائع في البحث.
- **طريقة العمل:**
- رفع الملفات إلى النظام.
- إدخال بيانات عنها (وصف، تاريخ، تصنيف).
- البحث والاسترجاع عند الحاجة.
- **أمثلة من مكتبات:Python**
- Flaskلتطوير الواجهة.
- SQLiteلتخزين البيانات.
- os وshutilللتعامل مع الملفات.

ز- **التقنيات الحديثة الاخرى :** تقنيات الذكاء الاصطناعي باستخدام لغة البرمجة بايثون التي تعبتر احدث تقنيات البرمجية للذكاء الاصطناعي والذي يستخدم في اغلب المؤسسات الحكومية .

س- **اخرى :** لايوجد

ثامنا : الموائمة و/او محاور أخرى

اود بتوضيح عملي في تضمين استخدام تقنيات الذكاء الاصطناعي وتعلم الالة بعمل نظام تميز صوت الاشخاص والتعرف على صورهم الذي قد تم فعلا البدء به

والوصول الى التعرف على الاشخاص وتميز صوتهم من خلال اعطاء اوامر مختلفة .

تاسعا : التجارب المستفادة

استفدت من خلال هذا البرنامج التدريبي في تمكين ٣ انواع خوارزميات في اعادة تاهيل مرضى السكتة الدماغية من خلال تصنيف التمارين ونوعها ومدى استفاده المريض من هذه الاليات في اعادة التاهيل للمرضى والذي هو جزء من رسالة الدكتوراه التي اقوم بها في جامعة العلوم الماليزية وسوف تكون الجهة المستفيدة هيئة البحث العلمي .

عاشرا : تقييم البرنامج التدريبي

أ- الامور التنظمية للدورة كانت جيدة جدا من ناحية الامور التنظيمية والترتيب والمختبرات بالدورة التدريبية والفندق واجراءات تاشيرة الدحول كانت جيدة جدا ولايوجد اي شي سلبي .

ب- المنهاج التدريبي كان جيد جدا من حيث المحتوى والاساليب المتبعة وقدرات المدرب والاساليب المتبعة وجو التدريب .

ج- أي ملاحظات حول تطوير البرامج اللاحقة.

من المفيد أن يكون جميع المشاركين ذو خبرة على

الاقل مستوطة للاستفادة من البرنامج التدريبي .

الحادي عشر : التوصيات والمقترحات:

من اهم التوصيات والمقترحات هي ان يكون هنالك برنامج تدريبي متقدم من نفس الجهة المانحة وخصوصا بعد معرفة امكانيتهم بتدريب برنامج متقدم في تعلم الالة وقد تم الحصول على شهادة مشاركة ورموز تذكارية وكذلك يجب اقامة دورة متكاملة من قبلي في مديرية التدريب لشرح جميع البرنامج التدريبي .

اولا : معلومات المشترك

اسم المشترك	نضال حسين محمد علي
التحصيل الدراسي والاختصاص	بكالوريوس علوم /علوم حاسبات
العنوان الوظيفي	معاون رئيس مبرمجين
اسم الجهة الحكومية	وزارة التعليم العالي /هيئة البحث العلمي
البريد الالكتروني	mahm783352@gmail.com
رقم الهاتف	07727553001

ثانيا : معلومات البرنامج التدريبي

عنوان البرنامج	برنامج متخصص في تعلم الاليات Special Program on machine learning
طبيعة البرنامج التدريبي	برنامج تدريبي متخصص في تعليم الاليات بلغة البايثون
البلد	الهند
الجهة الراعية	البرنامج مقدم من قبل التعاون التقني والاقتصادي الهندي (ITEC) التابع لوزارة الشؤون الخارجية لحكومة الهند
الجهة المنظمة	دولة الهند في مركز تطوير الحوسبة المتقدمة في نويدا Center for development of advanced computing (C-DAC)
مدة البرنامج	اسبوعين
التاريخ	من ٢٠٢٥/١/٢٢ الى ٢٠٢٥/٢/٤
الجهات الحكومية المشاركة في البرنامج	تم تنظيم البرنامج من قبل الحكومة الهندية والمشاركين في البرنامج من العراق و من مختلف دول العالم

IRAQ , BHUTAN ,CAMBODIA , ETHIOPIA, KAZAKHSTAN , KENYA , MALAYSIA MAURITIUS , NIGERIA , PARAGUAY, SOUTH SUDAN ,TAJIKISTAN ,UGANDA , UZBEKITAN	البلـــدان المشـــاركة الاخرى

ثالثا : محاور ومواضيع البرنامج التدريبي

Introduction to Machine Learning and Related concepts

Introduction to Machine Learning

Machine learning (ML) is the study of computer algorithms that improve automatically through experience and of data.it is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data , known as "training data" , in order to make predictions or decisions without being explicitly programmed to do so

Voice Assistants	some super cool use cases of Machine Learning
Stock Price Prediction	
Email Filtering	
Product Recommendations	
Predictive Maintenance	
Self Driving Cars	

1- Machine Learning vs Traditional Programming

- **Traditional programming**

In traditional programming, business requirements are converted into "deterministic" rules. These rules are then applied on input data to produce outcomes. The rules are deterministic implying that th y

do not change or learn from the data they ingest or outcomes they produce.

- Machine Learning

In Machine Learning, you will not write "deterministic" rules. Instead, you will use Machine Learning Algorithms to learn from the historical data fed and create a Model, which is essentially the set of rules that the algorithm has learnt from the data to solve a complex problem for which writing static rules is unsuitable.

2- Why Machine Learning

Can easily replace traditional programming methods where complex and very large set of combinations of rules (Emile spam Filtering)

Complex Problems where no traditional solution exist, e.g. Image Processing, Voice Processing, Autonomous Engines such as Self Driving Cars (Image Classification)

In situations with changing scenarios (Stock Market Prediction)

Getting insights about complex problems and large amounts of data (Fraud Detection)

3- Types of Machine Learning

Machine Learning Algorithms can be classified in the following ways:

Basis of Classifications of ML Algorithms

Whether or not they are trained with Labelled Target values

Supervised Learning

UN Supervised Learning

Semi-Supervised Learning

Reinforcement Learning

Whether or not they can learn incrementally on the fly

Online Learning

Batch Learning

Whether they work by simply comparing new data point to known data points, or instead detect patterns in the training data and build a predictive model

Instance-based Learning

Modal-based Learning

- Types of Machine Learning

  Classification based on whether they are labeled /unlabeled/ reward based

  - Supervised Learning : In supervised learning, the input data consisted of "labels" or the target values.

    E.g : k-nearest Neighbours , Linear Regression, Logistic Regression, Support Vector machines (SVMs) Decision Trees, Random Forests.

  - UnSupervised learning : In unsupervised learning, the input data does not include the labels.

    E.g : K-Means, DBSCAN, PCA, T-SNE, Apriori

  - Semi- Supervised learning: In Semi-supervised Learning, you would normally start with a supervised method and the system gradually move into an unsupervised learning.

    E.g: Deep Bellf Networks.

  - Reinforcement Learning :In Reinforcement Learning, an "Agent" tries to find an optimum path to solve a problem based on a "reward"/"penalty" basis. The target has to be achieved in the maximum reward path.

4- Types of Machine learning

Classification based on whether or not they can learn incrementally

- Batch learning :in batch learning the model is trained using all the available data at one go. this may take more time and computing resources if the volume of input data is large, hence it is typically done offline.
- Online learning : in online learning (also termed as incremental or out –of core learning),the model is trained incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches. Each learning step is faster than batch learning method and it ensures that a productionised models is kept up-to-date.

5- Types of Machine Learning

Whether they work by simple comparing new data points to known data points or instead detect patterns in training data and build a predictive model ,much like scientists do

- Instance based learning : in instance-based learning(also called memory-based-learning),the system learns from the examples fed to it ,then generalizes to new cases by comparing them to the learned examples (which are stored in the memory), using a similarity measure.
- Model based learning :in model-based learning ,the system learns from input examples to build a model of these examples then we use that model to make prediction .

6- Use cases of Machine Learning

- Medical imaging
- Stock Price Prediction
- Customer Segmentation
- Product Recommendation
- Sentiment Analysis
- Fraud Detection
- Customer Chum Prediction
- Robotic Assistants
- Autonomous Engines

7- Role of Data in Machine Learning

Data is in the heart of every Machine Learning Model. Learning Algorithms are applied on Data to learn patterns and create models to solve a problem, such as a Regression problem or a classification problem.

The quality of the input data is hence key to the success or accuracy of the machine learning model thus created.

Machine learning professional must take utmost care in choosing the right data and appropriate analysis and subsequent pre-processing to make the data suitable for ingestion to a specific Learning Algorithm to create the right model.

Models created by using rightly pre-processed data are subjected to validation/test to measure the accuracy/performance of the model.

Types of Data

- Numerical Data
- Categorical Data
- Time-Series Data
- Text Data
- Image Data

8- Challenges in Machine Learning

The Two key components of Machine Learning

Algorithm

Data

Hence the challenges in setting up a successful ML system is:

- Bad Algorithm
- Bad Data

9- Linear Regression

- Introduction to Linear Regression
- Understanding Cost Functions
- Linear Regression Assumption
- problem Statement
- Data pre-processing
- Building our LR Model
- Understanding Model performance
- Model tuning methods

10- Logistic Regression

- Introduction to Logistic Regression
- Estimating probabilities
- Logistic Regression cost Functions
- Softmax Regression
- Performance metrics
- Roc Curve and AUC
- Optimizing Logistic Regression model

11- Support Vector Machine (SVM)

- Introduction to support vector machine
- Linear SVM Classification
- Non Linear SVM Classification
- Polynomial Kernel Trick

12- Decision Tree Classifier

- Introduction to Classification Algorithms
- Introduction to Decision Trees
- Gini Index and Entropy
- Decision Tree in Practice
- Measuring Performance

13- Random Forest Classifier

- Recap on Decision tree
- Ensemble Techniques and Bagging Process
- Introduction to Random forest
- Gini Index and Entropy
- Steps in the building of a Random Forest classifier
- Random Forest in practice
- Measuring performance
- Out of Bag Error
- Credit Default prediction

14- Unsupervised learning with k-means clustering

- Unsupervised Learning
- Introduction to k-means
- K-means clustering process steps
- Centroid optimization/convergence
- Other consideration
- Optimizing number of clusters

رابعا : المنهاج التدريبي والمواصفات التخصصية

لقد تم اقامة البرنامج التدريبي في مركز تطوير الحوسبة المتقدمة في نويدا(C-DAC) تضمن البرنامج سلسلة من المحاضرات اليومية ولمدة اسبوعين ويتم الذهاب الى المركز من الساعة التاسعة صباحا وحتى الساعة الخامسة مساءا ماعدا ايام العطل الرسمية وكانت يومين في الاسبوع وهما يومي (السبت والاحد)كانا عطلة رسمية في دولة الهند ، وتناول المنهاج التدريبي المواضيع التي تم ذكرها في الفقرة السابقة وحسب الاختصاص وكان البرنامج التدريبي يهدف الى تعلم

- فهم اساسيات وانواع التعلم الالي types of machine learning
- التعلم الاشرافي المتقدم في مجال اللوغاريتمات
- استكشاف تقنيات التعلم الغير الخاضعة للرقابة
- التعلم الخاضع للاشراف باستخدام الاشجار
- مقدمة عن الشبكات العصبية

ويعد التعلم الالي (machine learning)هو دراسة خوارزميات الكومبيوتر التي تتحسن تلقائيا من خلال التجربة واستخدام البيانات وينظر اليه على انه جزء من الذكاء الاصطناعي وتبني خوارزميات التعلم الالي نموذجا يعتمد على بيانات العينة المعروفة باسم"بيانات التدريب" من اجل عمل تنبؤات او قرارات دون ان تتم برمجتها صراحة

خامسا :النشاطات الصفية والميدانية

أ- الصفية او النظرية

تم شرح المادة من قبل الاستاذ وتنفيذ البرامج مباشرة على الكومبيوتر حيث يجب على كل مشترك ان ينفذ البرامج على الكومبيوتر واظهار النتائج

ب- الميدانية

لاتوجد هنالك اي زيارات ميدانية ضمن البرنامج التدريبي

سادسا: التقارير والعروض التقديمية

لاتوجد هنالك تقارير اوعروض تقديمية مطلوبة من المشترك ضمن البرنامج التدريبي

سابعا: البرامجيات والتقنيات التكنولوجيا الحديثة

البرامجيات التي تم استخدامها هي التعلم الالي باستخدام لغة البايثون

Machine learning using python

حيث تعتبر لغة python من اكثر اللغات استخداما في هذا المجال بسبب سهولتها وقدراتها الواسعه في معالجة البيانات وانشاء النماذج التنبؤية حيث يعتبر التعلم الالي باستخدام بايثون واسعا ومتطورا ويوفر فرصا كبيرة للابتكار في مجالات مثل تحليل البيانات والذكاء الاصطناعي بفضل المكتبات القوية والمجتمع النشط اصبحت بايثون الخيار الاول للعديد من المطورين والعلماء

ثامنا: الموائمة او محاور اخرى

لاتوجد

تاسعا: التجارب المستفادة

- فهم كيفية استخدام التعلم الالي لحل المشكلات في العالم
- فهم انواع خوارزميات التعلم الالي واطار عمل بناء نماذج
- معرفة سبب اعتماد بايثون على نطاق واسع كمنصة لبناء التعلم الالي
- معرفة المكتبات الرئيسية في بايثون
- تعلم كيفية تحميل البيانات المنظمة على اطار البيانات
- تعلم انشطة اعداد البيانات مثل التصفية والتجميع والترتيب
- توزيع الاحتمالات واختبارات الفرضيات
- مفهوم الانحدار الخطي البسيط والمتعدد وتطبيقاته في التنبؤ

عاشرا: تقيم البرنامج التدريبي

- البرنامج التدريبي مقدم من قبل التعاون التقني والاقتصادي الهندي(ITEC) التابع لوزارة الشؤون الخارجية لحكومة الهند وقد تم انعقاد البرنامج في مركز تطوير الحوسبة المتقدمة (C-DAC)في نويدا حيث تحمل الجانب الهندي تكاليف السفر والاقامة وتوفير التاشيرة وتذاكر الطيران عن طريق السفارة الهندية في بغداد والاقامة والنقل الداخلي والوجبات الغذائية.
- حيث تم استقبالنا من مطار دلهي من قبل ممثل الجانب الهندي وايصالنا الى مقر الاقامة
- تضمن البرنامج التدريبي سلسلة من المحاضرات اليومية ولمدة اسبوعين ماعدا ايام العطل الرسمية وهما يومي (السبت والاحد)حيث كانتا عطلة رسمية في دولة الهند وقد خصص هذين اليومين كبرنامج ترفيهي للمشاركين في البرنامج التدريبي وذلك لزيارة الاماكن التاريخية والاثرية في دولة الهند
- شارك في البرنامج (٣٠)موظف من مختلف دول العالم ومن ذوي الشهادات والدرجات الوظيفية المختلفة
- وفي اليوم الاخير من البرنامج خصص لاجراء الامتحان الالكتروني النهائي في مركز (C-DAC)وتم منح المشاركين شهادة مشاركة واخذ الصور الجماعية والاحتفال مع وجبة غداء مع المسؤولين عن البرنامج.
- وبعد نهاية الفترة المحددة للبرنامج غادرنا مكان الاقامة الى مطار دلهي للعودة الى وطننا الحبيب.

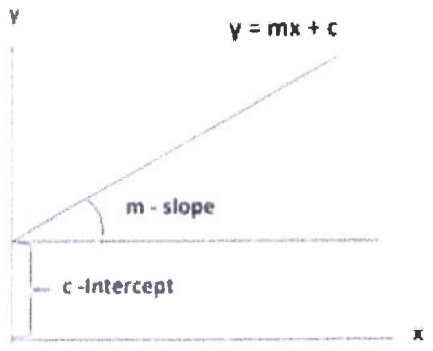
الحادي عشر: التوصيات والمقترحات

- البرنامج التدريبي كان جيدا ومتطور ومفيدا وخاصة لذوي الاختصاص في علوم الحاسبات والذكاء الاصطناعي ونقترح ان يكون هناك برنامج تدريبي مكمل لنفس البرنامج والمواضيع وذلك لكون الفترة كانت قصيرة بالنسبة للمواضيع التي تناولها البرنامج.

Simple Linear Regression with an Example

Year	Advertising Spend (US$)	Sales US$
2003	41000	22221000
2004	47000	25738000
2005	53000	24625000
2006	54000	28914000
2007	62600	32406000
2008	63000	34221000
2009	67600	32812000
2010	71800	35419000
2011	70000	37681000
2012	74400	35682000
2013	74800	38946000
2014	77000	37559000
2015	78000	39215000
2016	81900	41922000
2017	80900	40519000
2018	83000	41958000
2019	85500	43419000
2020	95000	?
2021	102000	?
2022	110000	?

Simple Linear Regression with an Example



- Predictors
- Beta Coefficients
- Independent and Target Variables
- Making Predictions
- Multiple Linear Regression
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_n x_n$

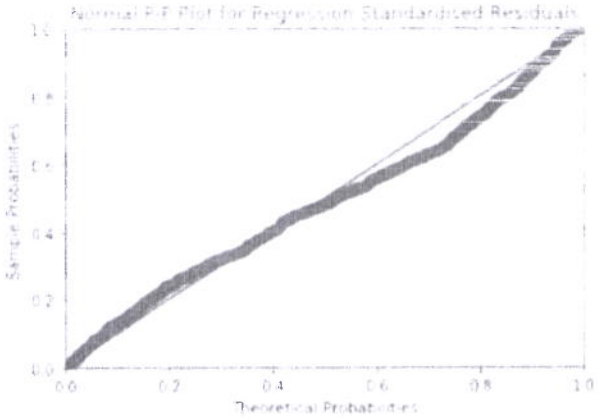
Testing the Model

Residual Normality

The P-P (Probability-Probability) plot further establishes the normality of the Residuals.

The normal probability plot or P-P Plot is a graphical technique for assessing whether or not a data set is approximately normally distributed. (here the dataset indicates the error terms.

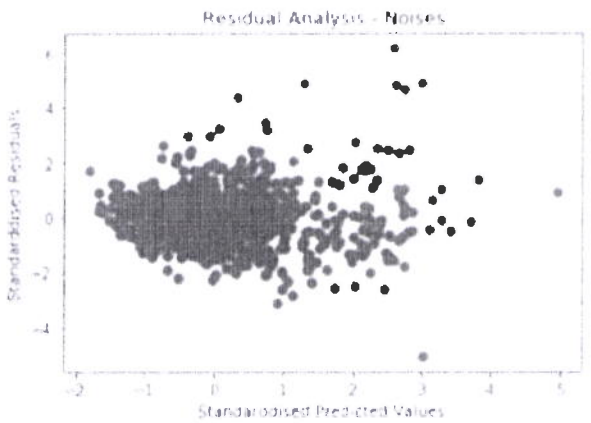
The data (Error terms) are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line.



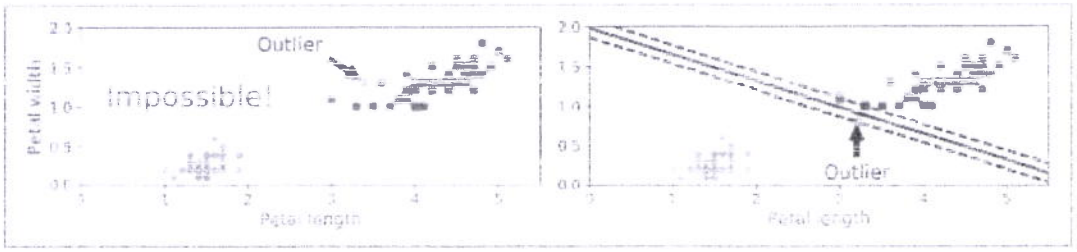
Testing the Model

Test of Homoscedasticity of residuals:

Variance of the residuals should not change for each predicted value of the target variable or there should not be a dependency pattern between the error term variance and predicted values. This is determined by plotting the residuals against the predicted values.



Hard Margin Classification



If we strictly impose that all instances be off the street and on the correct side, this is called ***hard margin classification***. There are two main issues with hard margin classification.

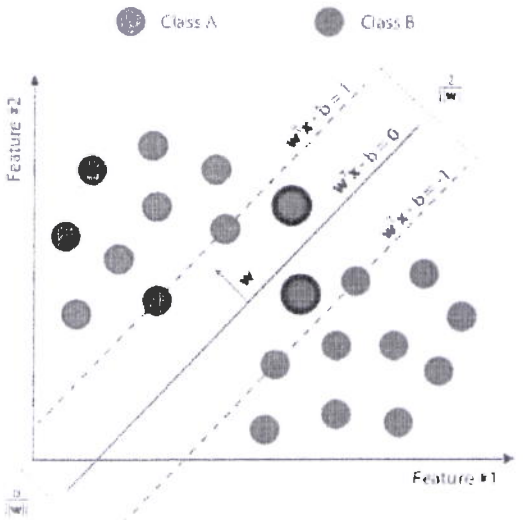
First, it only works if the data is linearly separable, and second it is quite sensitive to outliers.

Figure shows the iris dataset with just one additional outlier: on the left, it is impossible to find a hard margin, and on the right the decision boundary ends up very different from the one we saw in Figure without the outlier, and it will probably not generalize as well.

Soft Margin Classification

To avoid these issues it is preferable to use a more flexible model. ***The objective is to find a good balance between keeping the street as large as possible and limiting the margin violations*** (i.e., instances that end up in the middle of the street or even on the wrong side).

This is called ***soft margin classification***.



Bias and Variance

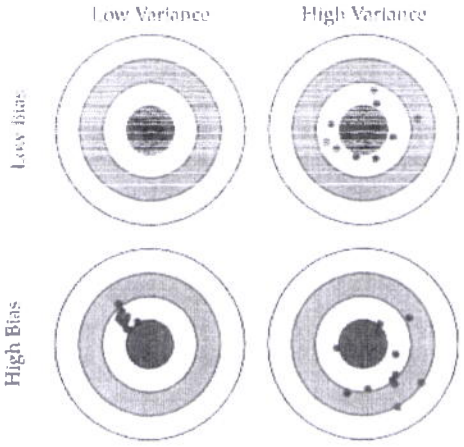
There are two types of Generalization Errors that can come out of Machine Learning Models:

Bias

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. **Model with high bias pays very little attention to the training data and oversimplifies the model.** It always leads to high error on training and test data.

Variance

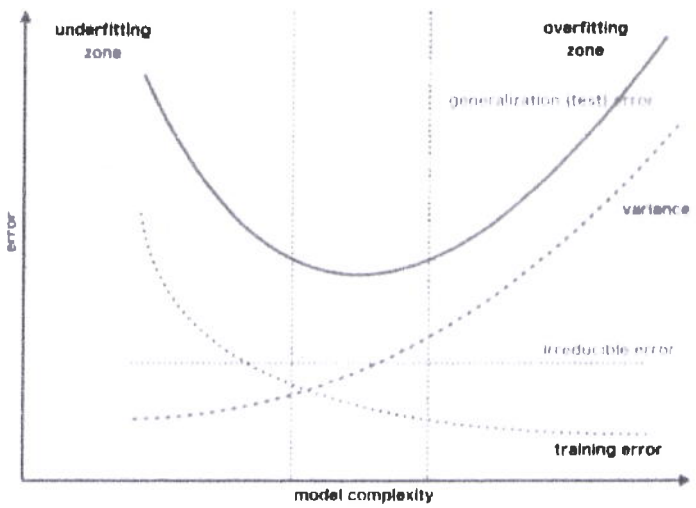
Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. **Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before.** As a result, such models perform very well on training data but has high error rates on test data.



Bias-Variance Trade-off

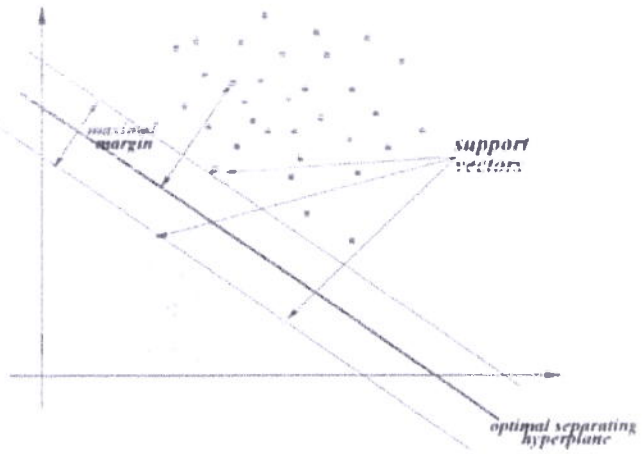
Irreducible error

This part is due to the noisiness of the data itself. The only way to reduce this part of the error is to clean up the data (e.g., fix the data sources, such as broken sensors, or detect and remove outliers).



Total Error = Bias^2 + Variance + Irreducible Error

Large Margin Classification



The SVM method can be explained with the help of the figure.

The data has two classes as noted by the colors and they are clearly linearly separable.

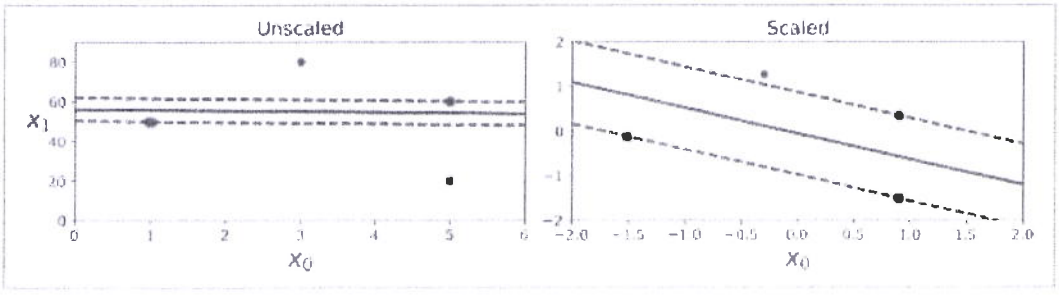
The goal of the SVM Classification challenge here will be to find the widest possible street between the classes.

This is called "**large-margin-classification**". The bold line in the middle is the "**decision boundary**".

Notice that adding more training instances "off the street" will not affect the decision boundary at all: it is fully determined (or "supported") by the instances located on the edge of the street. These instances are called the **support vectors**.

Sensitivity to Scales

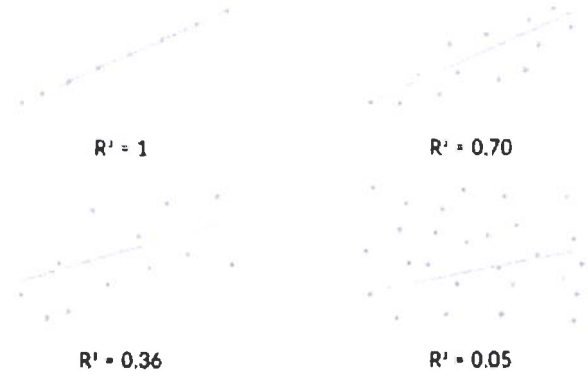
SVMs are sensitive to the feature scales, as you can see in Figure below: on the left plot, the vertical scale is much larger than the horizontal scale, so the widest possible street is close to horizontal.



Cost Functions

Residuals Calculation

Physical Significance of R^2

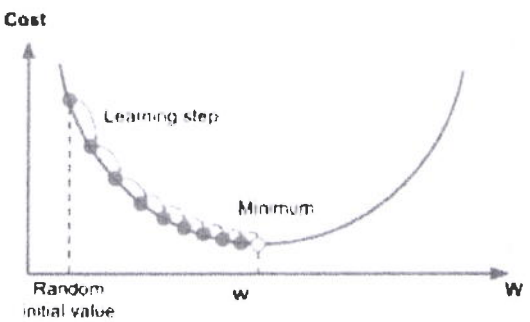


Gradient Descent

The Gradient Descent Method is an optimization technique that is used by Linear Regression Algorithms that aims at finding the best possible Linear Equation between the predictor and target variables by **minimizing the cost function**. Finding the best possible equation means finding the set of optimum β-coefficients ($\beta_1, \beta_2, \beta_3$) and intercept ($\beta_0$).

The way Gradient Descent Algorithms works is that it **assigns an initial set of values to the β-coefficients to the equation** and determines the predicted target (y) values for the input dataset.

Thereafter the **derivative of the equation is calculated**. The derivative refers to the slope of the equation at a given point. The slope will indicate which direction the values of the coefficients have to be moved in order to get a lower value of the cost function.



As the next step, Gradient Descent will **update the values of the β-coefficients towards the direction indicated by the derivativ** The update of the values is by a quantum that is known as the "**Learning Rate**" or the Alpha value. This value can be specified as part the Linear Regression as a Hyper Parameter to optimize the speed of the model building process. A higher Learning rate will make the model building faster but will have the risk of skipping the most optimum combination of coefficients. On the other hand, a lower Learning Rate will make the model building slower but will have the ability to attain a best set values for the final model.

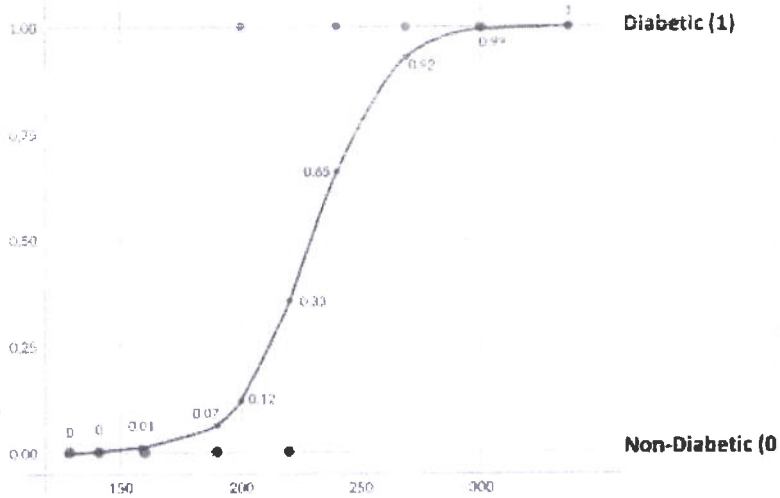
Introduction to Logistic Regression

Plotting the Probabilities

Sigmoid Curve

Sigmoid Function

$$y \text{ (Probability of Diabetes)} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$



Challenge:

How can you find the **best fit sigmoid curve?**

How to find the combination of β0 and β1 which fits the data best.

Introduction to Logistic Regression

Point no.	1	2	3	4	5	6	7	8	9	10
Diabetes	no	no	no	yes	no	yes	yes	yes	yes	yes

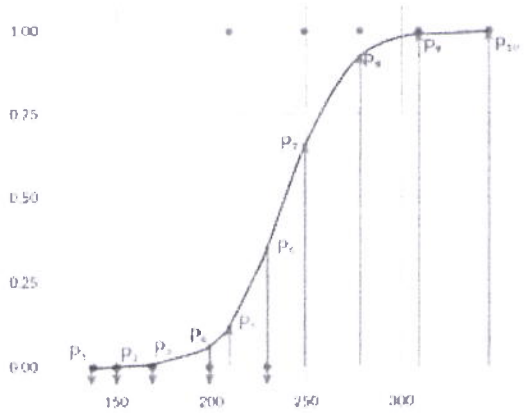
Cost Function?

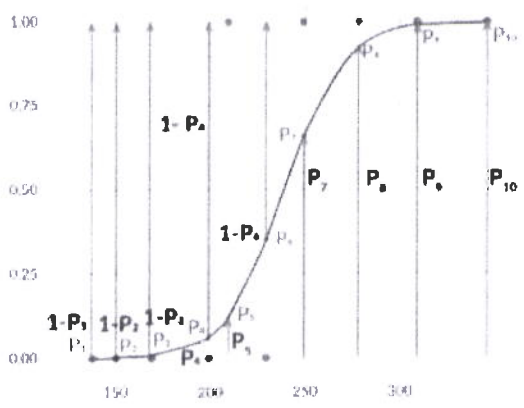
The best fitting combination of β0 and β1 will be the one which maximises the product:

(1−P1)(1−P2)(1−P3)(1−P4)(1−P6)(P5)(P7)(P8)(P9)(P10)

Maximum Likelihood Function

[(1−Pi)(1−Pi)------ for all non-diabetics --------]
** [(Pi)(Pi) -------- for all diabetics -------]*



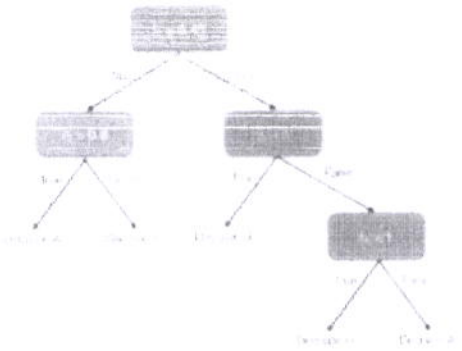


Decision Trees

Decision Trees are versatile Machine Learning algorithms that ***can perform both classification and regression tasks***, and even multioutput tasks. They are very powerful algorithms, capable of fitting complex datasets.

The goal of a Decision Tree is to create a model that predicts the value of a target variable by ***learning simple decision rules in the form of a tree inferred from the data features***.

Decision Trees are also the fundamental components of ***Random Forests****, which are among the most powerful Machine Learning algorithms available today.*



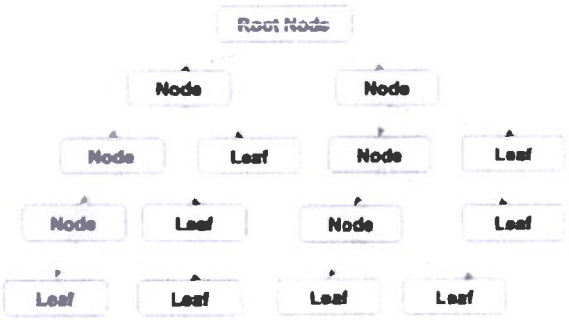
Example

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

A general algorithm for a decision tree can be described as follows:

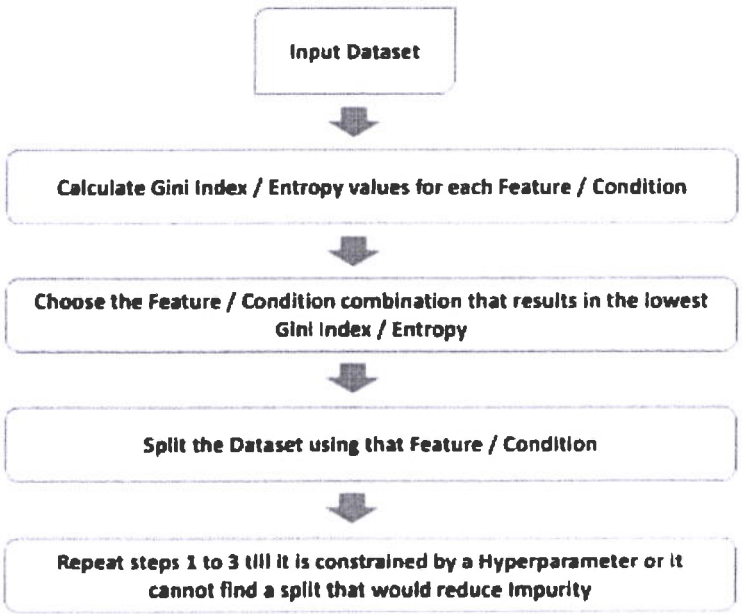
1. Pick the best attribute/feature. The best attribute is one which best splits or separates the data.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive to the answer.

Decision Tree Concepts



- **Root Node:** The representation of the Root Node is the first condition of split that the algorithm places with the aim of ultimately labelling all the observations correctly.
- **Decision Node:** Any further node down the tree other than the Root Node, where a decision of data split is being made based on a condition on one or more attributes is a Decision Node or simply a Node.
- **Splitting:** Splitting is the process of separating the input dataset at a node into multiple output sets based on the condition at that Node.
- **Leaf:** Observations that have received the target label at the end of branching through the tree / conditions are called Leaf. There is no further splitting from Leaf nodes.
- **Pruning:** Decision Trees unless controlled through Hyperparameters, have the tendency to run through the entire dataset and place / create conditions / nodes to split data to the minutest level to achieve a 100% accuracy thereby causing overfitting. Pruning is a method to limit the number of levels of a tree to avoid overfitting. There are multiple Hyperparameters that can be used to achieve this.
- **Branch / Sub-Tree:** Tree beneath a Node that's created as a result of conditions in that Node.
- **Parent and Child Node:** Nodes beneath a particular Node are child nodes and the Node in question is termed as the Parent Node.

Decision Tree Steps



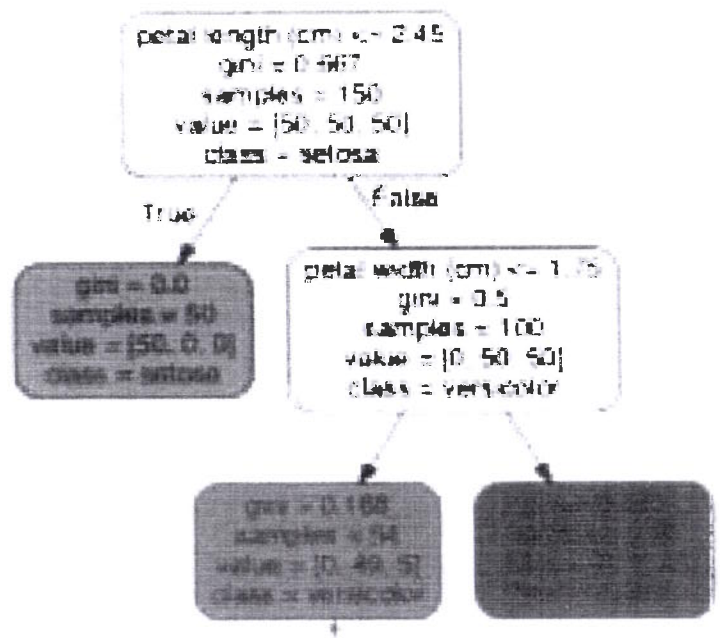
To begin with, the Decision Tree Algorithm aims to find a single Feature '***k***' and a Threshold 't_k' to split the Data.

The approach taken is to find out the Feature 'k', that yields the lowest value of ***Gini Impurity Index*** or '**Entropy**'.

Calculations from the IRIS Dataset

	sepallength	sepalwidth	petallength	petalwidth	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...					
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

Graphviz Representation



Gini Index $G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2 \Rightarrow 1 - (0/54)^2 - (49/54)^2 - (5/54)^2 = 0.168$

Entropy $H_i = - \sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log_2 (p_{i,k}) \Rightarrow -\frac{49}{54} \log_2 \left(\frac{49}{54}\right) - \frac{5}{54} \log_2 \left(\frac{5}{54}\right) = 0.445$

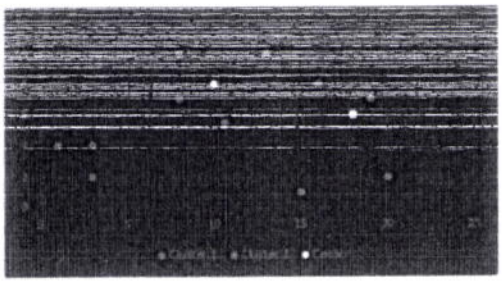
K-Means Clustering – Calculations-1

Iteration 1

X	Y	Cluster
8	10	1
20	2	2
16	8	2
8	7	1
1	4	1
13	10	1
15	1	2
19	7	2
3	4	1
3	2	1
11	6	1

Center	X	Y
1	10.0	6.0
2	18.0	4.0

SSE	143.6



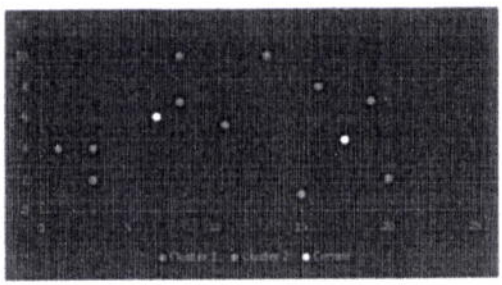
Min Squared Distance	Dist_C1	Dist_C2
8.0	2.0	10.0
20.0	14.0	4.0
8.0	6.0	2.0
5.0	2.0	9.0
9.0	9.0	17.0
13.0	3.6	6.4
14.0	8.6	5.4
2.0	9.1	1.4
65.0	8.1	25.1
85.0	9.2	15.5
6.6	2.6	7.4

Iteration 2

X	Y	Cluster
8	10	1
20	2	2
16	8	2
8	7	1
1	4	1
13	10	2
15	1	2
19	7	2
3	4	1
3	2	1
11	6	1

Center	X	Y
1	6.7	6.1
2	17.5	4.5

SSE	224.2



Min Squared Distance	Dist_C1	Dist_C2
17.2	4.2	11.0
12.5	14.0	3.5
14.5	9.5	3.8
2.2	1.6	9.8
16.3	6.0	16.5
50.5	7.5	7.1
18.5	9.8	8.3
8.5	12.4	2.9
17.7	4.2	14.5
30.0	5.5	4.7
15.9	4.0	7.0

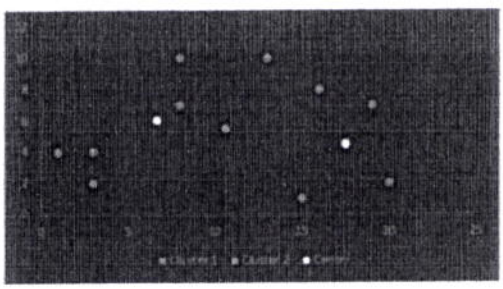
K-Means Clustering – Calculations-2

Iteration 3

X	Y	Cluster
8	10	1
20	2	2
16	8	2
8	7	1
1	4	1
13	10	2
15	1	2
19	7	2
3	4	1
3	2	1
11	6	1

Center	X	Y
1	5.6	5.4
2	16.6	5.6

SSE	204.8



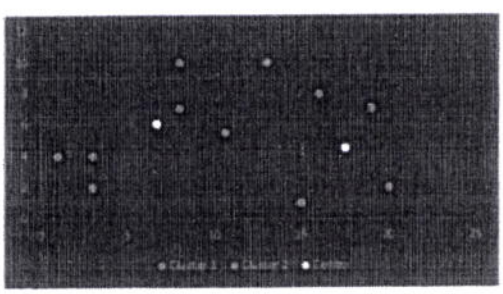
Min Squared Distance	Dist_C1	Dist_C2
26.8	5.2	9.7
24.5	14.8	5.0
6.1	10.7	2.5
8.3	2.9	8.7
23.2	4.8	15.7
32.3	8.7	5.7
23.7	10.4	4.9
7.7	13.5	2.8
8.8	3.0	13.7
18.4	4.3	14.1
25.0	5.0	6.0

Iteration 4

X	Y	Cluster
8	10	1
20	2	2
16	8	2
8	7	1
1	4	1
13	10	2
15	1	2
19	7	2
3	4	1
3	2	1
11	6	1

Center	X	Y
1	5.6	5.4
2	16.6	5.6

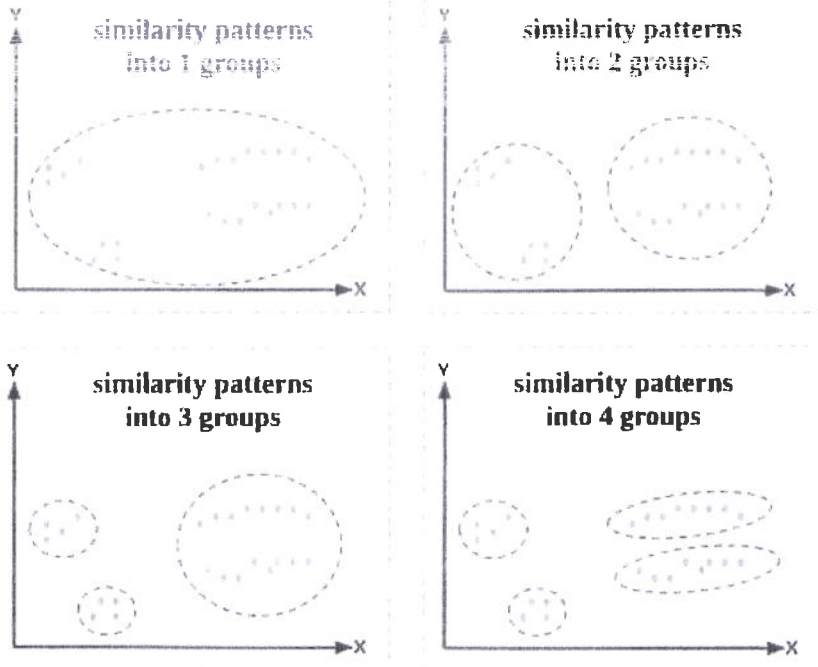
SSE	204.8



Min Squared Distance	Dist_C1	Dist_C2
26.8	5.2	9.7
24.5	14.8	5.0
6.1	10.7	2.5
8.3	2.9	8.7
23.2	4.8	15.7
32.3	8.7	5.7
23.7	10.4	4.9
7.7	13.5	2.8
8.8	3.0	13.7
18.4	4.3	14.1
25.0	5.0	6.0

K-Means Clustering

K-Means Clustering is an Unsupervised Learning Algorithm, which groups the unlabelled dataset into 'K' different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



K-Means Clustering

- The objective of K-means is to group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (*k*) of clusters in a dataset.
- A ***cluster*** refers to a collection of data points aggregated together because of certain similarities.
- We'll define a target number *'k'*, which refers to the number of ***centroids*** you need in the dataset. A centroid is the location (data point) representing the center of the cluster.
- In other words, the K-means algorithm ***identifies k number of centroids***, and then ***allocates every data point to the nearest cluster.***
- The ***'means'*** in the K-means refers to averaging of the data, that is, finding the centroid.

BSCAN (Density-Based Spatial Clustering of Applications with Noise) – imply Explained

agine you're at a party, and people are standing in groups, chatting. Some people are standing one. DBSCAN helps find groups (clusters) of people while ignoring the loners (noise).

ow DBSCAN Works (Step-by-Step Story)

1. Finding Crowds (Dense Areas)
   - DBSCAN looks for people standing close together.
   - If enough people are nearby, it considers this a group (cluster).
2. Expanding the Group
   - It checks if those people have even more people nearby and expands the group.
   - This continues until there are no more close people to add.
3. Ignoring the Loners (Noise)
   - Some people are standing too far from any group.
   - DBSCAN ignores them as noise because they don't belong to any dense region.

ey Features of DBSCAN

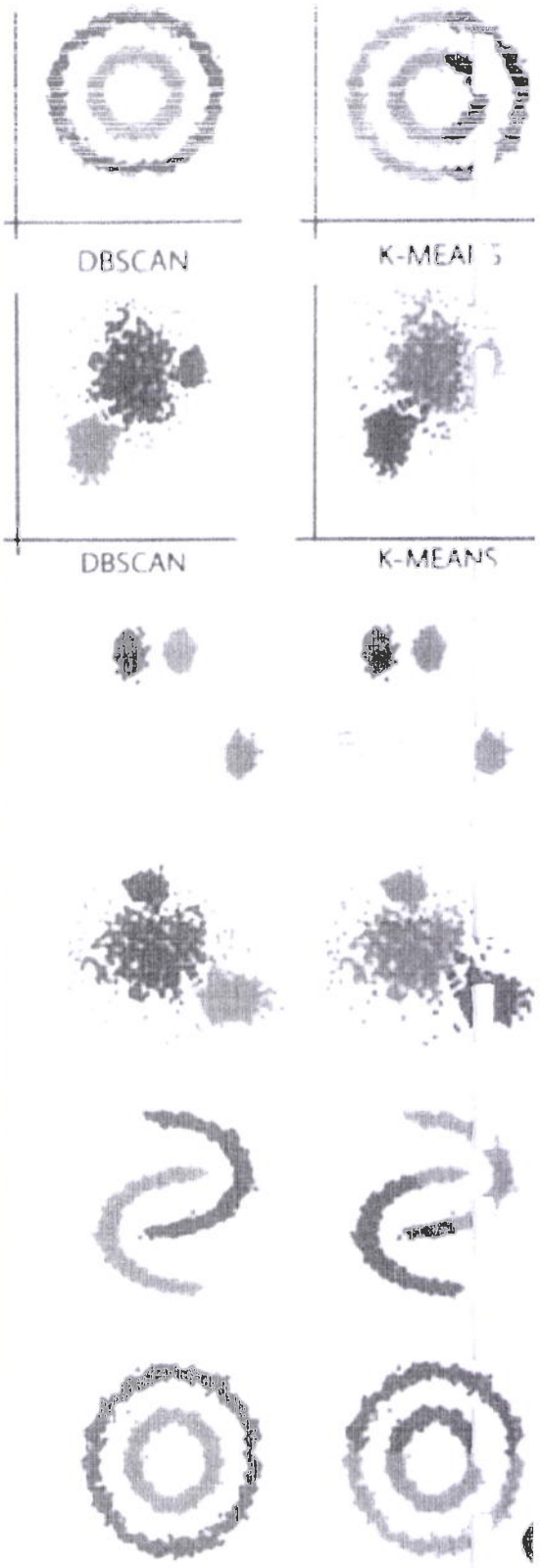
- Finds clusters of any shape (good for real-world data).
- Doesn't need to know the number of clusters beforehand (unlike K-Means).
- Automatically ignores outliers (noise).
- Works well when clusters have different sizes and densities.

omparison with K-Means

Feature	DBSCAN	K-Means
Cluster Shape	Any shape	Circular
Needs Number of Clusters?	✗ No	✓ Yes
Handles Outliers?	✓ Yes	✗ No
Good for Noisy Data?	✓ Yes	✗ No

hen to Use DBSCAN?

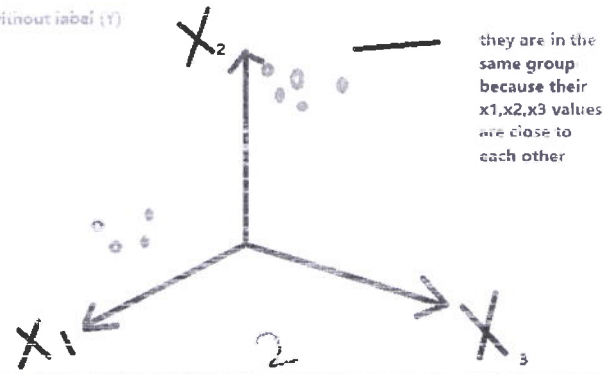
- When you don't know the number of clusters.
- When clusters have different shapes and densities.
- When your data has outliers (and you want to ignore them).



DataSet

x1	x2	x3

features without label (Y)



they are in the same group because their x1,x2,x3 values are close to each other

la to Find Distance Between Points in K-Means

ans Clustering, we use the **Euclidean Distance** to measure the distance a data point X (with coordinates x_1, x_2, x_3) and a cluster center C (with ates c_1, c_2, c_3).

n Distance Formula for 3D Space:

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + (x_3 - c_3)^2}$$

tion:

x_2, x_3 → Coordinates of the data point.

c_2, c_3 → Coordinates of the cluster center.

→ Distance between the point and the cluster center.

mula extends to **N dimensions** as:

$$d = \sqrt{\sum_{i=1}^{N} (x_i - c_i)^2}$$

Key Properties of an Ideal Clustering:

1. **High Intra-Cluster Similarity**
   - Data points **within the same cluster** should be **very close** to each other.
   - This ensures that each group consists of similar data points.
2. **High Inter-Cluster Separation**
   - Data points from **different clusters** should be **far apart** from each other.
   - This ensures that distinct groups are well-separated and meaningful.

Goal of Clustering:

- **Maximize similarity** within clusters.
- **Minimize similarity** between clusters.

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

3

ns is an iterative clustering algorithm that groups similar data points into **k clusters**. Let's e **k = 3** for this explanation.

of K-Means Clustering:

ialize Centroids

ndomly select **3 data points** from the dataset.

ese points act as the initial **centroids** of the clusters (C_1, C_2, C_3).

ign Data Points

r each remaining data point, **calculate the distance** from all centroids (C_1, C_2, C_3).

sign each data point to the **nearest centroid** (i.e., closest cluster).

date Centroids

mpute the **new centroid** for each cluster by taking the **average position** of all points in that ster.

ese new centroids are now C_1(new), C_2(new), C_3(new).

peat Until Convergence

back to **Step 2 & Step 3**, reassigning points and recalculating centroids.

ntinue this process **until the centroids stop changing** (i.e., convergence is reached).

Notes:

ures **data points in the same cluster are close to each other.**

ures **clusters are well-separated.**

e final clusters depend on the **initial random selection of centroids** (which is why K-Means++ used for better initialization).

K-Means++: A Smarter Way to Select Initial Centroids

Problem with Random Selection:

- In traditional K-Means, centroids are chosen randomly, which can lead to **poor clustering** and **slow convergence** if the starting points are not well spread out.

Solution: K-Means++

- K-Means++ improves the initialization by **choosing centroids that are far apart**, leading to **better and faster clustering**.

Steps of K-Means++ Initialization

1. **Select the First Centroid Randomly**
   - Choose **one** data point randomly from the dataset as the **first centroid**.
2. **Compute Distances for the Next Centroid**
   - For each remaining data point:
     - Calculate the **distance** from the **nearest** selected centroid.
     - The farther a point is from existing centroids, the **higher its chance of being selected** as next centroid.
3. **Select the Next Centroid Based on Probability**
   - Choose the next centroid **randomly but with weighted probability**:
     - Data points **farther away** from existing centroids are **more likely** to be selected.
4. **Repeat Until k Centroids Are Chosen**
   - Keep repeating **Step 2 & 3** until all **k centroids** are selected.

Why K-Means++ is Better?

- **Prevents bad starting positions** (reduces poor clustering).
- **Faster convergence** (reduces the number of iterations).
- **More stable results** (avoids local minima in optimization).

C.

ITEC

सत्यमेव जयते
Ministry of External Affairs
Government of India

सी डैक
CDAC

INDIAN TECHNICAL AND ECONOMIC COOPERATION

Government of India

MRS. NIDHAL HUSSEIN M.ALI MOHAMMED ALI

is hereby awarded the certificate in recognition of
successful participation in ITEC Programme

Specialised Programme on Machine Learning

(22nd January 2025 – 04th February 2025)

organised by

Centre for Development of Advanced Computing, Noida

under the auspices of the

ITEC Programme, Development Partnership Administration,
Ministry of External Affairs, Government of India.

Programme Coordinator	Scientist 'G' & GC	Executive Director

ITEC Certificate number **2024IRQ000302**. The genuineness of this certificate may be verified at www.itecgoi.in using this certificate number

प्रगत संगणन विकास केन्द्र
CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING
इलेक्ट्रोनिकी एवं सूचना प्रौद्योगिकी मंत्रालय, भारत सरकार की एक वैज्ञानिक संस्था
A Scientific Society of the Ministry of Electronics and Information Technology, Govt. of India

सी डैक
CDAC

अनुसंधान भवन
सी-56/1 संस्थागत क्षेत्र, सैक्टर-62
नोएडा-201309 (उ.प्र.) भारत
Anusandhan Bhawan,
C-56/1, Institutional Area, Sec-62
Noida-201309 (U.P.) India
फोन / Tel +91-120-2210800
www.cdac.in

Ref No.: 7(2)/2024-E&T

February 04, 2025

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **MRS. NIDHAL HUSSEIN M.ALI MOHAMMED ALI** from **IRAQ** had come to India under **ITEC** scheme of **Ministry of External Affairs, Government of India,** for undergoing **"Specialised Programme on Machine Learning"** from **22nd January 2025 to 04th February 2025** at **Centre for Development of Advanced Computing (C-DAC), NOIDA (Ministry of Electronics & IT, Govt. of India),** B-30, Institutional Area, Sector-62, NOIDA.

MRS. NIDHAL HUSSEIN M.ALI MOHAMMED ALI has successfully completed the above training and following modules have been covered during the course:

S.NO.	MODULE NAME
1	Types of Machine Learning, Supervised Learning , Unsupervised Learning, Supervised Learning with Trees, Introduction to Neural Network etc.

(Navneet Jain)
Programme Coordinator

प्रगत संगणन विकास केन्द्र
CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING
इलेक्ट्रोनिकी एवं सूचना प्रौद्योगिकी मंत्रालय, भारत सरकार की एक वैज्ञानिक संस्था
A Scientific Society of the Ministry of Electronics and Information Technology, Govt. of India

सी डैक
CDAC

अनुसंधान भवन
सी-56/1 संस्थागत क्षेत्र, सैक्टर-62
नोएडा-201309 (उ.प्र.) भारत
Anusandhan Bhawan,
C-56/1, Institutional Area, Sec-62
Noida-201309 (U.P.) India
फोन / Tel +91-120-2210800
www.cdac.in

Ref. 7(2)/2024-E&T

January 27, 2025

TO WHOMSOEVER IT MAY CONCERN

This is to certify that **MRS. NIDHAL HUSSEIN M.ALI MOHAMMED ALI** (Passport No.: B18293659) of **IRAQ** is a bonafide student of International Training Programme titled "**Specialised Programme on Machine Learning**" under the **ITEC** scheme of the Ministry of External Affairs, Government of India during **January 22, 2025 to February 04, 2025** at C-DAC, Noida. During this period **MRS. NIDHAL HUSSEIN M.ALI MOHAMMED ALI** will be residing at **Room No. 507, M/s Trot Hotels (Le Crescent), Indirapuram, Ghaziabad.**

(Navneet Jain)
Program Co-ordinator
Email- navneetjain@cdac.in
Mb.-9811900728