1. **Beskrivning av det problem som ni identifierat samt motivering varför det är ett problem och varför det är ett lämpligt att adressera med er datamängd.**

En dataanalys har utförts där ett fiktivt dataset framställt av dataforskare på IBM har använts. Den valda frågeställningen lyder "Vilka är huvudvariablerna som påverkar anställdas *Attrition*?". *Attrition* är ett boolean-värde i vårt dataset som antingen förekommer som *Yes* eller *No*, där *Yes* betyder att den anställde av någon anledning slutat på jobbet och *No* betyder att den anställde jobbar kvar. Som data science konsulter ville vi hjälpa företaget att förhindra att anställda slutar genom att försöka hitta mönster, dvs. vilka variabler som ofta leder till att anställda slutar på jobbet. Den inledande motiveringen som beskrivs i kodfilen är att frågeställningen är intressant eftersom allmänt sett vill företag att sina anställda ska jobba kvar så att företaget kan växa av detta, samt att det innebär en kostnad om en anställd slutar.

Enligt Frye et al. (2018), som också skapade en datamodell för att förutspå vilka variabler som leder till anställdas förfall, är *Attrition* en kostsam utmaning som flera arbetsgivare handskas med. Det är lätt att inse korrektheten med detta påstående eftersom en anställds förfall leder till att en ny person måste rekryteras, vilket bland annat innebär en rekryteringsprocess samt upplärningsperiod. Datamängden i vårt dataset är inte speciellt omfattande. Totalt sett innehåller datasetet 1470 rader och 35 kolumner. Dock ansåg vi att en överblick över vilka variabler som är mest avgörande för anställdas förfall inte kräver en kopiös datamängd, även om detta skulle ge en än mer ordentlig och empirisk grund för resulterande variabler.

Frye et al. (2018) refererar själva till studier som gjorts för att undersöka vad det kostar att förlora en anställd. En studie från 2012 visar att det kan kosta mellan 16 och över 213% av den anställdes årliga lön. Siffrorna varierar beroende på den anställdes position. Vidare hävdar författarna att det finns definitiva effekter av förlust av anställd, nämligen:

- Avgörande huruvida den lediga platsen ska bytas ut eller om arbetsuppgifter ska distribueras i verksamheten.
- Annonsering av den nya jobbpositionen till olika plattformar.
- Rekryterings- samt upplärningsprocess.
- Behandling av en försämrad moral bland kvarstående anställda.
- Fastställning huruvida verksamheten tolererar en lägre kompetensuppsättning från den anställdes ersättare.

Vi ser alltså hur även de som jobbar kvar drabbas genom att verksamhetens produktivitet minskas, vilket också kan leda till försämrad moral. Med tanke på problemets utsträckning fann vi det lämpligt att utforska vilka variabler som är karaktäriserande för förlust av anställda, trots att datamängden inte var speciellt stor.

2. **Diskutera svagheter och styrkor med vald(a) metod(er)**

Vår dataanalys genomfördes genom att först utföra en explorativ analys. Denna fungerade som en inledande undersökning av den data vi analyserat för att bekanta oss med den. Nästa steg var *Data* Wrangling där felaktiga klassbeskrivningar, utstickande värden (outliers), dubbletter och null-värden hanterades. Om ens dataset innehåller för mycket opålitliga, bullriga, irrelevanta och således onödiga data så kommer modellens träningsfas vara mindre givande (Kotsiantis, Pientelas och Kanellopoulos, 2006). Modellen kan i sådant fall inte identifiera beslutsregler lika enkelt och den data som borde rensats kommer i vägen. Efteråt genomfördes en s.k. *Data processing* i syfte att omvandla och förbereda data inför de algoritmer vi tänkt använda. Slutligen implementerades olika

maskininlärningsalgoritmer för att undersöka frågeställningen. Algoritmer som användes i dataanalysen beskrivs nedan, följt av förklaring till varför dem användes.

- Logistic Regression: en övervakad algoritm som används när målvariabeln är kategorisk och har en eller flera beroende variabler. Exempelvis kan denna algoritm avgöra huruvida ett Email bör hamna i skräppost eller inte. Algoritmen användes för att det är en välanvänd algoritm för binära klassifikationsproblem – problem där det finns endast två utfall (Moreira, Carvalho och Horvath, 2019).

- Decision Tree: en övervakad algoritm som skapar en modell som kan användas för prediktion av klasser och värden av en målvariabel genom att studera beslutsregler från inkommande träningsdata. Algoritmen användes eftersom den kan användas för såväl regressionsproblem som klassifikationsproblem. Algortimen kan också utbringa ett visuellt träd som kan granskas för att sätta sig in i modellens sätt att prediktera (Moreira, Carvalho och Horvath, 2019).

- RandomForestClassification: övervakad algoritm som använder en samling beslutsträd och kan ofta producera bra prediktioner utan inställningar för hyper-parameter. Algoritmen ansågs vara passande till följd av dess simpelhet, mångfald, samt att modellen kan användas för både klassifikation och regression (Moreira, Carvalho och Horvath, 2019).

- LGBMClassifier, AdaBoost och XGBoost implementerades också för att se om någon av dessa producerade ett mer givande resultat en de ovannämnda algoritmerna. Dessa algoritmer tillhör det s.k. Boosting-ramverket (Chengsheng, Huacheng och Bing, 2017). Algoritmerna medförde inte förbättrade resultat än de andra algoritmerna.

Nedan är en sorterad print på hur starkt samtliga variabler är korrelerade till *Attrition*.

```
TotalWorkingYears           0.195474
YearsAtCompany              0.190790
YearsInCurrentRole          0.171889
Age                         0.171781
JobLevel                    0.168421
YearsWithCurrManager        0.164155
MonthlyIncome               0.160598
MaritalStatus               0.156669
JobInvolvement              0.133460
EnvironmentSatisfaction     0.116592
WorkLifeBalance             0.078955
RelationshipSatisfaction    0.076040
JobSatisfaction             0.076020
DistanceFromHome            0.075613
Department                  0.069855
PercentSalaryHike           0.065935
DailyRate                   0.057675
JobRole                     0.051762
Education                   0.045979
MonthlyRate                 0.045583
TrainingTimesLastYear       0.035248
EducationField              0.035236
NumCompaniesWorked          0.023896
Gender                      0.014442
YearsSinceLastPromotion     0.014134
HourlyRate                  0.011265
BusinessTravel              0.004073
EmployeeNumber              0.002388
EmployeeCount                    NaN
Over18                           NaN
PerformanceRating                NaN
StandardHours                    NaN
```

Figur 1

Val av variabler utgick från korrelationsordningen beskriven i Figur 1. De första 16 korrelerade variablerna för *Attrition* valdes, förutom följande undantag:

- Ur ett etiskt perspektiv utelämnades *WorkLifeBalance* och *RelationshipSatisfaction* som berör den anställdas personliga liv.

- För att undvika multikollinearitet utelämnades *JobLevel, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole* och *YearsWithCurrManager*. Eftersom *Age* fortfarande är inkluderad kan *TotalWorkingYears* också anses vara redundant eftersom de uppfyller samma mening för prediktionsmodellen - den som är äldre har haft möjlighet att arbeta flera år än en anställd som är yngre. Ett ytterligare argument är att dessa variabler också är svåra att använda i förbättringssyfte för den anställda. Exempelvis är det svårt att göra något konkret åt att en anställd har jobbat på samma företag en viss tid (*YearsAtCompany*).

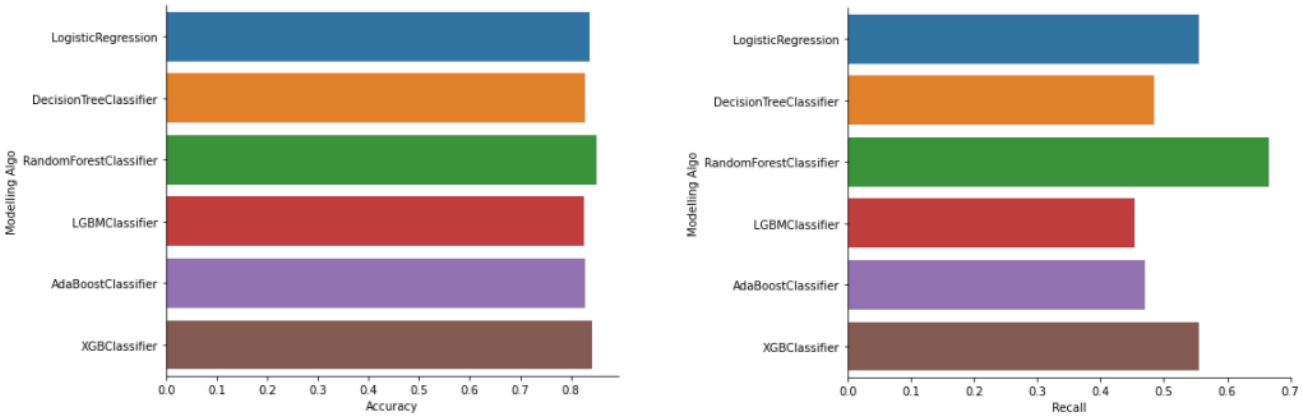
Variablerna som användes i analysen motiveras med att de hjälper prediktionsmodellen att identifiera typiska mönster hos en anställd som är på väg att sluta, samt att de kan användas för vidare utredning inom verksamheten. Ett exempel är att *JobInvolvement* kan förbättras för den anställde genom konversation mellan chef och anställd om arbetsinvolveringen. Följande variabler användes i dataanalysen för att träna modellerna att förutspå *Attrition*: *StockOptionLevel, Age, MonthlyIncome, MaritalStatus, JobInvolvement, EnvironmentSatisfaction, JobSatisfaction* och *DistanceFromHome*.

Dataanalysen framställd av Frye et al. (2018) visade att algoritmen *Logistic Regression* producerade den mest effektiva modellen. Deras inledande dataanalys visar omgående att "length of service" har starkast korrelation till förlust av anställda. Algoritmen hjälpte författarna förutspå anställdas förfall med en noggrannhet (*Accuracy*) på 74%. Vår kods resultat skiljer sig åt från författarnas, vilket är förståeligt med tanke på att olika dataset och variabler använts. Vårt dataanalysresultat visar att algoritmen *RandomForestClassifier* medför högst övergripande poäng på mätvärdenas resultat. Algoritmen kombinerar flera beslutsträd som utgår utifrån olika delprover av datasetet och nyttjar sedan medelvärdet i syfte att förbättra den prediktiva noggrannheten och kontrollera eventuell ”over-fitting” (Moreira, Carvalho och Horvath, 2019).

Mätvärdena *Accuracy* och *Recall* ansåg vi var speciellt vägande för det givna problemet. *Recall* beskrivs som den sanna positiva frekvensen och mäter procentandelen av följande förhållande: Antalet sanna positiva dividerat med summan av sanna positiva och falska negativa. *Recall* refereras även som känslighet och mäter alltså relevanta fall i datasetet (Moreira, Carvalho och Horvath, 2019). I vårt dataset betyder sanna positiva att modellen korrekt förutspått en positiv *Attrition* (*Yes*). Recall kan därmed ses som modellens förmåga att hitta alla relevanta datapunkter för vårt dataset.

Accuracy är det generella måttet för prediktionsmodeller och ger ett inledande resultat hos modellen. Den förutser emellertid inte den lokaliserade nivån av varje individuell klass som förutspås (Moreira, Carvalho och Horvath, 2019). Därför bör inte särskild vikt läggas på detta mätvärde för vårt dataset, vilket vi gjorde. Den mest lämpliga prediktionsmodellen förblir dock densamma – *RandomForestClassification*. Vårt val om att inkludera variabeln *Age* istället för *TotalWorkingYears* kan anses vara felaktigt eftersom den har starkast korrelation till *Attrition*, samt att Frye at al. (2018) också identifierade en snarlik variabel som den med starkast korrelation.

Nedan följer två plottar för algoritmernas *Accuracy* samt *Recall*.



Figur 2

Utifrån värdena på *Accuracy* och *Recall*, som syns i Figur 2, drogs slutsatsen att *RandomForestClassifier* var den mest passande modellen med de variabler vi utsett för att förutspå *Attrition*. Det var speciellt intressant att se skillnaden mellan *RandomForestClassifier* och *Decision Tree* eftersom den förstnämnda är en slags utökning av flera beslutsträd. Modellen producerad av *Decision Tree* fick en högre *Precision* och en längre *Recall* till följd av detta eftersom när *Precision* ökar så reduceras *Recall* (Moreira, Carvalho och Horvath, 2019).

3. Diskutera kvalitetsmässiga och etiska aspekter

1.Med ansatsen som helhet.

Det första steget i att få insikt i anställdas förfall är genom relevant data. Företag motstrider sig ofta utlämning av utnyttjade metoder för hantering av mänskliga resurser, även om metoderna använder anonyma data och är skapad inom verksamheten eller uppköpt (Frye et al., 2018). Detta är förståeligt eftersom företaget i sådant fall måste säkerhetsställa att deltagarna är informerade om detta, vilket rimligtvis leder till följdfrågor om vad denna data ska användas till och hur.

Det gemensamma dilemmat för insamlings- och interpreteringsaktiviteter av data som avser mänskligt beteende är frågan om vem som drar nytta av fördelarna, är det subjektet eller interpreteraren? En modell till för att effektivisera en verksamhet genom att studera mönster av anställdas förfall, skulle potentiellt kunna leda ett företag i en riktning där strävan efter ökad anställningsperiod blir prioriterad. Likaså kan en riktning tas som avser minskning av ersättning och andra förmåner när den anställdes "hållbarhetstid" räknats ut (Frye et al., 2018). Detta är uppenbara etiska dilemman som bör beaktas i dataanalysen.

Enligt Kopp et al. (2016) uppkommer de mest kritiska etiska svårigheterna till följd av problem kopplade till skada eller potentiell risk for forskningsdeltagarna. Detta faktum leder till att forskare som minst ska informera deltagarna om samtliga identifierade risker och i bästa mån se till att de begriper negativa konsekvenser som kan uppstå. På så vis presenteras valet för varje deltagare att antingen ge samtycke eller inte.

Vidare argumenterar Kopp et al. (2016) för att individens respekt måste vara i centrum för att god forskningsetik ska hållas. Författarna refererar till Belmont-rapporten som tydligt säger att individens

respekt omfattar två etiska principer. Den första säger att individer måste ses som autonoma varelser. Det som menas är att individer är fria och kan fatta sina egna beslut. Den andra principen innefattar skydd för individer med reducerad autonomi. Autonomi handlar om individens förmåga att självständigt hantera sina uppgifter och tillhörande data.

När deltagare informeras ska detta göras på ett tydligt sätt utan manipulering av förklarande som kan leda till tvetydliga tolkningar. En viktig aspekt är också att deltagarna informeras om rätten till att dra sig ur projektet när som helst, utan påföljder. Detta krav är essentiellt för skydd av individer som inte själva är fullt kapabla till att fatta informerade beslut till följd av psykologisk påfrestelse. Välgörenhet som koncept anses vara ett etiskt dilemma. I praktiken är konceptet komplext med tanke på att projektet handlar om en god gärning men att det fortfarande finns inbyggda risker när projektet handlar om mänskliga beteenden. Två allmänna regler som formulerats för välgörenhetskonceptet är: undvik att göra skada och maximera potentiella fördelar och minimera potentiella skador. Reglerna fungerar som komplement till verksamhetens välgörande handlingar (Kopp et al., 2016).

Den data som analyserades av Frye et al. (2018) är inte applicerbar på individnivå, utan för kluster av flera anställda. Det finns flera faktorer som påverkar förlust av anställda och när större datamängder tillämpas på mindre verksamheter eller på enskilda personer så förutsätts det att denne har samma beteende som den generella gruppen. Rimligtvis är detta ett problem för dataanalys. Det finns flertalet exempel i verkligheten som tyder på att generaliseringar och vice versa bör undvikas. Exempelvis bör vi inte anta att en viss person agerar på ett visst sätt på grund av dennes ursprung. De flesta håller nog med om att människor inte gillar att bli placerade i fack eller att generella slutsatser dras utifrån ens specifika handlingar.

2.Med den typ av data ni använder (samt även med den datamängd ni använt).

Personliga komponenter som hem- och arbetsbalans är variabler som kan påverka prediktionsmodellen för anställdas förfall. Om dessa personliga aspekter påverkar så är frågan om en personalavdelning (HR) bör ta hand om istället (Frye et al., 2018). Ett argument kan således göras för att sådana typer av variabler borde uteslutas från dataanalysen eftersom de annars kan påverka modellen. Även om modellen påverkas positivt när det kommer till den faktiska prediktionen så återstår risken om partiskhet.

Frye et al. (2018) nämner att det finns etiska implikationer av att använda en prediktionsmodell för data inom verksamheter. Detta var något vi själva reflekterade över under projektet och valde därför att utelämna två variabler *WorkLifeBalance* och *RelationshipSatisfaction*, som vi ansåg vara etiskt kontroversiella. Variablerna berör personliga komponenter av den anställdas liv och således exkluderades dessa. Anställda kan känna sig mindre bekväma med att uttala sina känslor i de fall där anställdas utlämnande ställs som krav för datainsamlingsprocessen. Detta fenomen är igenkänningsbar för de flesta. Exempelvis om en anställd får en fråga som avser nöjdheten med jobbet så kan denne känna sig mindre bekväm då denne inte vet om tillhörande chef kommer få ta del av resultaten (Frye et al., 2018).

Det finns fällor som prediktionsmodeller kan hamna i när mänskliga beteenden analyseras. Exempelvis om anställdas kön samlas in bland ens data, och om prediktionsmodellen finner att kvinnor har högre chans till förfall än vad män har, så kan modellen byggas med avseende på denna korrelation. Risken är då att könet förblir en lika avgörande faktor som exempelvis *MonthlyAmount* som beskriver månadslönen. Ur ett moraliskt perspektiv skulle detta absolut inte godtas som argument vid en anställningsprocess (Frye et al., 2018).

En anställds personliga liv, uppfattning av självvärde, samt sociala och ekonomiska situation kan kraftigt påverka resultatet av tillämpningen av generellt beteende gentemot små populationer.

Eftersom datamängden i vårt dataset är relativt litet så minskar modellens förmåga att prediktera allt för generella mönster hos de anställda. Datamängden blir därav till en positiv aspekt i detta avseende, även om en större datamängd troligtvis hade medfört en mer precis prediktionsmodell.

Problem med datakvaliteten innefattar bland annat saknade data, inkonsekventa eller utstickande värden (Kandel et al., 2011). Värdena håller en konsekvent struktur, vilket märktes i den explorativa analysen. Resterande aspekter hanterades i Data Wrangling-processen. Ytterligare ett sätt att förbättra datakvalitet är genom sociala interaktioner under olika faser av datalivscykeln (Kandel et al., 2011). Även om Data Wrangling förbättrar datakvalitet så leder detta inte till en fullständig datakvalitet. Därför är det viktigt att tänka på datakvalitet i övriga faser, som exempelvis under datainsamlingen. Attrition är ett väldigt obalanserat mätvärde då antalet *No* är betydligt mycket större än antalet *Yes*. Detta leder till att *Accuracy* blir till ett opålitligt mätvärde för prediktionsmodellerna (Moreira, Carvalho och Horvath, 2019).

3.Med vald(a) metod(er)/algoritmer

Vi genomförde anställds-specifika tester där nya data skapades för att representera en ny och fiktiv anställd. Sedan kunde modellen avgöra vad sannolikheten var för denne att sluta på jobbet eller inte, dvs. huruvida *Attrition* är *Yes* eller *No*. Efteråt försökte vi undersöka mönster och fann bland annat att ju längre en anställd är från att kunna gå i pension, desto mer sannolikt är det denne slutar. Detta fynd bekräftas även i dataanalysen genomförd av Frye et al. (2018).

Frye et al. (2018) reflekterar om att modellen skulle förbättras med hjälp av kvalitativa mätningar. Detta eftersom det inte finns någon tillhörande varians mellan två anställda med samma ekonomiska och arbetsmässiga förhållande. Det som skiljer sig åt i vårt dataset är hur den anställdes nöjdhet med miljö och arbetsinvolvering ser ut. Dessa variabler är viktiga för vår prediktionsmodell som på så sätt kan utnyttja korrelationsnivåerna bland dem. Detta öppnar även upp möjligheter för mer variation, samt att nöjdhet i en kontext kan leda till anställdsspecifika förändringar.

En negativ aspekt med *RandomForestClassifier*, som blev vår analys mest lämpliga modell, är att algoritmen är kostsamt för beräkningskraften (Moreira, Carvalho och Horvath, 2019). Till följd av datamängdens storlek blev detta inte ett problem. Det vi dock kunde undersökt noggrannare är vilket antal träd som utgav bäst resultat för algoritmen.

4. Diskutera och reflektera kring potentiella lösningar på etiska och kvalitetsmässiga problem.

Om ett företag är intresserade av frågeställningen vi framhävt så är det rekommenderat att bilda ett dedikerat team / kommitté för rådgivning om insamling av anställdas data. En rekommendation är även att ta hjälp av tredjepartsforskare som engagerar sig till att upprätthålla anonymitet. Vi har sett hur enkelt ett företag kan hamna i oetiska diken om målet endast är att reducera förlust av anställda. Det krävs en omfattande arbetsprocess för att genomföra en sådan typ av analys (Frye et al., 2018).

Tillämpning av en prediktionsmodell för anställdas förfall bör genomföras med en etisk lins. Om den anställde inte kan dra nytta av prediktionerna så tyder det på att modellen endast är av intresse för arbetsgivaren. Som tidigare nämnt, bör varje upptäckt modellen gör reflekteras över med avseende på vem eller vilka som kan få användning av modellimplementeringen. Det följer också att resultaten torde implementeras så nytta kan ges till anställda. Den stora risken är annars att organisationen erhåller och behandlar känsliga uppgifter och orsakar mer skada än nytta. När applicering av prediktionsmodeller sker med en etisk lins kan organisationer använda denna flitigt och på så sätt förbättra lönsamheten, samtidigt som anställdas mål tillgodoses.

En iterativ process för att kontrollera datakvalitet rekommenderas (Kandel et al., 2011). Under datainsamlingsprocessen kan verksamheten exempelvis kontrollera att data inte saknas, samt utvärdera utstickande värden omgående.

5.Slutsatser

Logiskt nog har vi sett att *MonthlyIncome*, *OverTime*, *StockOptionLevel* och *Age* har en hög prioritet för prediktionsmodellernas förutsägbarhet. Variablerna är logiskt kopplade till *Attrition*. Andra variabler har använts i syfte att ge förslag till verksamheten om förbättringar som kan göras för anställda. Ett par exempel följer nedan:

- Om en anställd är missnöjd med sitt jobb (*JobSatisfaction*) kan dennes chef eller HR-personal ha en konversation om hur detta kan förbättras.
- Om en anställd är missnöjd med distansen till jobbet (*DistanceFromHome*) kan den anställde få ett erbjudande om reducerad färdkostnad
- Om en anställd är missnöjd med arbetsmiljön (*EnvironmentSatisfaction*) kan den anställde bli frågad om hur hen vill att den ska bli förbättrad.

För att inkludera de etiskt problematiska variablerna ser jag gärna att verksamheten säkerhetsställer att datainsamlingen gjorts med en etisk lins. De bör även definiera dataanalysens mål och fastställa vilka som kan gynnas av resultaten och vilka som blir skyldiga för eventuella risker.

Om jag hade gjort om arbetet så hade jag valt att inkludera *TotalWorkingYears* som variabel samt lagt mindre vikt på *Accuracy* som inte är pålitligt för obalanserade dataset (Moreira, Carvalho och Horvath, 2019). Om en inledande dialog hade förts med verksamheten innan datainsamlingen så skulle jag verifierat att samtliga deltagare får information om hur dataanalysen kommer genomföras och vilket syfte den ämnar uppfylla. Med samtycke från deltagare kan även variabler som avser personliga komponenter av dennes liv användes för prediktionsmodellen. Jag hade också rekommenderat att verksamheten inför en dedikerad kommitté för att säkerhetsställa god rådgivning för insamlingen av data.

Litteraturförteckning

Chengsheng, T., Huacheng, L. & Bing, X. (2017). *AdaBoost typical Algorithm and its application research* (pp. 1).

Frye, A., Boomhower, C., Smith, M., Vitovsky, L., & Fabricant, S. (2018). *Employee Attrition: What Makes an Employee Quit?*. SMU Data Science Review, 1(1), 9.

Moreira, J. (2019). A General Introduction To Data Analytics.

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., Van Ham, F., Riche, N. H., ... & Buono, P. (2011). *Research directions in data wrangling: Visualizations and transformations for usable and credible data. Information Visualization*, 10(4), 271-288.

Kopp, C., Layton, R., Gondal, I., & Sillitoe, J. (2016). *Ethical considerations when using online datasets for research purposes. In Automating Open Source Intelligence* (pp. 131-157). Syngress.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). *Data preprocessing for supervised leaning. International Journal of Computer Science*, 1(2), 111-117.