Chapter 4: Fejlett adatelemzés és a Machine Learning

Big Data & Analytics

Cisco | Networking Academy®
Mind Wide Open™

Chapter 4 - Sections & Objectives

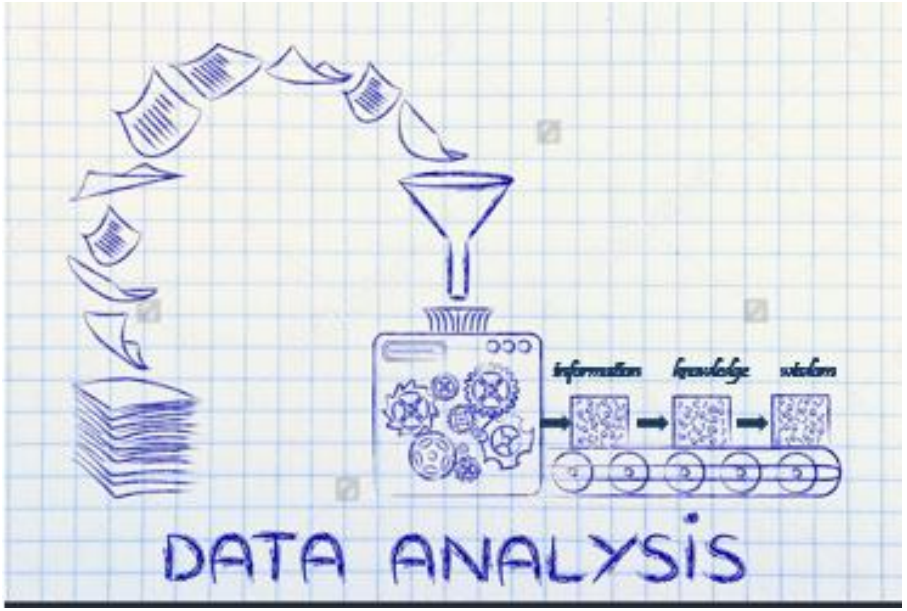
- 4.1 Prediktív analitika
  - A jövőbeli eredmények valószínűségének azonosítása adatok, statisztikai algoritmusok és gépi tanulási technikák felhasználásával, múltbeli adatok alapján.
- 4.2 Modell értékelése
  - A prediktív analitikában használt különböző értékelési mérőszámok vizsgálata.
- 4.3 Felkészülés a 4. fejezet laborjaira

4.1 Prediktív analitika Analytics

Cisco | Networking Academy®
Mind Wide Open™

Looking Ahead

- Jellemzők, amelyek megkülönböztetik a Big Data-t az adatoktól:
  - Volume
  - Velocity
  - Variety
  - Veracity
- A Big Data-t olyan előrejelző modellek létrehozására használják, amelyek választ adnak a következőkre:
  - Mi fog történni?
  - Hogyan cselekedjünk?

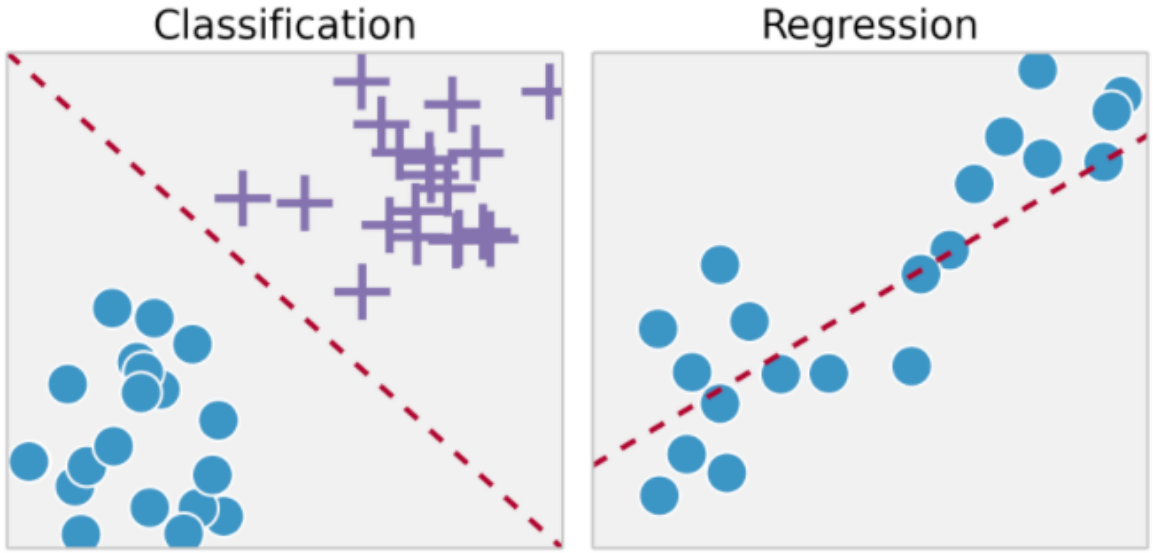


Mi a gépi tanulás?

- Kevin Patrick Murphey meghatározása szerint a gépi tanulás "...olyan módszerek összessége, amelyek képesek automatikusan mintákat felismerni az adatokban, majd a feltárt mintákat felhasználni a jövőbeli adatok előrejelzésére, vagy más típusú döntéshozatalra bizonytalanság mellett".
- A gépi tanulási algoritmusok az adott feladatok ismételt elvégzése alapján javítják teljesítményüket bizonyos feladatokban. A gépi tanulási módszereket számos alkalmazásban alkalmazzák, többek között a beszédfelismerésben, az orvosi diagnosztikában, az önvezető autókban, az értékesítési ajánlómotorban és sok másban.

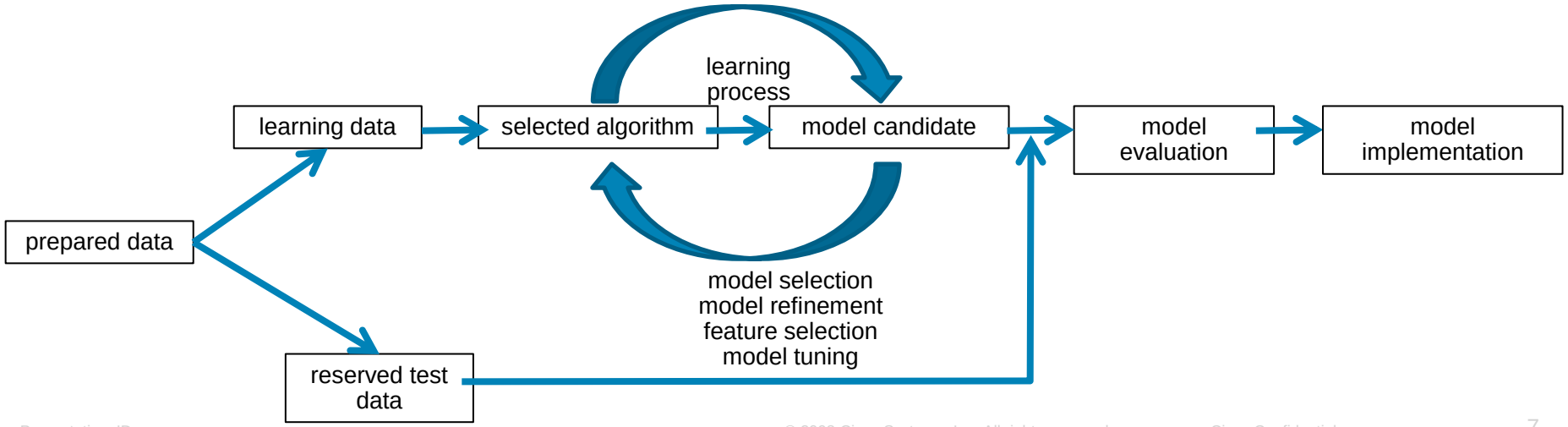
A gépi tanulási elemzés típusai

- A gépi tanulási algoritmusok két fő kategóriája:
  - Supervised – általánosan használt prediktív analitika. A regressziós és osztályozási problémák megoldására használják.
  - Unsupervised – önállóan fedeznek fel mintákat az adatokban. A nem felügyelt módszerekkel megoldott problémák példái a klaszterezés és az asszociáció.



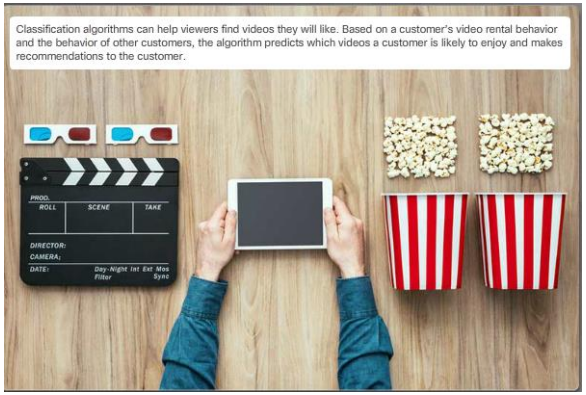
A Machine Learning Process

- A gépi tanulási megoldások fejlesztése a következő lépésekre egyszerűsíthető le:
  - Step 1 – Az adatok előkészítése
  - Step 2 – Hozzon létre egy tanulási készletet
  - Step 3 – Hozzon létre egy tesztkészletet
  - Step 4 – Hozzon létre egy hurkot
  - Step 5 – a megoldás tesztelése
  - Step 6 – A megoldás végrehajtása



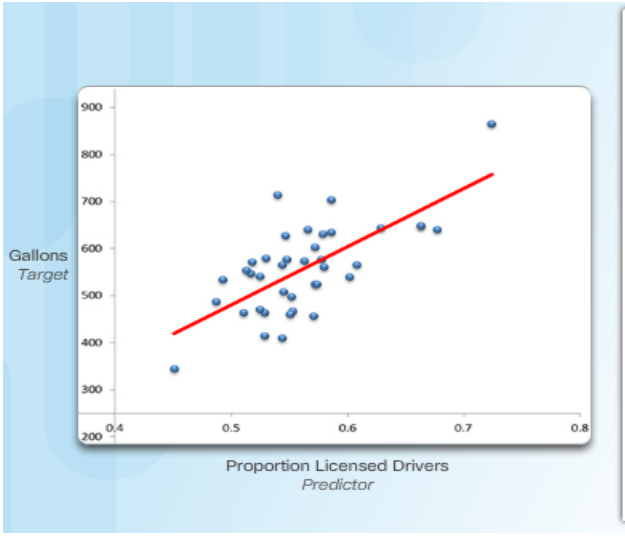
Common Applications of Machine Learning

- A prediktív analitikai algoritmusoknak számos alkalmazási területe van, többek között a szórakoztatóipar, a mezőgazdaság, az orvostudomány és a kiskereskedelmi értékesítés területén is alkalmazzák az analitikai technológiát.



Regression Analysis

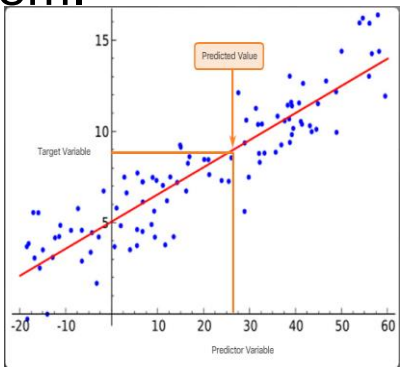
- A regresszióelemzés az egyik legrégebbi és leggyakrabban használt statisztikai módszer az adatok elemzésére.
- A regresszió fő célja egy vagy több független változó (prediktor változó(k)) és egy függő változó (célváltozó) közötti matematikai kapcsolat meghatározása.

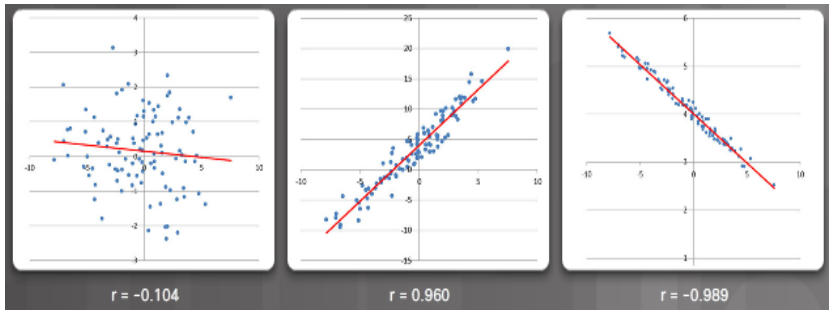


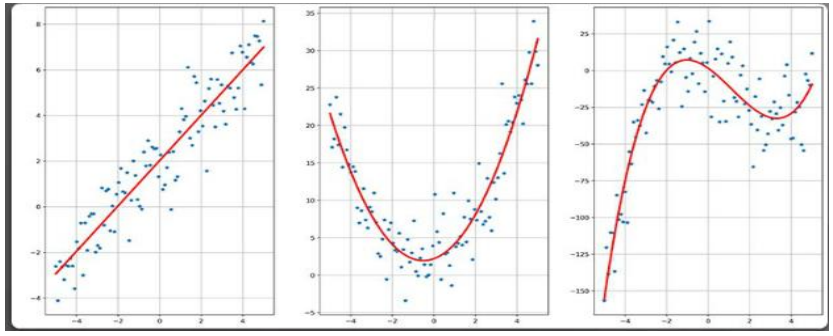
Proportion of Population with Driver's Licenses	Consumption of Gasoline (Millions of Gallons)
0.525	541
0.572	524
0.58	561
0.529	414
0.544	410
0.571	457
0.451	344
0.553	467
0.529	464
0.552	498
0.53	580
0.525	471
0.574	525
0.545	508
0.608	566
0.586	635

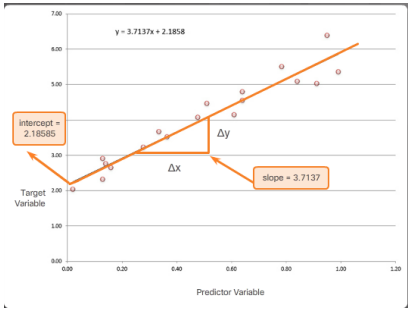
Linear Regression

- A lineáris regressziók a legegyszerűbbek mind számítási, mind matematikai szempontból.
  - A lineáris kifejezés azt jelenti, hogy a regressziós függvény mindig más függvények súlyozott átlagának felhasználásával próbál illeszkedni az adatokhoz, függetlenül attól, hogy ezek a függvények lineárisak-e vagy sem.







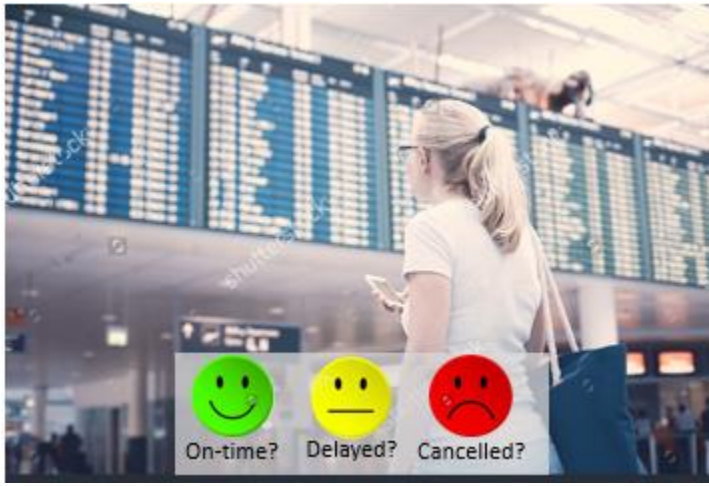


Applications of Regression Analysis

- A regressziós elemzésnek számos alkalmazása van. Gyakran használják az üzleti és pénzügyi elemzésben a múltbeli adatokkal, hogy tájékoztassák a jövőbeli stratégiákról.
- Felhasználható a gazdasági tendenciák előrejelzésére és a gazdasági növekedés irányítására irányuló politikai intézkedésekhez.
- Az ügyfelek viselkedését is meg lehet előre jelezni, hogy meghatározzák a normális és az esetlegesen csalárd viselkedést a biztosítás és a fogyasztói hitelek területén.

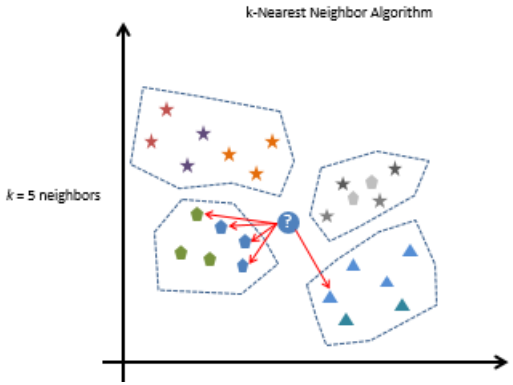
Classification Problems

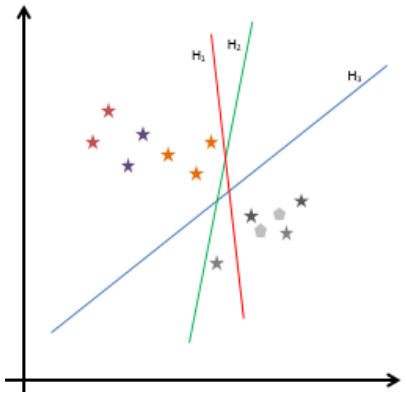
- Az osztályozás regressziós problémának tekinthető, ahol a célváltozó diszkrét, és egy olyan osztályt képvisel, amelybe egy emberi szakértő besorolta az adatmintát.
  - Például egy webes utazási vállalat szeretne megbízhatósági minősítést adni az általa az ügyfelek számára talált járatokra vonatkozóan. A különböző modellek kipróbálásával sikerült meghatározni, hogy az adathalmazban szereplő változók közül melyek a legrelevánsabbak az osztályozások szempontjából. Ezt úgy is nevezik, hogy melyek a legnagyobb diszkriminatív erővel rendelkező változók. Csak ezeket a releváns jellemzőket vonjuk ki az adatokból, és használjuk fel az osztályozó tréningjéhez.

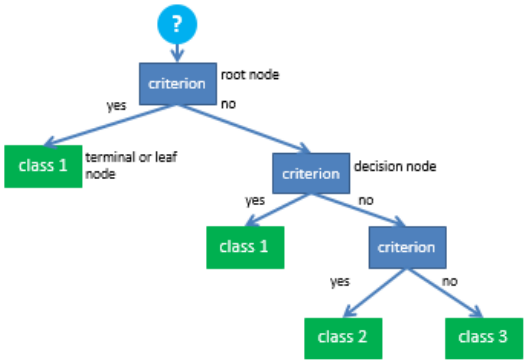


Classification Algorithms

- **k-nearest neighbor (k-NN)** - A k-NN valószínűleg a legegyszerűbb osztályozó, amely a gyakorló példák közötti távolságot használja a hasonlóság mértékeként. A k-NN osztályozó működésének szemléltetéséhez képzeljük el, hogy minden mintának két jellemzője van, amelyek értékei egy 2D-s ábrán ábrázolhatók.
- **Support vector machines (SVM)** - A támogató vektor gépek (SVM) a felügyelt gépi tanulási osztályozók példái. Ahelyett, hogy a kategóriához való tartozást más pontoktól való távolságokra alapoznák, a támogató vektor gépek kiszámítják azt a határt vagy hipersíkot, amely jobban elválasztja a csoportokat.
- **Decision trees** - A döntési fák az osztályozási problémát a jellemzők értékei alapján hozott döntések halmazaként ábrázolják. A fa minden egyes csomópontja egy jellemző értékének küszöbértékét jelenti, és a képzési mintákat két kisebb halmazra osztja fel.







Applications of Classifications

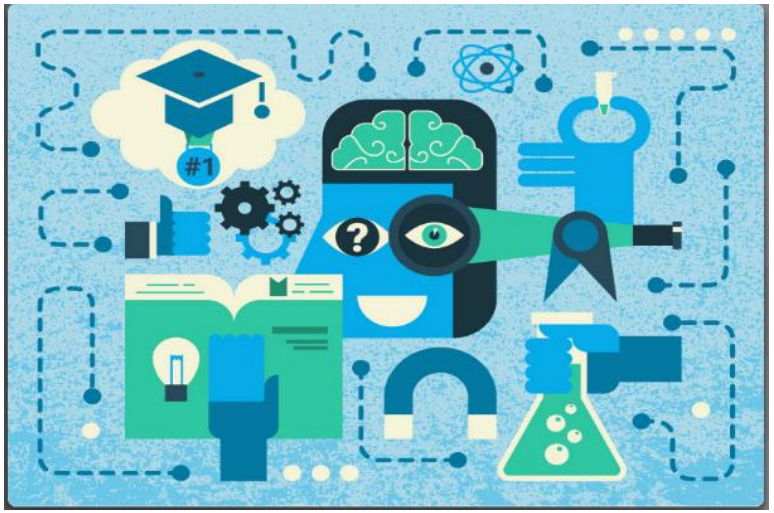
- Az osztályozási algoritmusoknak számos alkalmazása van. Például:
  - **Risk Assessment** - Az osztályozási rendszerek segítségével meghatározható, hogy számos tényező közül melyik járul hozzá a különböző kockázatok valószínűségéhez.
  - **Medical Diagnostics** - Az osztályozó rendszerek az irányított kérdések segítségével olyan döntési fát állíthatnak össze, amely segíthet a különböző betegségek és a betegségek kockázatának diagnosztizálásában.
  - **Image Recognition** - A kézírás-felismerés során egy rendszer a kézzel írt számjegyek azonosításán dolgozhat.

4.2 Model Evaluation
Modell értékelése

Cisco | Networking Academy®
Mind Wide Open™

Issues in Using analysis

- A tudományos felfedezés hat lépésből álló folyamata a következő:
  - Ask a question about an observation
  - Kutatás végzése.
  - Alakítson ki hipotézist
  - A hipotézis tesztelése
  - lemezze a kísérletekből származó adatokat a következtetés levonásához.
  - Az eredmények kommunikálása



Validity

- Bár az érvényesség típusainak leírására számos kifejezést használnak, a kutatók jellemzően négyféle érvényességet különböztetnek meg:
  - **Construct validity** - Does the study actually measure what it claims to measure?
  - **Internal validity** - Was the experiment designed correctly? Does it include all the steps of the scientific method?
  - **External validity** - Can the conclusions apply to other situations or other people in other places at other times? Are there any other causal relationships in the study that might account for the results?
  - **Conclusion validity** - Based on the relationships in the data, are the conclusions of the study reasonable?

Reliability

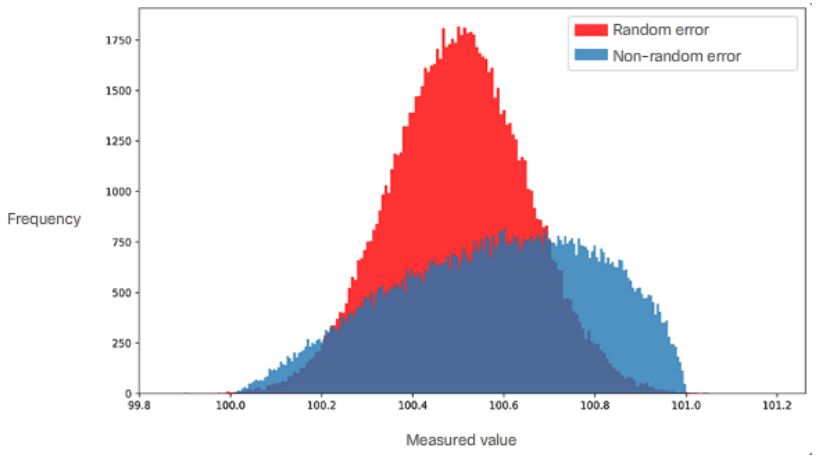
- A Reliable experiment or study means that someone else can repeat it and achieve the same results. Researchers distinguish between four types or reliability:
  - **Inter-rater reliability** - How similarly do different people score on the same test?
  - **Test-Retest Reliability** - How much variation is there between scores for the same person taking a test multiple times?
  - **Parallel-Forms Reliability** - How similar are the results of two different tests that are constructed from the same content?
  - **Internal Consistency Reliability** - What is the variation of results for different items in the same test?

Error in Data Analytics

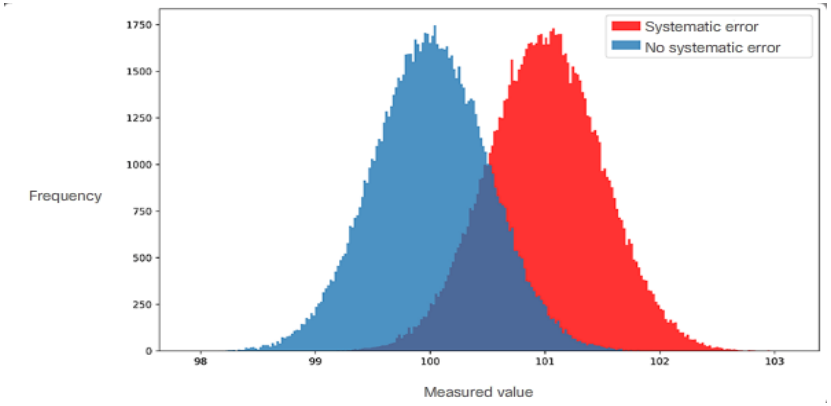
- Errors, and more in general, uncertainty, affect the data analytics process at different levels:
  - The first type of error is the **measurement error**. Any device for taking measurements is limited in its precision. Therefore, all measurements have a built-in error component.
  - Another type of error is the **prediction error**. In supervised learning, the prediction error is quantified as the difference between the value predicted by the model and the observed value.

Types and Sources of Measurement Error

- Measurement errors can be categorized into these three groups:
  - **Gross errors** - These are caused by a mistake in the instrument being used to take the measurement, or in recording the result of the measurement.
  - **Random errors** – These are caused by factors that randomly impact the measurement over a sample of data.
  - **Systematic errors** – These are caused by instrumental or environmental factors that impact all measurements taken over a given period of time.



Random errors



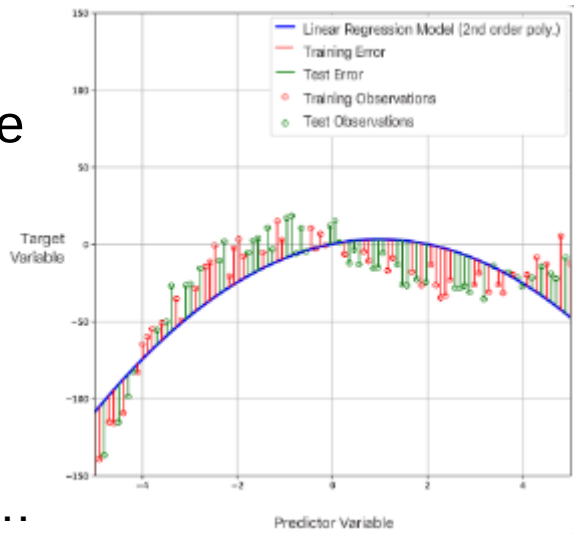
Systematic errors

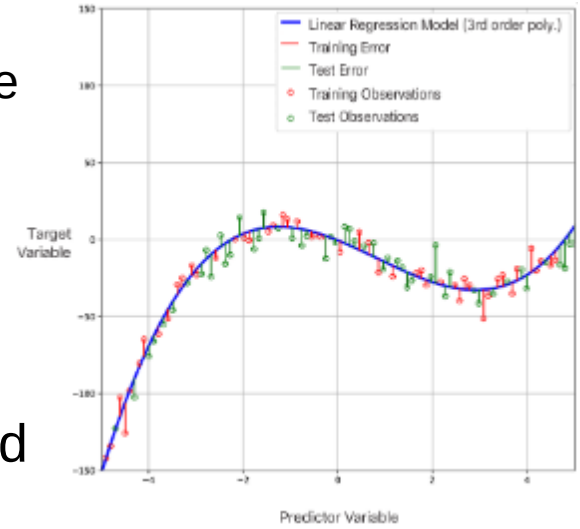
Random Error Distribution

- **Random errors** tend to create a normal distribution around the mean of the observation. It is possible to build a statistical model of the error, in which case regression and classification algorithms can easily take it into account.
- **Systematic errors** tend to shift the distribution of the observations (right side of the figure) in one direction or another. A systematic error is therefore harder to deal with, because the true value is not known, so the only way to detect a systematic error is to use another measurement system that we deem more reliable.

Errors in Predictive Analytics

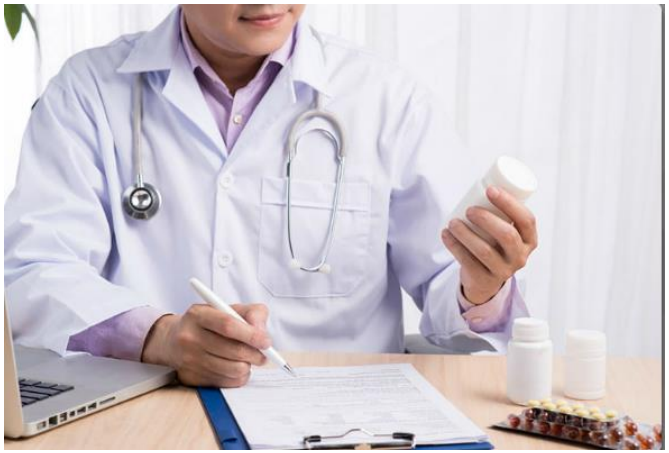
- **Prediction error is a difference** between the value predicted by the regression or classification model, and the measured value.
- **Prediction error is the distance** between the regression function, and the data points. The prediction error has **two components**
  - The first component is caused by the choice of model… we make an assumption on how the data is distributed, which is inevitably an approximation.
  - Even when the chosen model perfectly reflects the true distribution, there will still be differences between predicted and actual values because of the measurement error.
- In machine learning, the first cause of prediction error is often called **bias** of a model, while the second is **variance**. One cannot minimize both, and this situation is often called the **bias-variance tradeoff.**





Misleading Research

- Az érvényesség, a megbízhatóság és a hibák hatásának megértése egy adatmintában fontos első lépés annak biztosításához, hogy következtetései szilárd kutatási terven alapuljanak..
- Misleading, bad, or erroneous research is more common than you may think. In fact, John P.A. Ioannidis states that most research findings are false.



Guidelines for Evaluating Results

- There are several guidelines you can following when evaluating the results reported by a research study or a data analysis report:
  - **Statistics** - Does the study have a large enough sample size to support the findings?
  - **Research design** - Did the architects of the study follow generally accepted methods of research design?
  - **Duration** - Does the research appropriately account for the impact on time?
  - **Correlation and causation** - Just because two variables are correlated does not mean that one caused the other.
  - **Alignment to other studies** - Do the results confirm or align with other studies in the field?
  - **Peer review** - Has the study been reviewed by experts in the same field?

4.3 Preparation for Chapter 4 Labs

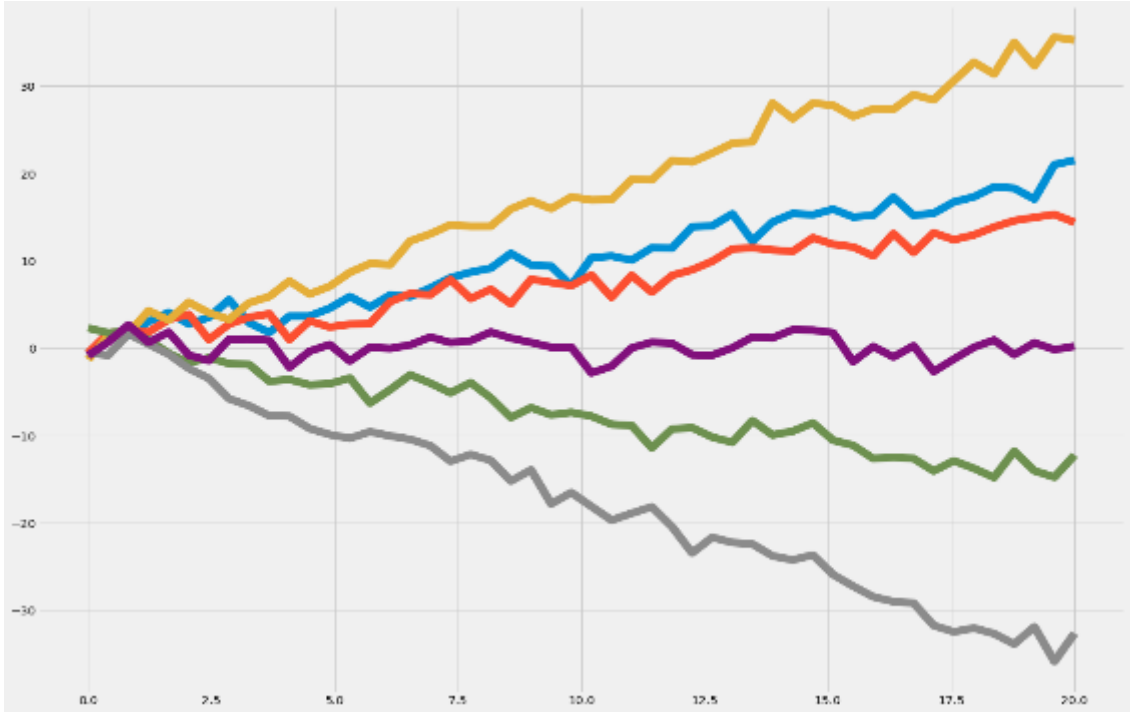
Cisco | Networking Academy®
Mind Wide Open™

Using scikit-learn for Regression Analysis

- **scikit-learn is a machine learning library for Python** built on NumPy, SciPy, and matplotlib
- Az első laborban regresszióelemzéssel tekintheti meg az internetes forgalom növekedésére vonatkozó historikus adatokat. Számszerűsíteni fogja az év és az internetes forgalom mérése közötti kapcsolatot.

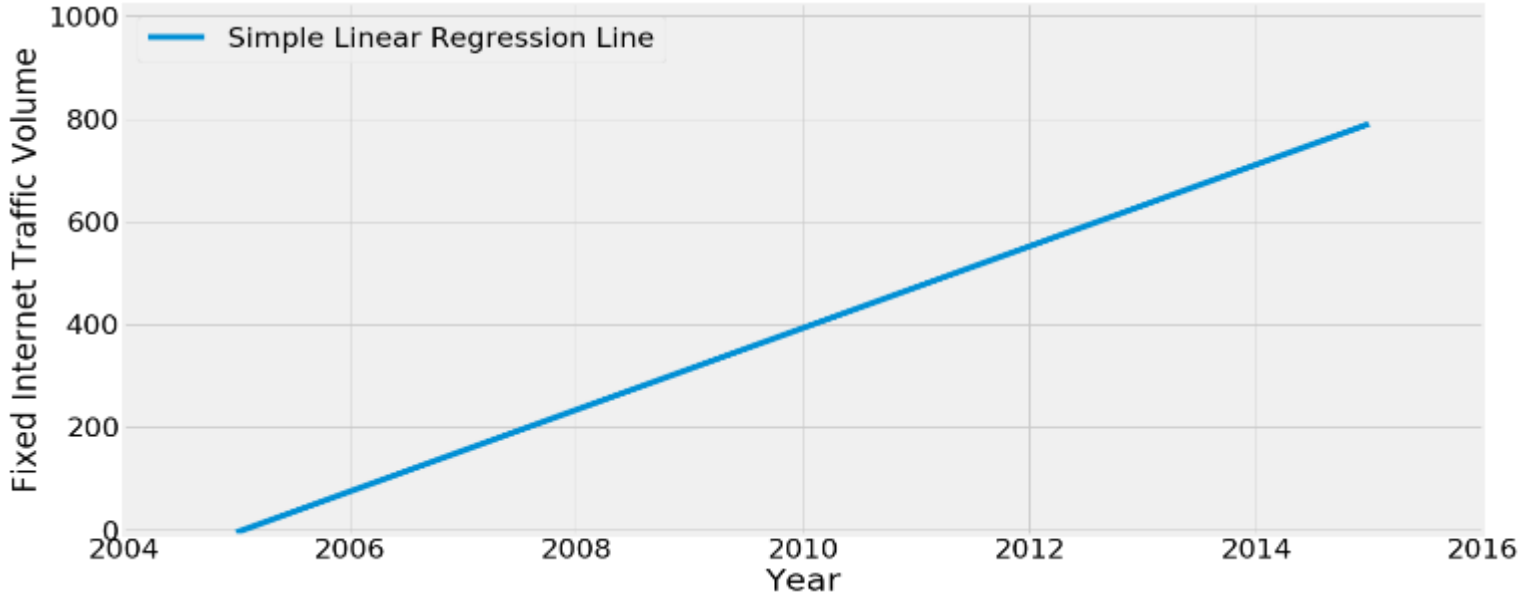
Style Sheets for Plots

- Telepíteni fogja a pandas, numpy és matplotlib programokat. A matplotlib könyvtár különböző stílusokat tartalmaz a grafikonok megjelenítéséhez.



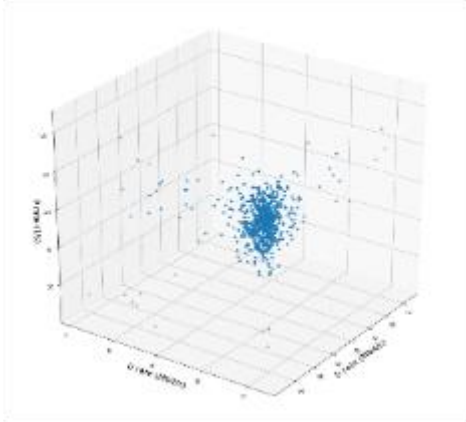
Fitting the Data

- A lineáris regresszió elvégzéséhez Pythonban a Numpy osztályát, a polyfit-et hívja meg. Bár a polyfitnek sok argumentuma van, csak az x, y és deg értékeit fogod meghatározni. Az x és y értékét fogjuk használni az x és y tengelyhez. A polyfit használatával az ábrán látható egyszerű lineáris regresszió ábrázolását végezheti el. A deg értéke határozza meg az illeszkedés mértékét...



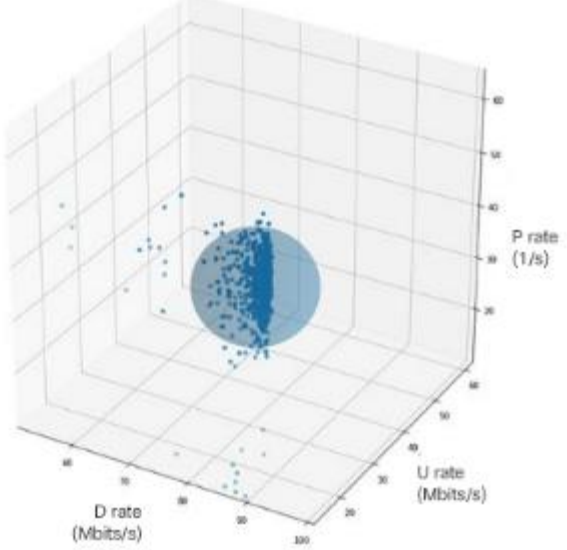
Plotting in 3D

- Az adatokat három dimenzióban fogja megjeleníteni. Ehhez az mplot3d könyvtár mpl_toolkits osztályának telepítésével bővíted a matplotlib könyvtárat. Ezután az internetmérő adataiból egy 3D-s ábrát készítesz, amely három tengelyen jeleníti meg a következő adatokat: letöltési sebesség (x tengely); feltöltési sebesség (y tengely); és ping sebesség (z tengely). Ez a vizualizáció azt fogja megjeleníteni, hogy a legtöbb pingelés sebessége hol csoportosul.Translated with www.DeepL.com/Translator (free version)



Visualizing the Boundary for Anomalies

- Az adatanomáliákat okozhatják a mérés, az átvitel vagy a tárolás során bekövetkező sérülések vagy torzulások. Ezeket az értékeket kiugró értékeknek tekintjük. Olyan mértékben eltérnek a várt értékektől, hogy torzíthatják az elemzés eredményeit.
- Az anomáliákat gyakran gondos mérlegelés után eltávolítják az adathalmazból.
- A gömb mutatja a normál adatok és az anomális adatok közötti döntési határt.



4.4 Summary

Cisco | Networking Academy®
Mind Wide Open™

Summary

- A Big Data-t a mennyiség, a sebesség, a változatosság és a hitelesség jellemzi.
- Példák a felügyelt gépi tanulási megközelítésekre, azaz: regresszió és osztályozás.
  - A regresszió egy vagy több független változó és egy függő változó közötti múltbeli kapcsolatot használ a függő változók jövőbeli értékeinek előrejelzésére.
  - Az osztályozási modellek osztályozóként ismertek. Számos osztályozó algoritmus létezik. Példa: k-közelebbi szomszéd, támogató vektorgép és döntési fa.
- A fejezet a tudományos módszer által az értékelési modell validálására használt hatlépcsős folyamatot tárgyalja.
- Az érvényesség négy típusa a következő: konstrukció, belső, külső és következtetés.
- A megbízhatóság négy típusa a következő: inter-rater, teszt-reteszt, párhuzamos formák és belső konzisztencia.
- A hiba a megfigyelés tényleges értéke és a mért érték közötti különbség.