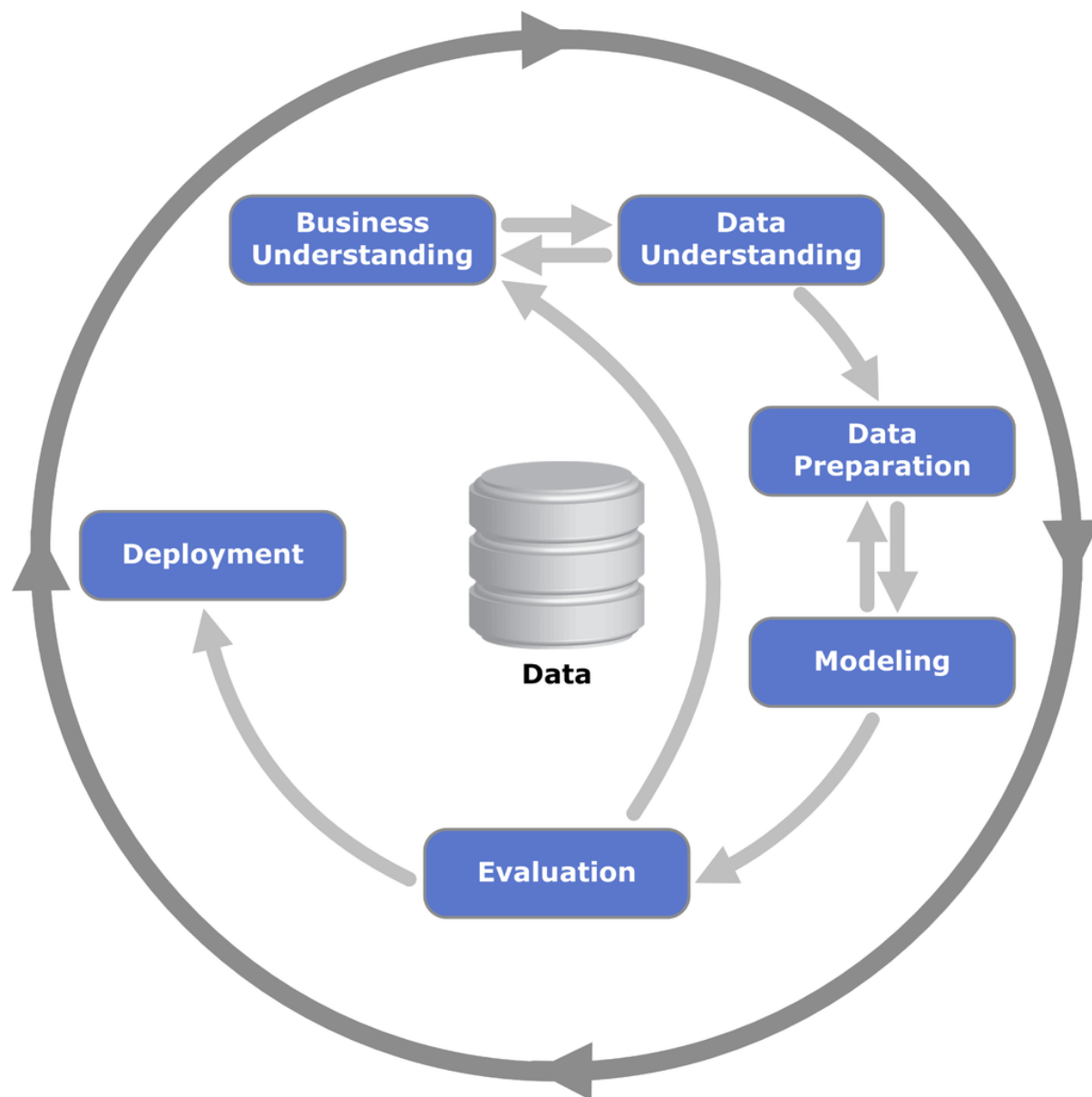


1 Metodologie CRISP-DM. Exemple de proiecte DS

Proiectele de DS sunt diverse si greu de prins intr-o schema unitara. Totusi, exista un consens larg asupra pasilor care trebuie sa se execute intr-un astfel de proiect. Mai mult, exista o metodologie care da secventa de pasi de urmat. Metodologia se numeste CRISP-DM (Cross-industry standard process for data mining). Este un standard deschis, convenind pasii pe care expertii in Data Mining ii urmeaza. Conceput in 1996, a fost extins in 2015 de catre IBM prin Analytics Solutions Unified Method for Data Mining/Predictive Analytics (ASUM-DM).



Sursa: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining#/media/File:CRISP-DM_Process_Diagram.png

Sunt 6 faze majore intr-un proces de DS/DM:

1. Business understanding



2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

▼ 1.1 Pasul 1. Determinarea obiectivelor de business (Determine business objectives)

1. Ce se dorește de la proiect?

- setarea obiectivelor - traducerea din perspectiva de business în obiective data science. De exemplu, se dorește determinarea tendinței clienților de a se muta la un competitor; decidera asupra căror clienți li se va adresa o anumită campanie (cui trimiteți pliante/solicitări de donații etc.)
- perspectiva DS: specificarea tipului de problemă: clasificare, regresie, descriere, sumarizare, grupare, detectare de outliers etc.
- definirea metricilor de succes - cum se decide dacă proiectul se încheie cu succes sau nu, metode măsurabile de cuantificare a profitului, pe scurt măsurarea impactului procesului (parte din Key Performance Indicators, KPIs)
- producerea unui plan de acțiune - pași care urmează să se execute în restul proiectului, inclusiv alegerea uneltelor și a tehnicilor folosite

2. Estimarea situației curente

- inventar de resurse: personal (experti în domeniu, pentru domain knowledge; suport tehnic; experți ML/DS, ingineri DS); date (fișiere CSV, date relaționale, data warehouse, data lake, REST endpoints etc.); resurse hardware (platforme HW, CPU/GPU, stocare, resurse cloud, backup, modalitate de comunicare); software (unelte gratuite/plătite - deja disponibile sau achiziționabile, alt software relevant);
- cerințe, presupuneri, constrângeri:
 - calendar al activităților, securitatea datelor/privacy, deliverables (rapoarte, proof of concept etc.)
 - presupuneri care trebuie validate - coerența datelor
 - constrângeri: resurse necesare, timp de rulare, modalitate de deploy, constrângeri tehnologice
- managementul riscului - care sunt porțiunile riscante din proiect? există alternative? care e costul lor?
- terminologie - termeni de business, dar și termeni de ML (accuracy, precision, recall etc.)
- costuri și beneficii - pentru fiecare rezultat potențial, care sunt costurile de proiect pentru atingerea lui?

3. Definirea scopurilor

- scopuri de business - descrierea rezultatelor intenționate, legate de obiectivele de business (creșterea numărului de clienți, limitarea pierderilor dintr-un proces etc.)
- metrici de DS - gradul de acuratețe atins, scor de tip mean squared error/mean absolute error etc.

4. Producerea unui plan de proiect

- pași de executat, durată, resurse cerute, intrări, ieșiri, dependințe; iterațiile trebuie explicit date; detectarea zonelor de risc + alternative

- evaluare initiala a tehnicilor si uneltelor folosite; necesar SW; deciziile pot avea impact de durata mare

▼ 1.2 Pasul 2. Intelegerea datelor

1. Lista surselor de date
2. Descrierea datelor
 - care sunt datele disponibile, in ce format, efort necesar pentru convertirea lor; verificarea faptului ca satisfac niste cerinte de calitate minimale
3. Explorarea datelor - interogari particulare, vizualizarea datelor, rapoarte intermediare
 - relatii intre perechi sau multimi mici de attribute
 - rezultatele unor agregari simple
 - subpopulatii de date (grupare pe genul persoanelor, provenienta geografica, localizare, nivel de educatie, cunostinte de ordin financiar etc)
 - analiza statistica simpla
 - finalizare cu raport de explorare a datelor - primele chestiuni descoperite, impactul lor preconizat asupra proiectului si metodelor utilizate (de ex: ocurenta valorilor lipsa, clase debalansate etc.)
4. Verificarea calitatii datelor
 - asigura toate cazurile cerute?
 - contin erori? chestiune in care domain knowledge e esential; detectare de valori eronate, outliers
 - missing values? daca da, cum se trateaza (stergere de date/attribute, missing value imputation)
5. Raport asupra calitatii datelor
 - lista rezultatelor obtinute in urma investigatiilor din acest pas
 - sugerarea rezolvarilor - dependenta de domain knowledge, tipuri de probleme, unelte DS disponibile

▼ 1.3 Pasul 3: pregatirea datelor

1. Selectarea datelor
 - date selectate dupa relevanta
 - motivele pentru care unele date se exclud trebuie sa fie clar documentate
 - constrangeri tehnice (cantitatea de date) considerate
 - caracterul privat al datelor
2. Curatarea datelor
 - modalitate de selectare/filtrare
 - missing value imputation (valori default sau sub-proiecte pentru estimarea valorilor lipsa)
 - raport al etapei de curatare a datelor
3. Constructii auxiliare
 - selectarea sau extragerea trasaturilor (chestiuni diferite)
 - generarea de inregistrari suplimentare (clienti fara comenzi efectuate, alte situatii preluate din realitate)
4. Integrarea datelor

- jonctiuni de date, concatenare
- agregari de date (numar de achizitii, valoarea totala a cumparaturilor, valoare media/mediana) -> posibil sa duca la noi trasaturi informative

▼ 1.4 Pasul 4. Modelarea

1. Selectarea tehnicii de modelare:
 - documentarea tehnicilor ce urmeaza sa fie folosite
 - presupuneri/cerinte asupra datelor (fara valori lipsa, statistici minimale) - in ce conditii tehnicile de la pct anterior functioneaza
 - adaptarile tehnicilor standard pentru cazurile concrete existente
2. Design-ul pasului de test
 - cum se masoara performanta modelelor (metrici ML/DS)
 - descrierea etapei de validare si testare
 - modalitatea de impartire in train/validation/test subsets
3. Construirea modelelor
 - antrenarea modelelor, determinarea hiperparametrilor, estimarea erformantelor
 - salvarea modelelor rezultate
 - descrierea modelelor si a modificarilor specifice
 - documentarea dificultatilor intampinate (conversia datelor, durata de antrenare/validare, dependente intre pasi etc.)
4. Estimarea performantei modelelor
 - raportarea rezultatelor, interpretarea lor, explicarea comportamentului modelelor
 - revizuirea hiperparametrilor + a valorilor candidat specifice; reantrenare modele

▼ 1.5 Pasul 5: evaluarea

1. Evaluarea rezultatelor
 - confruntarea cu obiectivele de business - estimarea castigului realizat
 - daca e posibil: aplicare modele pe piata
 - considerarea rezultatelor colaterale obtinute, impactul asupra proiectului, considerarea lor ca metrice
 - aprobarea modelelor - cele care indeplinesc cerintele de business devin **modele aprobate**
2. Revizuirea (review)
 - reevaluarea pasilor procesului, lessons learned, posibilitate de extindere, verificarea compatibilitatii cu datele actuale (data privacy, prevederi locale sau generale etc.)
3. Determinarea pasilor urmasori
 - finalizare proiectului si mutarea in productie (deployment)? reluarea iteratiilor?
 - evaluarea resurselor ramase poate influenta decizia
 - se produce: lista de posibile actiuni, motivul de alegere al fiecarei optiuni
 - decizia finala, cu motivatie

1.6 Deployment

1. Planificarea pasului de deployment

- se sumarizeaza pasii prin care modelele adoptate ajung in "productie". Se poate decide deployere pe cloud, pregatirea de masini virtuale setate corespunzator, software containers (Docker), rescrierea modelelor in alte limbaje (C++/Java) etc;
- estimare de efort (timp, oameni, resurse software, calificari - e.g. DevOps)
- sumar al strategiei de deploy, inclusiv pasii necesari, preconditionii

2. Monitorizare si mentenanta

- se evita utilizarile necorespunzatoare, perioadele de nefunctionare, input necorespunzator, interpretarea eronata a productiei sistemelor DS
- se consulta jurnalele de activitate (logging)
- se monitorizeaza performanta modelelor (timpii de reactie, calitatea predictiilor, incarcarea sistemelor); se poate detecta degradarea performantei modelelor, de exemplu din motive de concept drifting;
- plan de monitorizare: actiuni, etape, frecventa de actiune/interogare a starii sistemelor; KPIs;

3. Producerea raportului final

- poate fi doar un sumar al proiectului si experientele dobandite
- poate fi un raport extensiv al rezultatelor obtinute
- artefacte: raport final + prezentare finala

4. Revizuirea proiectului

- documentarea experientei; lectii invatate; greseli in pasii urmati;
- poate utiliza rapoartele realizate de-a lungul pasilor

2 Exemplul 1

[Predictia supravietuirii pe Titanic \(https://www.districtdatalabs.com/how-to-start-your-first-data-science-project\)](https://www.districtdatalabs.com/how-to-start-your-first-data-science-project)

3 Exemplul 2

[Data science workflow example](https://github.com/aakashtandel/misc_projects/blob/master/Data%20Science%20Workflow%20Project)

https://github.com/aakashtandel/misc_projects/blob/master/Data%20Science%20Workflow%20Project