**Data Science**

Data Science böyük məlumat həcmi və kompleks datasetlər üzərində statistika, riyaziyyat, kompüter elmləri və məlumat analizi metodlarını tətbiq edərək məlumatdan məna çıxarmağa nail olmağı məqsəd tutan tədqiqat sahəsini ifadə edir. Data Science, böyük məlumat həcmi toplamaq, emal etmək, təhlil etmək, model yaratmaq və məlumata əsaslanan anlamli məlumatları çıxarmağa nail olaraq qərar vermə proseslərini inkişaf etməyə mərkəziləşir.

1. Məlumat Toplanması: Məlumat alimləri müxtəlif mənbələrdən, o cümlədən sensorlar, sosial media, biznes əməliyyatları və s. məlumat toplayır.
2. Məlumatların Təhlili: Nümunələri, meylləri və əlaqələri müəyyən etmək üçün hazırlanmış məlumatlara statistik üsullar və maşın öyrənmə alqoritmləri tətbiq edilir.
3. Məlumatların vizuallaşdırılması: Məlumat alimləri öz tapıntılarını aydın şəkildə təqdim etmək və maraqlı tərəflərə mürəkkəb məlumatları anlamağa kömək etmək üçün qrafiklərdən, qrafiklərdən və digər vizuallaşdırma vasitələrindən istifadə edirlər.
4. Proqnoz Modelləşdirmə: Maşın öyrənmə alqoritmlərindən istifadə edərək, məlumat alimləri tarixi məlumatlar əsasında gələcək nəticələri və ya meylləri proqnozlaşdırmaq üçün modellər qururlar.
5. Məlumatların şərhi: Məlumat alimləri öz analizlərinin nəticələrini şərh edir və tapıntıları qeyri-texniki maraqlı tərəflərə çatdırmaqla onlara verilənlərə əsaslanan qərarlar qəbul etməyə kömək edir.

**Variable** - naməlum və ya qiymətə verilən simvolik addır. Zamanla və ya müxtəlif vəziyyətlərdə dəyişə bilən kəmiyyəti təmsil edir. Dəyişənlər müxtəlif hesablamalarda və əməliyyatlarda məlumatları saxlamaq və manipulyasiya etmək üçün istifadə olunur.

**Week 1**

***Variable*** – data setimiz variable yığılır. Hər sətir observation, hər column bir variable (dəyişkən)

Variable 2 olur: numeric(rəqəmlə(maaş)), categorical(sözlə(təsil pilləsi))

Quantitative eyni Numeric Qualitative eyni Categorical

Categorical 2 olur: nominal(ardıcıl münasibət yoxdu(gender, vəzifə - mühasibat, IT)) ordinal(ardıcıl münasibət(təhsil pilləsi orta, bakalavr, magistr,PhD))

Numeric 2 olur: continuous(kəsr, daha dəqiq ölçmək mümkündür. Geniş diapozonda vahid ala bilir ölçmək ilə təyin edilir.(65.6 kilo)), discrete(tam ədəd(10 nəfər))

Cotinuous 2 olur: Ratio və İnterval

**Statistika – təhlil**

Məlumatların toplanması, təhlili, şərhi, təqdimatı və təşkili ilə bağlı riyaziyyat elminə aiddir. O, sorğuların və təcrübələrin dizaynı baxımından məlumatların toplanmasının planlaşdırılması da daxil olmaqla, məlumatların bütün aspektləri ilə məşğul olur.

Statistikanın Əsas Aspektləri:

1. Məlumatların Toplanması: Statistikaya sorğular, təcrübələr və müşahidələr kimi müxtəlif mənbələrdən ədədi və ya kateqoriyalı ola bilən məlumatların toplanması daxildir.

2. Məlumatların Təhlili: Məlumat toplandıqdan sonra statistika bu məlumatları təhlil etmək və şərh etmək üçün üsullar təqdim edir. Buraya verilənlər daxilində nümunələri, meylləri və əlaqələri müəyyən etmək üçün üsullar daxildir.

3. Məlumatların şərhi: Statistik təhlillər məlumatların şərhinə imkan verir, tədqiqatçılara və qərar qəbul edən şəxslərə əldə edilən məlumatdan mənalı nəticələr çıxarmağa imkan verir.

4. Proqnoz Modelləşdirmə: Statistika mövcud məlumatlara əsaslanan modellər yaratmaq üçün istifadə olunur, daha sonra gələcək hadisələr və ya tendensiyalar haqqında proqnozlar və ya proqnozlar vermək üçün istifadə edilə bilər.

Statistikanın İstifadəsi:

1. Scientific Research: Statistika müxtəlif elmi fənlərdə eksperimental məlumatları təhlil etmək, fərziyyələri təsdiqləmək və təbiət hadisələri haqqında nəticə çıxarmaq üçün istifadə olunur.
2. Business and Economics: Müəssisələr bazar araşdırması, məhsul təhlili və maliyyə proqnozu üçün statistikadan istifadə edirlər. İqtisadçılar iqtisadi meylləri təhlil etmək və proqnozlar vermək üçün statistikadan istifadə edirlər.
3. Sports: Statistika idmanda oyunçu və komandanın təhlili, performansın proqnozlaşdırılması və oyun strategiyasının hazırlanması üçün geniş istifadə olunur.
4. Social Sciences: Sosiologiya, psixologiya və politologiya kimi sahələrdə statistika insan davranışını, ictimai rəyi və sosial meylləri təhlil etmək üçün istifadə olunur.
5. Government and Policy Making: Statistika hökumətlər və siyasətçilər tərəfindən əhali meylləri, təhsil və səhiyyənin planlaşdırılması kimi məsələlərlə bağlı əsaslandırılmış qərarlar qəbul etmək üçün istifadə olunur.

**Subfields of statistics**

***Descriptive*** - mürəkkəb məlumat dəstlərini sadələşdirilmiş şəkildə qrafik şəkildə təqdim etmək üçün istifadə olunur və daha asan şərhə kömək edir.

***Inference*** – kiçik kütlədən alınan məlumatı böyük kütlədən alınan cavab kimi göstərmək (unidəki 200 tələbədən sorğu 80 min tələbəyə bərabər edirsən)

***Risk and Probability*** – istənilən bir hadisənin baş vermə hadisəsinə əsaslanır(kredit gecikmə rayon və şəhər, qadın və kişi arasında)

***Correlation and relationship*** - Korrelyasiya iki və ya daha çox dəyişən arasında statistik əlaqəni ölçür. Aralarındakı əlaqənin gücünü və istiqamətini göstərir.

***Modelling*** - təhlil etmək və proqnozlaşdırmaq üçün real dünya hadisələrinin sadələşdirilmiş təsvirlərinin yaradılmasını nəzərdə tutur.

***Sampling*** - a subset(altçoxluq) of a population. Yaxşı sample əldə etmək üçün əhalinin yaxşı təmsilçi(representative) olması gərəkdir. Əgər 100 nəfərdən 50% qadın 50% kişidirsə sample-da da 50% qadın 50% kişi olmalıdır. Bunu isə random sample edir. Random sample ilə o nəticətini almaq olur.

**Descriptive Statistics**

Telecom da ARPU anlayışı ARPU- average revenue per user ( məsələn 1 aylıq balansına ödədiyin ödəniş)

Mesures of location

***Mean*** = average(outlier tərəfindən asanlıqla təsirə məruz qalır) – toplayıb şəxslərin sayına bölmək

***Median*** – artan sıra ilə düzürsən ortadakı ədəd mediandır. Əgər say cüt olanda ortadakı iki ədədin average

***Mode*** - ən çox təkrarlanan

***Percentile*** - faizlik(faizlə ifadə), median = 50 pecentile (10 nəfər var 1ci adam 10 percentile, 2ci adam 20percentile və ya 2000 işçi var. Ən yüksək maaş alan top 10 neçə nəfərdi. 10percantile (2000\*10)/100 = 200)

***1st quartile*** = ¼hissə, 25 percentile (yarının yarısı) (ən az maaş alan 25 nəfər)

***3rd quartile*** = ¾ hissə, 75 percentile (ən çox maaş alan 25 nəfər)

Measures of variability

***Standard deviation*** – stabillik göstəricisi, bütün yayınmaların ortalması, xəta göstəricisi. Yuxarı olanda qeyri – stabilik və keyfiyyətin aşağı olması, standard deviation aşağı olanda stabilik. (iki nəfərdən birincisi imtahandan 70 70 70 70 70, ikincisi 60 80 50 90 70 bal yığıb hər iksinin ortalaması 70dir. Birinci stabildi standard deviation = 0, ikincidə isə standard deviation (yayınma)) = 15.81. Excel = STDEV

***Normal distributions*** – Statistikanın ən vacib paylanması (insana verilən ağıl, güc, boy, qan təzyiq). Qrafikdə çox vaxt orta hissə daha yuxarı olur. Məsələn, 100 nəfərin zəkası ölçülür - çox kütlə orta zəkaya sahib olacaq, bir qisim hədsiz az adam yüksək IQ-yə, bir qisim hədsiz az adam təfəkkür pozğunluğu olacaq.

***Outlier*** – Dataset-də həddən artıq böyük, həddən artıq kiçik dəyər, anormal case. Z-score ilə təyin olunur. Düzgün qərar verməyimiz üçün və normal paylanma etmək üçün outlier çıxılmalıdır. -3dən aşağı, +3dən yuxarı olanda outlier olur və bunlar silinməlidi. Z-score -5 və +5 olduqda extreme-dir. *Nümunə*: ortalama maaş 1000 manatdır. Std devision = 500. Mən 1500 manat maaş alıram z-score=1, 500 manat maaş alıram z-score= -1, 2000 manat alıram z-core=2.

**Week 2**

**Normal distributions**

Normal distributions da mean, mode, median bərabər və ya bir-birilərinə çox yaxın olur. Simmetrik olur. Lakin bütün simmetrik olanlar normal payalnılmır. Məsələn: uniform distribution, binomial distribution.

Non-parametric data normal paylanmayıb, parametric data normal payalanıb.

***Skewness*** – datanın meyilliyini ölçür. 3cür olur: left-skewness , simmetric, right- skewness.

* Left-skewness(negative) - dataların çoxu sağda olur, quyruq neqativdədi. Mean < Median.
* Right- skewness(positive) – dataların çoxu solda olur, quyruq pozitivdədi. Mean > Median, çünki outlier çoxalıb mean dəyişib.
* Simmetric – Normal distributions

*Example:* insanların çoxu az maaş alır. Datanın çoxu solda yerləşir right skewness

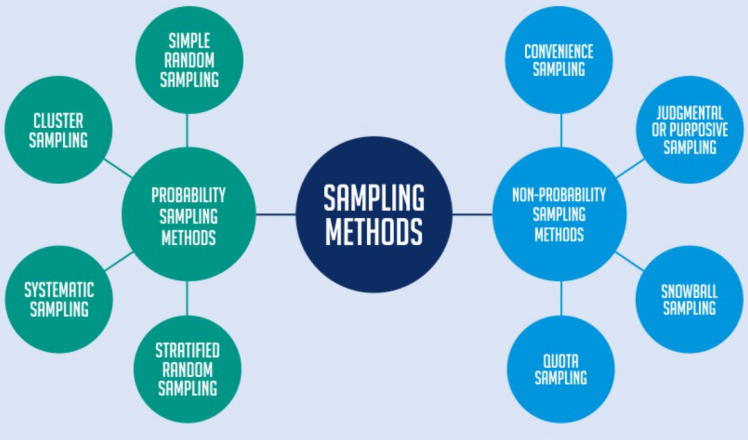
***Kurtosis*** – normal distribution da tepenin yüksekliyini göstərir.

Standard deviation artıqca datanın kurtosis artır. Əgər standart deviation bizdə 0 bərabər olsa bu uniform distribution olur.

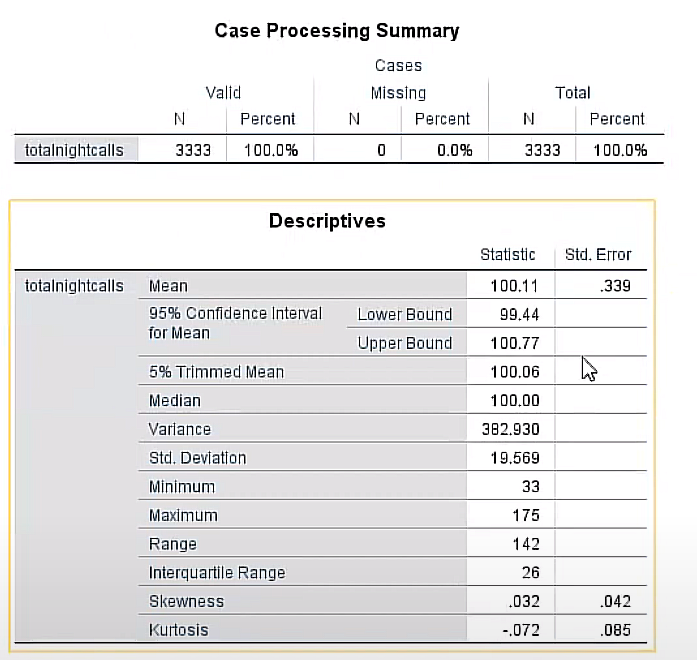
***Empirical rule*** – standart deviation = 200, mean = 1000. Datamın 68% -> 800-1200 arası, 95.4% -> 600-1400 arası və 99.7% -> 400-1600 arası yerləşəcək.

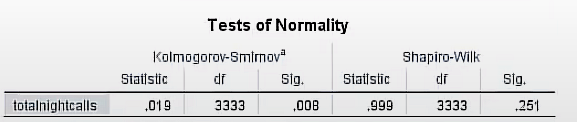
***Standard normal distribution*** - mean = 0, stdv=1.

***Sampling distribution*** – sample size nə qədər çoxdusa nəticə daha dəqiq olacaq.



***Sample size artıqca sample error azalır.***





**N = 3333 variable sayı, simple size. 100% datam missing value yoxdu. Cases missing percent 0% datam tamdır. Mean(ortalama) = 100. 95% *confidence interval* for mean yəni 95% əminliklə deyirəm ki, population meani 99.4 və 100.7 dəyişir. Confidence intervalım nə qədər çox olsa(97.5 və 107.6) range artır və səhv etmə ehtimalım azalır ki, mean həmin aralıqda yerləşir. Variance = Standard deviation kvadratı. Əgər sənin datan da sample size 50dən azdısa shapiro-wilk test nəticəsinə baxırsan amma 50 dən yuxarıdısa hər ikisinə(kolmogorov və shapiro) baxmaq olar. Skewness=0, kurtosis = 0, mean = meadian , sig = 0.251 > 0.05 nəticədə datam normal paylanıb.**

***Z-score fərqi: z-score normal paylanıb və sample sizə 150dən yuxarıdır.***

**Hypothesis Testing**

Hypothesis testing statistikada fundamental anlayışdır və Data Science sahəsində həlledici rol oynayır. Əgər biz nəyisə test edə biliriksə və ya isbat edə biliriksə bu *hypothesis* sayılır. Yox əgər biz onu heç cür test edə bilmiriksə bu fərziyyə olaraq qalır.

Bu, məlumat nümunəsi əsasında əhali haqqında nəticə çıxarmaq üçün sistematik bir üsuldur. Hypothesis testinin əsas məqsədi nümunə məlumatlarında əhali haqqında iddianı (hypothesis) dəstəkləmək üçün kifayət qədər dəlil olub-olmadığını müəyyən etməkdir.

**Steps of Hypothesis**

1. ***Formulate Hypotheses:*** Proses iki fərziyyənin formalaşdırılması ilə başlayır: sıfır hipotezi (H0) və alternativ fərziyyə (H1 və ya Ha). Null hypothesis ödənilmədiyini, alternativ hypothesis isə yoxlamaq istədiyiniz iddianı təmsil edir.
2. ***Null Hypothesis(H0):*** Biz H0 reject etməliyik. Çünki sıfır hipotezi tez-tez heç bir təsirin, fərqin olmadığını və ya əhali arasında əlaqə olmadığını ifadə edən standart bir fərziyyəni təmsil edir. Əlaqə yoxdur, qanuna uyğunluq yoxdur, statistik qərar yoxdur. Random baş verib. Məsələn genderin income aslılığı random baş verir. H0 p-value əsasən baxılır.
3. ***Alternative Hypothesis (Ha):*** Alternativ hipotez sınamaq istədiyiniz iddianı əks etdirir, bir qanuna uyğunluq var və o təsirini, fərqini və ya əlaqəni mövcudluğunu göstərir. H1 mənalı bir əlaəq olduğunu göstərir. H0 reject edəndə H1 bizim üçün məqsədə uyğun olacaq.
4. ***Collect Data:*** Məlumat maraq dairəsini təmsil edən sampledan toplanır. Sample data əhali haqqında nəticə çıxarmaq üçün istifadə olunur. Əgər papulation 40% kişi, 50% qadın bölünübsə bu sampleda da belə olmalıdır.
5. ***Statistical Test:*** Verilənlərin typena və hypothesis əsasında statistik test seçilir. Bu sample datadan hesablanmış numeric dəyər olan test stafistikası yaradır.
6. ***Calculate P-value:*** Test statistic p-value hesablamaq üçün istifadə olunur. Əgər 95% əminlik varsa alfa 0.05, 99% əminlik varsa alfa 0.01 olacaq.
7. ***Make a Decision***: Əgər p-value 0.05dən və 0.01dən kiçikdirsə siz sıfır hypothesis reject edə bilərsiniz. Əgər p-dəyəri kiçik deyilsə, sıfır hipotezini fail to reject (rədd etməmək).

**Normality vs Correlation tests**

***Normality tests:*** Normallıq testləri verilmiş datasetdə normal distribution əməl edib-etmədiyini qiymətləndirmək üçün istifadə olunur. Normallıq testində null hypothesis (HO) adətən məlumatların normal distribution gəldiyini bildirir. Testlə əlaqəli p-value kiçikdirsə (adətən seçilmiş əhəmiyyət səviyyəsindən azdırsa (bir çox hallarda bu, 0,05-dir, çox vaxt alfa kimi qeyd olunur), siz null hypothesis reject edə bilər və məlumatların normal distribution əhəmiyyətli dərəcədə kənara çıxdığı qənaətinə gələ bilərsiniz. Normal disrtibution da null hypothesis həmişə müsbət case kimi götürürük. Digər caselərdə negativedir (əlaqə yoxdur, təsir yoxdur və s). Buna görədə null hypothesis 0.05dən kiçik olanda reject etdiyimiz üçün, normal distributionda p-value 0.05dən yuxarı götürülür.

***Bu halda null hypothesis (H0) datasetin normal şəkildə paylanmasıdır (normal distribution).***

***Correlation tests:*** Correlation testləri iki və daha çox dəyişən arasında statistically significant relationship olub olmadığını müəyyən etmək üçün istifadə olunur. Correlation testindəki null hypothesis (H0) adətən dəyişənlər arasında korrelyasiya olmadığını bildirir. Əgər correlation testi ilə əlaqəli p-dəyəri kiçikdirsə (adətən seçilmiş əhəmiyyət səviyyəsindən az), siz null hypothesis reject və dəyişənlər arasında statistik əhəmiyyətli əlaqənin olduğu qənaətinə gələ bilərsiniz.

***Bu halda null hypothesis (H0) dəyişənlər arasında correlation olmamasıdır.***

**P-value**

"Probability value " qısaldılmışı p-valuedur, statistik hypothesis testində null hypothesisə əksinə olacaq dəlilərin gücünü qiymətləndirməyə kömək edən rəqəmsal ölçüdür. P-value məqsədi odur ki, H0 reject edir, doğurdanda əlaqə olduğunu, statistik cəhətdən bunun bir izahı olduğunu və random baş vermədiyini göstərsin. Əgər p-value böyükdürsə, bu, null hypothesis ehtimal altında müşahidə olunan məlumatların xüsusilə olası olmadığını göstərir. Bu, null hypothesis failure to reject (edilməməsinə) səbəb ola bilər. Bununla belə, p-value null hypothesis nəticəsi haqqında birbaşa məlumat vermədiyini başa düşmək çox vacibdir.

Kiçik p-qiyməti alternativ hypothesi(H1) doğru olduğunu sübut etmir; sadəcə olaraq məlumatların null hypothesis ilə uyğunsuz olduğunu göstərir.

Bundan əlavə, böyük p-value null hypothesis doğru olduğunu sübut etmir; bu, sadəcə olaraq, məlumatların null hypothesis ilə ziddiyyət təşkil etmədiyini göstərir.

P-value məlumatlarınızda müşahidə edilən fərqlərin və ya təsirlərin statistik əhəmiyyətli olub-olmadığını və ya onların təsadüfi nəticəsində baş verə biləcəyini qiymətləndirməyə kömək edir. Əgər p-value kiçikdirsə (alfadan azdır), bu, müşahidə edilən nəticənin təsadüfi nəticəsində olma ehtimalının az olduğunu göstərir və məlumatlarda mənalı təsir və ya əlaqə olduğunu göstərir. Digər tərəfdən, əgər p-value böyükdürsə, bu, müşahidə edilən nəticənin təsadüfi nəticəsində ola biləcəyini göstərir və əhəmiyyətli təsir iddia etmək üçün kifayət qədər sübutunuz olmaya bilər.

**One-Sample T-test**

One-sample t-test, the mean of a single sample məlum və ya hypothesizdən asılı olaraq population meandən significantly different olub olmadığını müəyyən etmək üçün istifadə olunur.

1. Sample sayı populationdan random olaraq götürülür.
2. Bizim variable normal paylanmışdır.
3. Datamızda significant outlier yoxdur.

Report:

1. P-value (Sig < 0.05)
2. The mean difference between the sample mean and the population mean

*Example*: Təsəvvür edin ki, bir cofeshop da işləyirsiniz və orta ölçülü kapuçinonun ölkə üzrə orta qiyməti 3,50 dollardır. Qoyulan bu qiymət population üçün aktualdırmı? Qiymət yuxarı və ya aşağı deyilki? Bütün satdığın kapuçinoların qiymətlərini population ilə müqayisə edirsən və one-sample t-test keçirirsiniz.

Null hypothesis(H0): kofeshop da orta ölçülü kapuçinoların orta qiyməti 3,50 dollardır, qiymət aktualdır (fərq yoxdur).

Alternativ hypothesis (Ha): kofeshop da orta ölçülü kapuçinoların orta qiyməti 3,50 dollar deyil (fərq var). Qiymət çox aşağı və ya çox yuxarı qoymusan. Nə dərəcədə fərq edir ona baxırsan.

Interpretation: One-sample t-tesdən alınan p-value kiçikdirsə (tutaq ki, 0.05 azdır) H0 reject edirsən , bu o deməkdir ki, kofeshop da orta ölçülü kapuçinoların orta qiyməti 3,50 dollardan significantly fərqlidir.

**Independent Samples T-test**

Independent samples t-test, bir-birindən significantly different olub olmadığını müəyyən etmək üçün iki independent qrupun ortalamalarını müqayisə edir.

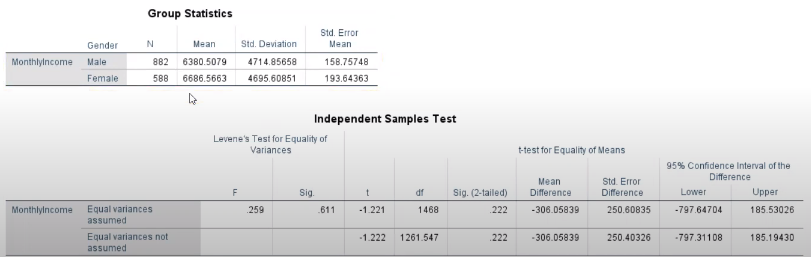
1. Independentlar arası connection olmamalıdır.
2. Dependent variable normal paylanmalıdır.
3. Significant outlier olmamalıdır.
4. Variance eyni olmalıdır. Paylanması çox fərqlənməməlidir.

*Example:* Siz müəllimsiniz və bilmək istəyirsiniz ki, *iki fərqli* sinifin imtahan ballarında ciddi fərq var yoxsa yox. Bir sinif ənənəvi üsullarla, digər sinif isə yeni öyrənmə texnikasından istifadə edirdi. Siz imtahan ballarını toplayır və independent samples t-test keçirirsiniz.

Null hypothesis(H0): İki sinfin orta imtahan balları eynidir (tədris texnikasına görə no difference).

Alternativ hypothesis(Ha): İki sinfin orta imtahan balları eyni deyil (tədris texnikasına görə difference var). Demək mən bu tədris üsullundan istifadə edə bilərəm çünki fərq var qanuna uyğunluq var.

Interpretation: Independent samples t-testindən alınan p-value kiçikdirsə (0,05-dən az), bu o deməkdir ki, iki sinfin orta imtahan balları significantly dərəcədə fərqlidir.



Qadınla kişilərin ortalama maaşı müqayisə olunur. Sig=0.22 > 0.05 significant deyil. Lower mənfi və upper müsbət olduğu üçün population mean o qədər də fərqlənmir. Yəni qadınların və kişilərin ortalama maaşı arasında fərq çox yoxdu. Cədvəldə mean arasında(6380 və 6686) fərq göstərilsədə populotion üçün significantly different deyil.

**Paired Samples T-test**

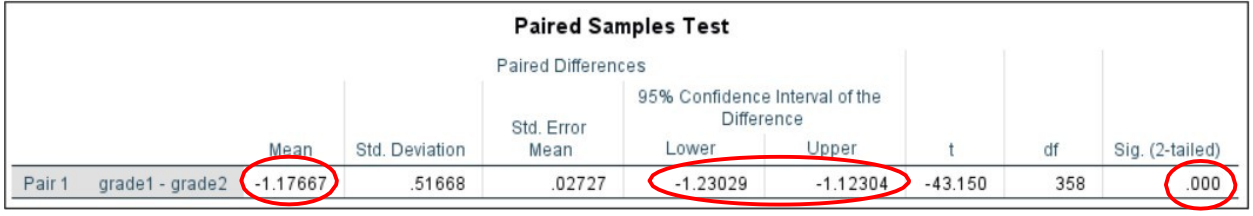
Paired samples t-testi bir-birindən significantly different olub olmadığını müəyyən etmək üçün iki əlaqəli və ya paired qrupun meanslərin müqayisə edir. Bir insanın əvvəlki və sonrakı çəkisi, bir tələbənin eyni imtahanı həm online həmdə offline verdikdə nəticələri, bir dərmanın birincidə dozanın aşağı ikincidə isə dozanın yuxarı olduqda nəticələri paired sample t-test ilə müqayisə edilir.

*Example:* Misal: Siz çəki atmaq üzrə məşqçisisiniz və yeni məşq proqramının müştərilərinizin fitness səviyyələrini yaxşılaşdırmaqda effektiv olub olmadığını bilməyəcəksiniz. Siz yeni proqramdan əvvəl və sonra müştərilərin fitnes ballarını ölçür və paired samples t-testi keçirirsiniz.

Null hypothesis (H0): Proqramdan əvvəl və sonra orta fitness balları eynidir( proqrama görə heç bir irəliləyiş yoxdur).

Alternativ hypothesis (Ha): Proqramdan əvvəl və sonra orta fitness balları eyni deyil (improvement due to the program).

Interpretation: Əgər paired sample t-testindən alınan p-value kiçikdirsə (0,05-dən az), bu o deməkdir ki, proqramdan əvvəl və sonra orta fitness ballarının significantly differentdir və bu proqramın effektiv olduğunu göstərir.



Məsələn 2 imtahan keçirilir. Sig >0.05, mean yoxdur (mean = 0). İki imtahanında dərəcəsi eynidir.

Əgər sig < 0.05, mean varsa göstərilən mean qədər imtahan dərəcəsi arasında fərq var.

Mean var amma sig >0.05 imtahanların arasında fərq yoxdu sadəcə bu random baş verib.

**ANOVA Test**

ANOVA can help us determine whether there are significant differences between the means of three or more groups, for a continuous variable.

1. Üç və daha çox kateqoriya ilə independent variable categorical olmalıdır.
2. Dependent variable numeric olmalıdır və hər bir qrup normal paylanmalıdır.
3. Dependent variable da outlier olmamalıdır.
4. Dependent variable da hər qrupda variances beraber olmalıdır. Bu şərt yerinə yetirilməzsə, Welch testi adlanan F testinin robust versiyasından istifadə etməliyik.

*Example:* Tutaq ki, siz üç fərqli pəhrizin (A, B və C) arıqlamağa təsirini araşdıran bir tədqiqatçısınız, müəyyən müddət ərzində hər bir pəhrizə əməl edən iştirakçılardan çəki məlumatlarını topladınız. Üç pəhriz arasında kilo itkisində significant difference olub olmadığını müəyyən etmək istəyirsiniz.

Null Hypothesis (H0): A, B və C pəhrizləri üçün arıqlama vasitələrinin hamısı bərabərdir (o significant difference in weight loss between diets).

Alternative Hypothesis (Ha): Ən azı bir pəhrizin orta çəki itkisi digərlərindən fərqlidir (significant difference in weight loss between diets).

Interpretation: ANOVA testindən alınan p-value kiçikdirsə (0,05-dən az), bu, üç pəhriz arasında kilo itkisində significant difference olduğunu göstərir. Bu, hansı pəhrizlərin bir-birindən fərqli olduğunu dəqiq söyləmir; bu, yalnız pəhrizlərdən ən azı birinin o fərqli təsirə malik olduğunu və ya çəki itirdiyini göstərir.

**CIRSP DM METHODOLOGY** – Layihənin gedişatın da köməklik edir. Report yazmalıyıq (datanı hardan əldə etdiyini , references, icazə aldığın şəxs və s.)

* ***Business Understanding*** - Əsas diqqət layihələrin biznes məqsədlərini, tələblərini və məhdudiyyətlərini başa düşməkdədir. Bu, həll edilməli olan əsas biznes probleminin müəyyən edilməsini və məlumatların işlənməsinin biznes məqsədlərinə çatmağa necə kömək edə biləcəyini müəyyən etməyi əhatə edir.

Həll etməyə çalışdığımız əsas biznes problemi nədir?

Layihələrin konkret məqsədləri və gözlənilən nəticələri hansılardır?

Biznes problemini həll etmək üçün hansı məlumatlar mövcuddur və ya toplana bilər? (sources, survey genişləndirmək and etc)

Layihənin məhdudiyyətləri nələrdir?

* ***Data Understanding*** - Mövcud məlumatları araşdırır və onlarla tanış olursunuz. Siz məlumatları toplamalı və keyfiyyətini qiymətləndirməli, məlumatların strukturunu, əlaqələrini və potensial problemləri başa düşməlisiniz.

Hansı məlumat mənbələri mövcuddur və biz onlara necə daxil ola bilərik? (hər hansısa şirkətdən əldə etmək)

Dataset də hansı xüsusiyyətlər (dəyişənlər) var və onlar nəyi təmsil edir? ( hansılar lazımlıdır hansılar lazım deyil, adları nədir)

Məlumatda hər hansı çatışmayan və ya səhv dəyərlər varmı? (adı, soyadı var lakin genderi qeyd etməyib onu araşdrımalıyıq)

Məlumatın keyfiyyəti ilə bağlı həll edilməli olan hər hansı problem varmı? ( Female/Male əvəzinə F, M yazmaq, outliets, missing value, duplicates and etc.)

* ***Data Preparation*** - Modelləşdirmə üçün məlumatların hazırlanması. Buraya məlumatların təmizlənməsi, dəyişdirilməsi və müvafiq xüsusiyyətlərin seçilməsi daxildir. Məqsəd modelləşdirmə üçün yaxşı hazırlanmış verilənlər bazası yaratmaqdır.

Çatışmayan məlumatları doldurmaq və ya silinməsi yolu necə olunacaq?

Yeni müvafiq column yaratmalıyıqmı (feature engineering)? (date column 20.02.1997 verilib age column yaradıram)

Hansı inputları daxil edirsən featues kimi? ( x input və y output, x independet və y dependent)

Məlumatların normallaşdırılması və ya standartlaşdırılması tələb olunurmu? ( z-score tapmaq )

* ***Modelling*** – Data tam hazırdır, modelimə tam yararlıdır. Modelimi qururam.
* ***Evaluation*** – Modelimin yaxşı işləyib-işləmədiyini bilmək üçün evaluation hissədə baxıram. Qiymətləndirirəm. Evaluate edirəm yaxşıdır.
* ***Deployment*** - Model artıq productiona hazırdır. Saytda deploy etmisən və population həmin sayta daxil olaraq öz datalarını qeyd edirlər.

1saat 53dk - 2 saat 6dk qədər.

**Week 4**

**DATA QUALITY**

Data quality dəqiqliyə, tamlığa və ya hər hansısa digər məlumatla əlaqəli fəaliyyətləri aiddir. Yüksək keyfiyyətli məlumatlar və etibarlı qərarlar qəbul etmək, dəqiq fikirlər yaratmaq və etibarlı modellər yaratmaq üçün vacibdir. (missing value və outlier yoxdur, istənilən kimidir).

Data nadir hallarda mükəmməldir. Beləliklə, məlumatların toplanması zamanı və toplandıqdan sonra bəzi məlumatların keyfiyyət problemi yaranır. Məsələn:

* ***Missing data:*** Bir və ya daha çox variable qeydə alınmadıqda və ya müəyyən müşahidələr üçün əlçatan olmadıqda baş verir. Məlumatların hazırlanması zamanı düzgün istifadə edilmədikdə, qərəzli təhlillərə və qeyri-dəqiq nəticələrə gətirib çıxara bilər. (Mean – numeric, mode – numeric, categorical variable boşluqları doldurmaq üçün istifadə olunur və ya araşdırılır. ($null$, 999 kimi qeyd olunur))
* **Inaccurate Data:** Məlumatların daxil edilməsi, məlumatların ötürülməsi və ya verilənlərin inteqrasiyası zamanı səhvlər baş verdikdə qeyri-dəqiq məlumatlar yaranır. Yanlış təhlillər və etibarsız məlumatlara əsaslanan səhv qərarlar ilə nəticələnə bilər. (( Female/Male əvəzinə F/M, 175cm əvəzinə 1.75 və ya Xaçmaz əvəzinə Khacmaz yazmaq)
* ***Inconsistent Data:*** Uyğun olmayan məlumatlar eyni dataset daxilində və ya müxtəlif datasetlər arasında format, ölçü vahidləri və ya kodlaşdırma sxemləri ilə fərqlənən verilənlərə aiddir. Uyğunsuzluqlar müxtəlif mənbələrdən alınan məlumatların müqayisəsi və birləşdirilməsində çaşqınlıq və çətinliklərə səbəb ola bilər. (175 inch-i 175cm kimi və ya 5m-i 5km qeyd etmek)
* ***Duplicate Data:*** Verilənlərin daxil edilməsi xətaları və ya verilənlərin inteqrasiyası zamanı eyni qeydlər verilənlər bazasında bir dəfədən çox olduqda təkrarlanan məlumatlar baş verir. Dublikat məlumatlar sayları artıra, statistik məlumatları təhrif edə və təhlildə artıqlığa səbəb ola bilər. ( unikal dəyərlər ID, Fin, CustomerID, AZE kodu və s. tək bir şəxsə aiddir)
* ***Outliers:*** Outliers məlumatların qalan hissəsindən əhəmiyyətli dərəcədə yayınan məlumat nöqtələridir. Bunlar həqiqi müşahidələr və ya məlumat daxiletmə xətaları ola bilər. Outliers statistik təhlili təhrif edərək, mean və standart devation kimi ölçülərə təsir edə bilər. *Outlier 3 cür olur:*

1. Natural outliers : Təbii olan hadisələr. Sinifdə 30 nəfərdən bir nəfər 90 alıb, digərləri 70 alıb. Bir bölgədə illik yağıntıların datasetin də, xüsusi bir il ərzində qeyri-adi dərəcədə güclü yağışlar baş verir və bu, datanın artmasına səbəb olur.
2. Measurement errors: Bir insan hündürlüyü 180m qeyd etməlidir 180cm qeyd edib. Yaş column iki ədədlidir 500 yazıb və ya çəkidə 55 yerinə 5.5 yazıb.
3. İntentional outliers: Nümunə: Dərman Effektivliyi Araşdırması

Dərman effektivliyinin öyrənilməsi zamanı tədqiqatçılar statistik analizlərinin möhkəmliyini qiymətləndirmək və dərmanın ekstremal hallarda necə təsir göstərdiyini başa düşmək üçün məlumatlara qəsdən kənar göstəricilər təqdim edirlər. Dərman təhlükəsizliyi hədlərini müəyyənləşdirin.

Metadata – Data haqqında məlumat. Məsələn bir qutudu, içindəki ilə bağlı üstündə qeydlər yazılıb (rəngi, ölçüsü, forması və s.). Column başqa cür təsvir olunub amma column da başqa şeylər qeyd olunub bu da data quality aiddir.

**DATA QUALITY REPORT**

Məlumatın keyfiyyəti hesabatı müəyyən bir təhlil və ya layihədə istifadə olunan məlumatların keyfiyyətinə dair ümumi məlumat verən sənəddir.

1. Introduction:

Məlumatın keyfiyyəti hesabatının məqsədi nədir?

Hesabatda əhatə olunan məlumatların əhatə dairəsi nədir (məsələn, xüsusi verilənlər toplusu, müddət)? (adı, əhatə etmə müddəti, haqqında məlumat)

1. Data sources and collection methods:

Məlumatların mənbələri hansılardır? İnternal or external?

Məlumatlar necə toplanıb (məsələn, sorğular, sensorlar, əl ilə daxiletmə)?

1. Data Quality Metrics: (Dəqiqlik, Tamlıq(missing value), Etibarlılıq, Consistency(female-f), Unikallıq(dublikat olmamaq) və s.)

Verilənləri qiymətləndirmək üçün istifadə olunan əsas məlumat keyfiyyət göstəriciləri hansılardır?

Hər bir məlumatın keyfiyyət göstəricisi necə hesablanır və ya ölçülür?

Hər bir tarix keyfiyyət göstəricisi üçün məqbul həddlər və ya standartlar hansılardır?

1. Data Accuracy:

Verilənlər düzgünlük və dəqiqlik baxımından nə dərəcədə doğrudur?

Məlumatın keyfiyyətinin qiymətləndirilməsi zamanı hər hansı təfərrüat daxiletmə xətaları və ya qeyri-dəqiqliklər aşkar edilibmi?

Məlumatların düzgünlüyünü təmin etmək üçün hansı tədbirlər görülür?

1. Data Completeness:

Məlumat nə qədər tamdır? Çatışmayan dəyərlər və ya natamam qeydlər varmı?

Verilənlərin hansı hissəsi çatışmır və hansı dəyişənlərdə və ya sahələrdə?

İtkin məlumatlarda hər hansı bir nümunə və ya meyl varmı?

* + Conclusion:

Məlumatların keyfiyyəti hesabatının əsas nəticələrini ümumiləşdirin.

Məlumatın keyfiyyətinin əhəmiyyətini və onun qərar qəbul etmə və təhlilə təsirini təkrarlayın.

***Controlling outliers***

Coerce – outlieri outlier olmayana yaxınlaşdıracaq. Məsələn 100 outlier deyil, 101 outlierdir. 101i 100ə yaxınlaşdıracaq.

Discard – tamamilə bütün outliers silir. Bu yaxşı method deyil ara da natural outliers silinir.

Nullify – outlieri empty kimi əvəz edir. Null deyer edirsen və mean, mode ilə onu doldurursan.

Coerce outliers/ discard extremes – outliers sıxır, extremes olanları -5 və +5 ondan kənar olanları silir.

***Handling missing values***

$Null$ value – numeric boşluqlar və ya düzgün olmayan value

White space – categorical aiddir, boşluqdur və length 1dir

Empty string – categorical aiddir heç value yoxdur

Blank value – numericlərin içinə categorical value qarışıb və ya categoricalların içinə numeric value qarışıb onu type modundan dəyişirsən.

*3 cür aradan qaldrımaq olar*:

Remove fields - əgər 70-80% missing value olanda

Remove cases – sətirdə doldura bilmədiyin column çoxdusa onda sətiri sil

Impute missing value

* Fixed bütün hallar üçün eyni dəyəri istifadə edir. Fixed dəyər constant, mode, mean və ya median ola bilər (seçimlər sahənin ölçmə səviyyəsindən asılı olaraq dəyişəcək).
* Random normal və ya vahid paylanmaya əsaslanan təsadüfi qiymətdən istifadə edir.
* Expression çatışmayan dəyərləri müəyyən etmək üçün öz tənliyini yaratmağa imkan verir.
* Algorithm C&R Tree modeli tərəfindən proqnozlaşdırılan dəyərdən istifadə edir.

**CLEM – Control Language for Expression Manipulation**

Select – age > 30 olanları göstərir

Balance – target yes/no ibarətdir. Columnda 500 variable var. 400 yes, 100 no bu inbalancedır. Bunu balanslaşdırmaq lazımdır.

Aggregate – group by edirik

Derive – yeni column yaratmaq

Filler – replace

***Deleting duplicates***

1. 3500
2. 2500
3. 56
4. 1
5. 56151

Create a composite record for each group – eyni ID olan toplayır. 1 6056, 2 1, 3 56151

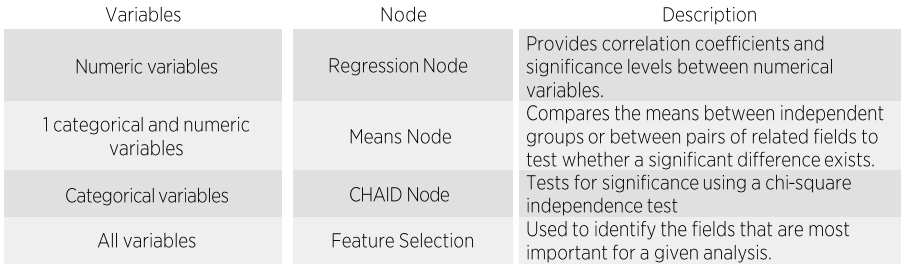
Include only the first record in each group – ilk fərqli olanı saxlayır. 1 3500, 2 1, 3 56151

Discard only the first record in each group – dublikatı olub-olmadığın yoxlayır. 1,2,3,4,5,6 bu IDləri yoxlaya-yoxlaya gedir, ilk fərqli olanı silir, dublikat yoxdusa 0 qaytarır.

***Feature selection node*** – istədiyin field target edirsən və ona təsir edən featureları sənə göstərir.

**Week 5**

**Diagnostic Analytics**



**Machine Learning**

Insanın verdiyi alqortim ilə dataları öyrənir və toplayır. İnsan bir əşyanı rəngindən, iyindən,dadından tanıyırsa machine da onun alqoritimin ötürür və machine həmin datanı tanıyır.

*İki tip* öyrənmə yolu var: *Supervised* və *Unsupervised*

Supervised də bir-bir alqoritmi verirsənki bu almaldır bu armuddur. Unsupervised də isə alqoritmi verirsən ona uyğun özü tapır. Məsələn bir uşağa meyvələri verirsən özü onu hər meyvənin adına görə uyğunlaşdırır.

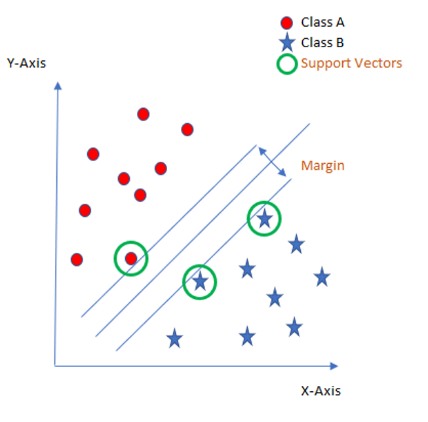
Supervised 2 yerə bölünür: classification and linear models

***Classification models*** – sənin targetin hansısa classa aid olmasını öyrənirsən. Hər hansısa classa aiddir ya yox. İki cür class olur: binary və multi. Binary class da 2 dənə output olur. ( 1 və 0, yes/no). Multi class da output 3 və daha çoxdu(5 şəkil verirsən özü nəticədə 5 şəkilin hansı meyvə olduğun göstərir)

***CHAID modelling -*** suallar nəticəsində decision tree qurur və targetə təsir edən columlar arsında əlaqəni göstərir***.***

***Classification & Regression Trees (CART) modelling***  *-* ilk öncə dataları classification edir(bir-birindən ayrırır) daha sonra decision tree qurur.

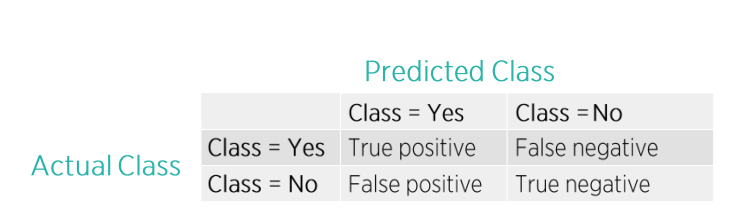
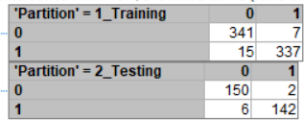
***Support Vector Machine(SVM)*** – dataları bir-birindən ayırmaq üçün threshold ilə bölür. Thresholda yaxın olan və onun sahəsinə düşən data pointləri düzgün classlara bölə bilmir. Nəticədə data pointləri wrong class edir.



***Partition*** *–* Datanı train və test bölür. Train datanın çox hissəsin götürərək datadan düzgün model qurulmasında və qalan dataları götürərək datamızı test edir.

***Auto classifier*** – datamızın düzgün şəkildə bölərək proqnoz vermək üçün bizə ən yaxşı modelləri göstərir.

***Confusion matrix parameters***

* *

İki sinifə bölürük: Yes=Postivie class = 0 No=Negative class = 1

Yes-lə Yesin kəsişməsi yəni 0-ilə 0 kəsişməsi true verir. Çünki modelim 341 ədəd 0 classından 0 düzgün seçib. True Positive olur.

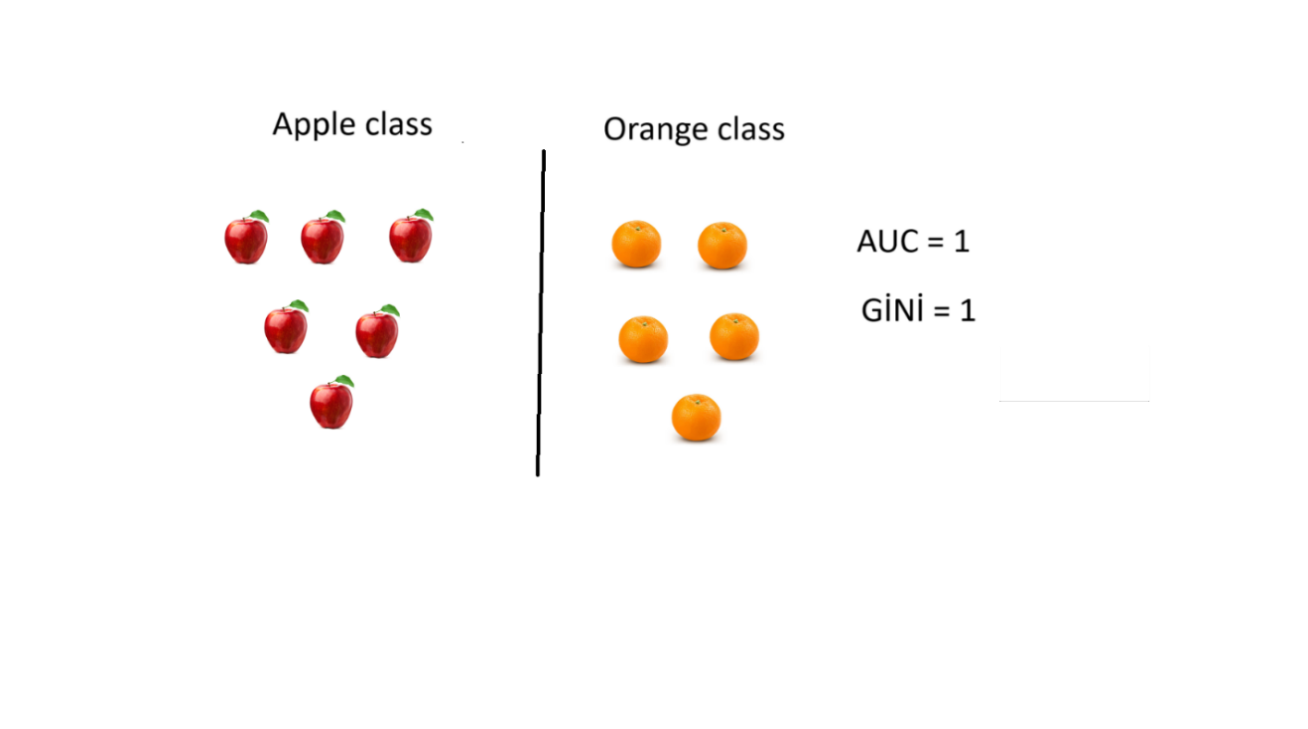
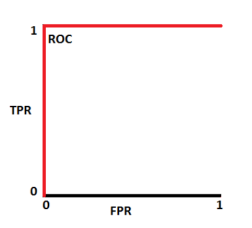
Yes-lə No kəsişməsi yəni 0-ilə 1 kəsişməsi false verir. Çünki modelim 15 ədəd 0 classında olan 1-ləri 0 deyə seçib. False Positive olur.

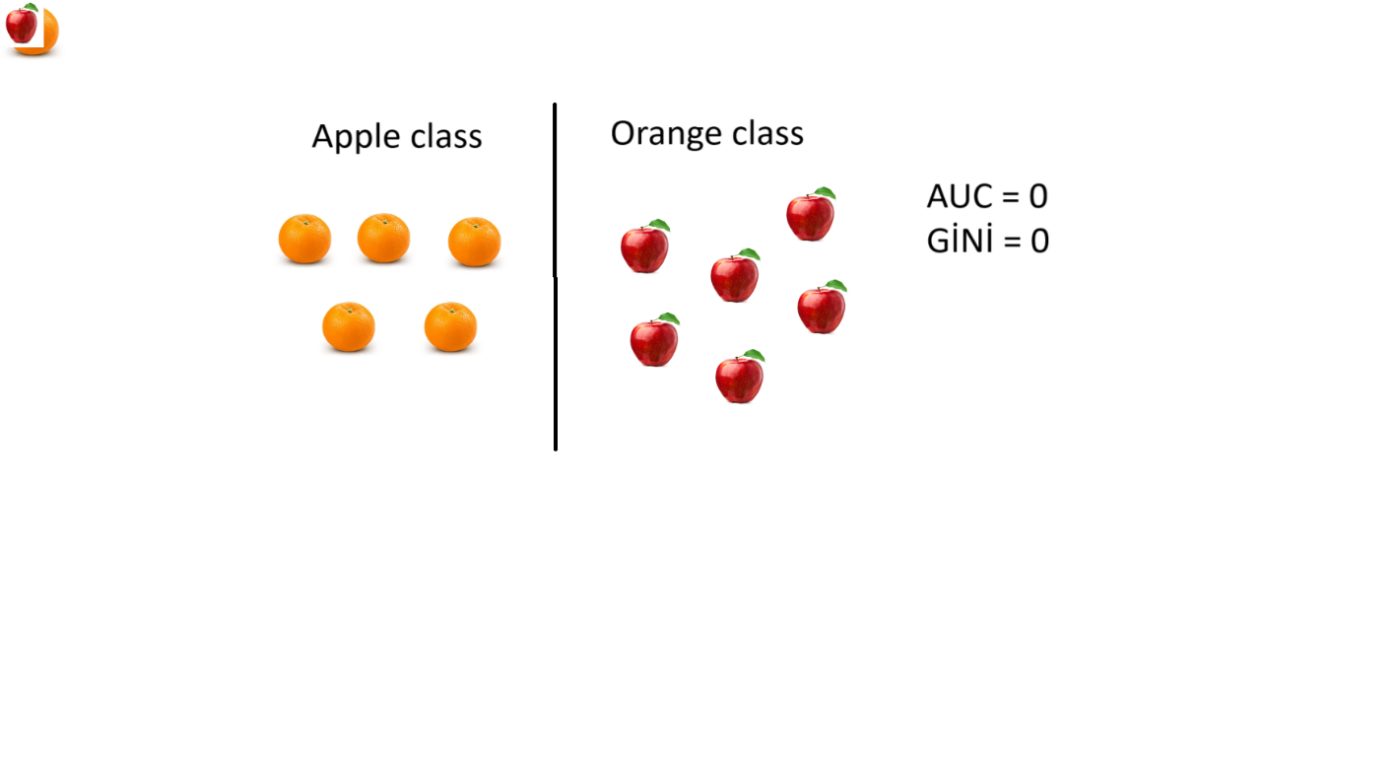
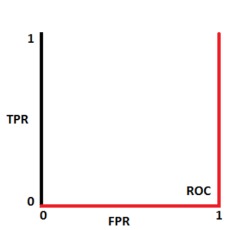
No-la Yesin kəsişməsi yəni 1-ilə 0 kəsişməsi false verir. Çünki modelim 7 ədəd 1 classında 0-ları 1 deyə seçib. False Negative olur.

No-la No-nun kəsişməsi yəni 1-ilə 1 kəsişməsi true verir. Çünki modelim 337 ədəd 1 classında 1 düzgün seçib. True Negative olur.

***Performance measures –*** *Pericision* bir tərəfdə olan classda variableların neçəsin düzgün proqnoz verilmə ehtimalı. *Recall* bütün bir classda olan variableların neçəsi düzgün praqnoz verilib***.*** *F1 score* pericion və recall-da daha çox ehtimal faizi olması üçün düzgün çəkilən xətt. <https://www.youtube.com/watch?v=qWfzIYCvBqo>

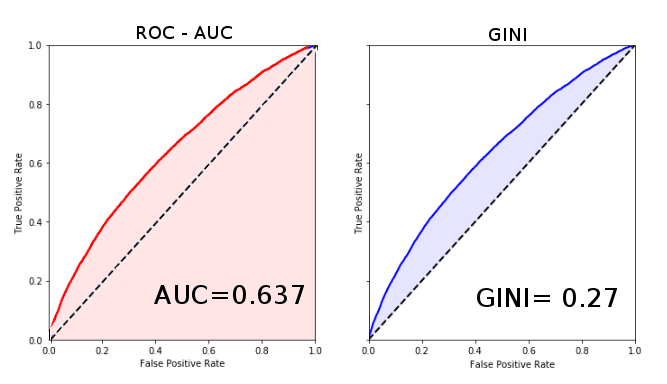
***AUC /GİNİ -*** AUC və GİNİ datalarımın bir-birindən nə dərəcə doğru ayırıb onu göstərir. Məsələn Yes-I No-dan, 0 1-dən, iti pişikdən və s. Əgər AUC və GİNİ 1disə demək datalarımı bir-birindən düzgün ayırıb.

AUC = 0.5 – 1 aralığında olduqda datamın yaxşı bölündüyün göstərir. Gini = 2AUC -1 -> 0 və 1 aralığında olur.

***ROC*** – datalarımı bir-birindən nə dərəcə doğru ayırıldığın göstərən graphdır. Datam kvadrat sahədə paylanır. ROC xətti yuxardan kvadratı tamalayanda demək datam bir-birindən tam doğru ayrılıb. Aşağı hissədən kvadratı tamamladıqda demək modelim datamı tamda biri birindən düzgün bölməyib.



**Week 6**

***Formatting data*** - Modeli qurmazdan əvvəl son addım olaraq, müəyyən məlumatlar üçün xüsusi format tələb edib-etmədiyini yoxlamaq üçün faydalıdır. Məsələn bir məhsul olan ardınca başqa bir məhsul alır onu sort edərək sıralayır. Model sizin üçün sorting apara bilsə belə, modelləşdirmədən əvvəl Sort node istifadə etmək emal vaxtına qənaət edə bilər. (Asc, Desc)

**Regression models**

Classification bu taskı edə bilərsən ya yox? Regression bu taskı neçə saniyə edə bilərsən?

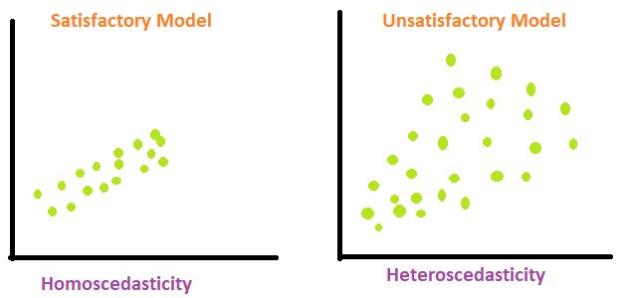
***Linear regression*** dependent variable ilə bir və ya bir neçə independent variable arasındakı əlaqəni təhlil etmək üçün istifadə edilən statistik modelləşdirmə üsuludur. O, verilənlərə ən yaxşı uyğun gələn və independent variables qiymətləri əsasında dependent variable dəyərlərini proqnozlaşdırmaq üçün istifadə edilə bilən xətti tənliyi tapmaq məqsədi daşıyır. Dependent və independent variablearim numeric olmalıdır. Outlier-a qarşı həssasdı. Modeldən əvvəl outlier-ları təmizləmək lazımdır. Təmizlənməsə xətt outlierları nəzərə almaq üçün düzgün çəkilməyəcək. Linear regression model qurmaq üçün, trendi tapmaq üçün və variable arasında əlaqəni baxmaq üçün istifadə olunur.

*Assumptions:*

* Residuals(xəta) meanleri 0 olmalıdır.
* Residualların homoscedasticity və bərabər variance
* Residuallar arasında autocorrelation olmamalıdır.
* Multicollinearity olmamalıdır.
* Residuals normal olmalıdır.

***Autocorrelation*** *-* Bir dəyərin özündən əvvəlki dəyərlə əlaqəsi olmasıdır. Tutalım satış dəyərləri verilib birinci gün 1000, ikinci gün 2000, üçüncü gün 3000. 3cü gün 3000 olması üçün 2ci gün 2000 olmalıdır ki, 1000 üzərinə gəlib 3000 olsun. 1000nin 2000lə, 2000nin 3000lə autocorrelation var.Əgər Durbin-Watson testi 2ə yaxındırsa, variablelar arasında autocorrelation yoxdur.

***Homoscedasticity*** *– Prediction dəyərlərimlə residuals arasında plot edirəm. Əgər əlaqə* homoscedasticitydirsə satisfactory modeldir. Yəni model qura bilərəm.



***Multicollinearity*** – Cədvəldə 3 column yaratmısan: hours of study, minute of study, days of study. Hər üçüdə eyni domene aiddir, üçüdə study duration göstərir. Aralarında bir-birilərinə təsirləri var. Multicollinearity olmaması üçün ya bir column saxlıyırsan yada hamısın bir columnda cəmləşdirirsən. VIF dəyəri çıxır, əgər 10dan yuxarıdısa iki column multicollinearitydır.

Regression model 2 cür olur: *Simple linear regression* and *Multiple linear regression*

***Sample linear regression*** - linear regression kimi tanınan sample linear *bir* independent variable və dependent variable əhatə edir. Sample linearregression üçün tənlik aşağıdakı kimi yazıla bilər:

Y = kx + b + ξ

Y dependent variable

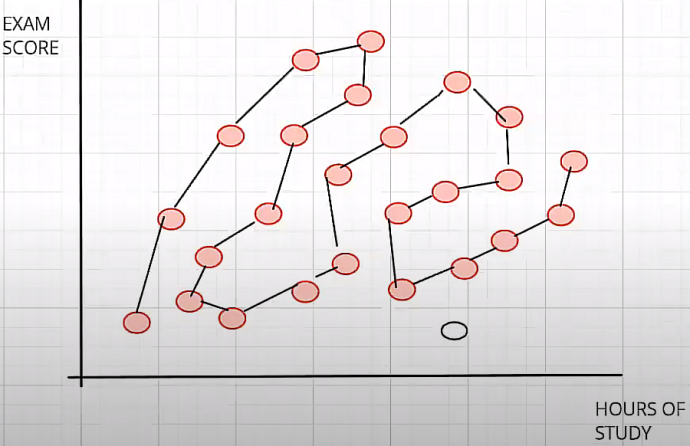
X independent variable

b kəsişmədir(intercept) (X 0 olduqda Y-nin qiyməti)

k slope (X-də bir vahid dəyişikliyi üçün Y-də dəyişiklik)

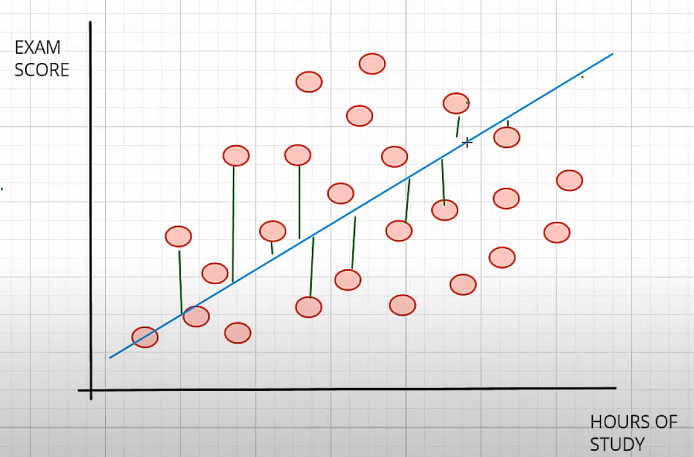
ξ error terminidir (qalıqlar və ya izah olunmayan dəyişmə)

*Example:* Siz müəllimsiz iki gün imtahan keçirirsiz. Siz cədvəldə ilk imtahan üçün hər bir uşağın imtahana neçə saat hazırlaşdığını və imtahan nəticələrini qeyd edirsiniz. Sabahakı imtahana görə şagird soruşsa ki neçə saat hazırlaşsam neçə alaram cədvələ baxıb: model şəkilində dəqiq və ya indivudual olaraq ortalama təxmini qiymət deyə bilərsən. Model olaraq bir tənlik qurursan Y = 60 + 5x. Sabah ki imtahan üçün şagirdə heçnə oxumasa min 60, 2 saat oxusa 70 alacağın deyirsiz. Lakin şagird imtahanda 74 qiyməti alır. 4 burda xətadır(error). Prediction olan 70dir, actual olan 74. Düstur dəyişilir və Y = kx + b + ξ olur. Yəni düstura əsasən model qurur. Əgər yeni şagird olsa düsturdakı 60, 5 və error dəyişilir. Çünki ən ideal düsturu almağa çalışıram, düzgün prediction verə bilmək üçün.



Y = 60 + 5x qrafiki. Overfittingdir. Hamısı əzbərləmişəm xətti ona görə qurmuşam. Yeni dəyər əlavə etdikdə error olacaq.

*Ən optimal düstur və ən optimal xətt odur ki, errorlarım min olsun. Əgər x = 0 olsa b minimum dəyər alacaq və y xəttin kəsəcək. K dəyər versək xətt dəyişiləcək. Yəni ən optimal düstur və ən optimal xətt tapmaq üçün k və b-yə optimal dəyər verməliyəm ki, error min olsun. Error = actual–prediction. Xətti dəyişdirirəm və düstura(*Y=kx + b + ξ) *əsasən ortalama hesablayıram. Ən kiçik ortalamaya əsasən xəttimi çəkirəm bu ən ideal xətt olur.*



*Qırmızılar və göy xətt arasındakı yaşıl xətt errordu və məsafə azdır. İdeal xətt və qrafik bu cür qurulur.*

***Multiple linear regression*** - linear regression kimi tanınan sample linear *bir neçə* independent variable və dependent variable əhatə edir. Multiple regressionda X-lar çoxaldıqca xətti daha ideal çəkmək olur. Çünki düsturda hər şey nəzərə alınır. Təbiətdə linear regression hadisələri çoxu multiple linear regressiodır. Prediction actuala daha yaxın olur. Multiple linearregression üçün tənlik aşağıdakı kimi yazıla bilər:

Y = 60 + 5x + 7\*(last exam score) + 0.5\*(participation)

*Example:* Evin qiymətin prediction verəndə otaq sayı, otağın kvadratı və s. nəzərədə alınır.

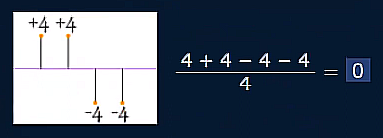
***Auto numeric node*** - say əlavə edirsən ona uyğun ən yaxşı modelləri sənə göstərir.

***Ensemble modelling*** - auto numeric node ilə əldə etdiyimiz ən yaxşı modelləri bir-birinə birləşdirik ki, daha yaxşı nəticə əldə edək.

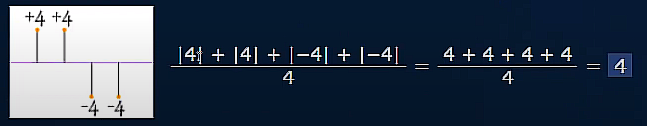
**Regression Model Evaluation**

***R square*** - 0 və 1 olur. R square əgər 1sə 100% sənin model qurmağın hər kəsə ortalama qiymət söyləməyindən yaxşıdır. R square əgər 0 olsa model qursanda, ortalama qiymət söyləsəndə eyni olacaq. Yəni sənin xətaların iksində də bərabər olacaq. Bu halda istəsən modeldə qurmaya bilərsən. Tutalım R square = 0.7 olanda mənim modelim hamıya ortalma deməkdən 70% yaxşıdır. Çünki X nəzərən prediction veririk. Yəni x-lar kömək edirki 70% düzgün prediction verim, digər 30% isə əgər əlavə column olsa idi (əlavə x-lar) daha düzgün prediction verə bilərdin.

***Mean Absolute Error (MAE)*** – Bütün errorları götürür, modula salır, hamısın toplayır və ortalamasın tapır. K hansı dəyərində, xəttin hansı versiyasında error azdısa xətti elə çəkirəm. MAE = 0 olmalıdır ki, error 0-a bərabər olduğun görüm.



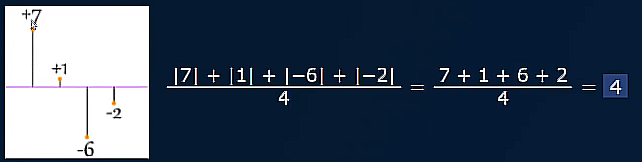
Bu halda MAE = 0 olur bu düzgün şəkildə hesablanma deyil.



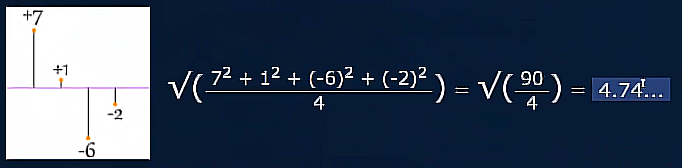
Bu errorlar eyni dərəcədə olanda doğrudur. Hamısı 4dür.

***Mean Squared Error (MSE)*** - Bütün errorları götürür kvadrata yüksəldir, onları toplayır ortalamasın tapır.

***Root Mean Squared Error (RMSE)*** - Bütün errorları götürür kvadrata yüksəldir, onları toplayır ortalamasın tapır kök altına salır.

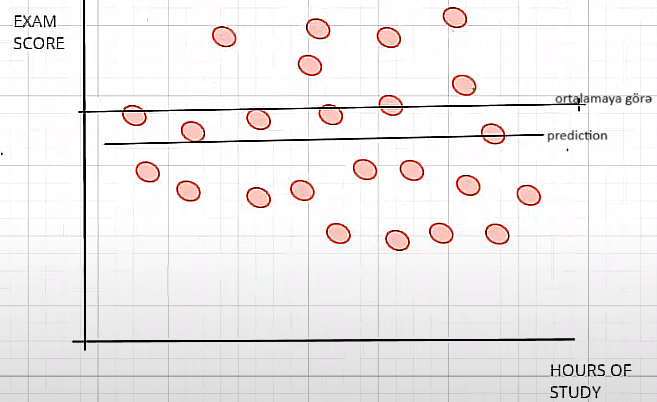


Errorlar eyni dərəcədə olmadığı üçün MAE doğru nəticə vermir. Outlier var. Buna görə RMSE istifadə edirik.



Daha dəqiq nəticə verir çünki böyük dəyərlərin square daha yüksəkdir.

*MAE və RMSE baxırsan əgər dəyərlər bir-birinə yaxındırsa 4.74 erroru nəzərə ala bilərsən. Yox əgər MAE = 4 və RMSE = 7 onda demək sənin modelində outlier var. Çünki outlierlar squareları nəticəni böyüdür.*

**

R square və MAE hər iksidə o zaman 0 bərabər olurki data pointlərim ortalama ətrafın da paylanıb.

**Week 7**

**Unsupervised learning**

Maşın öyrənməsində unsupervised learning, burada model unlabeled dataset üzərində öyrədilir və məqsəd açıq-aşkar təlimat olmadan verilənlər daxilində patterns, strukturları və ya əlaqələri tapmaqdır.

*Example:* müxtəlif rəngli toplardan ibarət böyük bir qutunuz var, ancaq onlar haqqında heç nə bilmirsiniz - rənglər üçün heç bir etiket və ya ad yoxdur. Tapşırıq bu topları oxşarlıqlarına görə qruplara ayırmaqdır. Hər bir topun hansı qrupa aid olduğunu deyən müəlliminiz yoxdur: bunu özünüz müəyyənləşdirməlisiniz.

Unsupervised learning də kompüteriniz buna bənzər bir şey edir. Bu, bir dəstə məlumat alır və bu nümunələrin və ya qrupların nə olması lazım olduğunu söyləmədən, onun daxilində nümunələri və ya qrupları tapmağa çalışır. Kompüter məlumat nöqtələri arasında oxşarlıq axtarır və onları birləşdirir. Bu, marketinq kampaniyası üçün oxşar müştərilər tapmaq və ya mətndə oxşar sözləri qruplaşdırmaq kimi datanız haqqında maraqlı şeyləri kəşf etməyə kömək edə bilər.

***Unlabeled dataset*** – etiketlənməmiş. Məlumat daxil edildiyi datasetdə hər bir data kategorilənmemiş və ya təyini edilməmişdir. Unlabeled datasetlər, ümumilikdə model təcrübələr aparmaq üçün istifadə edilir və bu məlumatlar üçün əlavə analiz və tədqiqat tələb olunur. Column adı yoxdu nə nədi deyə.

*Differences Supervised Learning and Unsupervised learning:*

Now, let's compare unsupervised learning to supervised learning:

*Guidance:*

Supervised: In supervised learning, sizə düzgün cavabları deyən müəlliminiz var. Siz həm inputun, həm də istədiyiniz outputun təmin olunduğu nümunələrdən öyrənirsiniz.

Unsupervised: In unsupervised learing, sənin müəllimin yoxdur. Düzgün cavablar verilmir. Siz onun xas nümunələrini tapmaq üçün məlumatları araşdırırsınız.

*Labels:*

Supervised: Labeled examples modeli nə gözlədiyini öyrətmək üçün istifadə olunur. Məsələn, siz pişik və itlərin şəkillərini labeled edə bilərsiniz ki, onları tanımaq üçün bir model qura.

Unsupervised: Öyrənməyə istiqamət verən heç bir labeled yoxdur. Model datalarla öz strukturunu və ya münasibətlərini tapır.

*Purpose:*

Supervised: Məqsəd adətən yeni, görünməmiş inputlar üçün düzgün outputlar proqnozlaşdırmaqdır. Proqnozlaşdıra bilirəm bir targetim var.

Unsupervised: Məqsəd çox vaxt data-nı araşdırmaq, oxşarlıqlar, fərqlər və ya klasterlər tapmaq və ya məlumatların mürəkkəbliyini azaltmaqdır. Var olmuş datalarla bir nəticəyə gəlirəm.

*Types of Algorithma:*

***Clustering algorithms*** are a fundamental part of unsupervised learning. Onlar oxşardata pointləri müəyyən xüsusiyyətlərə və ya xarakterikalara əsaslanaraq qruplaşdırmaq məqsədi daşıyır. Klasterləşmə, data pointlər arasında patterns və əlaqələri müəyyən etməklə datasetə məxsus strukturu anlamağa kömək edir.

***Association rules*** are another important concept in dataların öyrənilməsində və machine leaning də , xüsusən də market basket analysis sahəsində daha bir vacib anlayışdır. Tez-tez tranzaksiya və ya koteqorik datalarda tapılan, datasetdəki elementlər arasındakı əlaqələri aşkar etməyə kömək edir. Association rules "Müştəri A elementini alırsa, başqa hansı əşyaları alacaq?" kimi suallara cavab verməyə kömək edir. Bu qaydalar məhsul placements, promations və recommendations optimallaşdırmaq üçün bizneslər üçün dəyərli anlayışlar təmin edə bilər.

*Real life examples*

**Clustering Algorithms**

***Customer Segmentation:*** Müəssisələr tez-tez clustering algorithmdən istifadə edərək, müştərilərini davranışlarına, seçimlərinə və satınalma nümunələrinə əsasən fərqli qruplara bölürlər. Bu, müxtəlif müştəri seqmentləri üçün marketinq strategiyalarının uyğunlaşdırılmasına kömək edir. Məsələn, bir e-ticarət şirkəti müştəriləri "Tez-tez Alış-veriş edənlər", Sövdələşmə Axtaranlar" və "Yüksək səviyyəli Bayers" kimi kateqoriyalara qruplaşdırmaq üçün clustering (klasterləşdirmədən) istifadə edə bilər.

***Tibbi Təsvirdə Segmentation:*** Tibbi görüntüləmədə clustering algorithmi, MRT taramasında müxtəlif toxumaların müəyyən edilməsi kimi şəkillərin içərisində müxtəlif strukturları bölmək üçün istifadə edilə bilər. Bu, həkimlərə diaqnoz və müalicənin planlaşdırılmasında kömək edir.

***Social Network Analysis:*** Clustering algorithmləri sosial şəbəkələrdəki icmaları müəyyən edə bilər. Bu, insanların ümumi maraqlara və ya qarşılıqlı təsirlərə əsaslanaraq necə əlaqəli olduğunu və qruplaşdırıldığını anlamağa kömək edir. Məsələn, sosial media platformasında dost qruplarının müəyyən edilməsi.

***Xəbər Məqaləsinin Kateqoriyaya salınması:*** Oxşar xəbər məqalələrini məzmununa görə qruplaşdırmaq üçün clustering istifadə edilə bilər. Bu, xəbər platformaları üçün məzmun tövsiyəsi və təşkilində kömək edir.

***Kibertəhlükəsizlikdə anomaliyaların aşkarlanması:*** Clustering algorithmləri şəbəkə trafikində qeyri-adi nümunələri və ya davranışları aşkar etməyə kömək edə bilər. Potensial kiber təhdidləri və ya hücumları göstərə biləcək anomaliyaları müəyyən edir.

**Association Rules**

***Market Basket Analysis:*** Association rules ən məşhur nümunələrindən biri pərakəndə satışda market basket analysisdir. Məsələn, uşaq bezləri alan müştərilərin tez-tez körpə formulası və uşaq salfetləri aldığını aşkar etmək. Bu məlumat mağaza planının optimallaşdırılması və məqsədyönlü tanıtımlar üçün istifadə edilə bilər.

***Onlayn Məhsul Tövsiyələri:*** E-ticarət platformaları müştərilərə baxış və alış tarixçələrinə əsaslanaraq məhsulları tövsiyə etmək üçün association rules istifadə edir. Müştəri noutbuka baxarsa, platforma noutbuk çantaları, siçanlar və ya proqram təminatı kimi əlaqəli elementləri təklif edə bilər.

***Səhiyyə Müalicə Modelləri:*** Müalicə kombinasiyalarında modellər tapmaq üçün association rules tibbi datalar tətbiq edilə bilər. Bu, tibb işçilərinə müəyyən şərtlər üçün hansı dərmanların və ya müalicələrin tez-tez birlikdə təyin olunduğunu anlamağa kömək edə bilər.

***Web Page Analysis:*** Web analtikasında association rules istifadəçilərin birlikdə ziyarət etməyə meylli olduğu səhifələrin ardıcıllığını aşkar edə bilər. Bu məlumat vebsayt dizaynına və məzmunun təşkilinə istiqamət verə bilər.

***Fırıldaqçılığın aşkarlanması:*** Maliyyə əməliyyatlarında association rules saxtakarlığı aşkar etməyə kömək edə bilər. Müəyyən qeyri-adi əməliyyatların ardınca tez-tez digər spesifik hərəkətlər gəlirsə, bu, potensial saxtakarlığı göstərə bilər.

Association rules 2 alqoritmə bölünür: Apriori algorithm və Sequence algorithm. Antecedents, Consequents, Confidence, Rule support, Antecedent support, Lift *hər ikisinədə* məxsusdur.

***Apriori alqoritmi*** dataların əldə edilməsi və association rules öyrənilməsində klassik və geniş istifadə olunan alqoritmdir. O, xüsusi olaraq tranzaksiya datasetdən tez-tez element dəstlərini və association rules aşkar etmək üçün nəzərdə tutulmuşdur. Alqoritm iterativ olaraq candidate element dəstlərini yaratmaqla və minimum support tresholda cavab verməyənləri budamaqla işləyir. Bir element dəstinin dəstəyi həmin element dəstinin göründüyü tranzaksiyaların nisbətidir.

*Examples:* Bir ərzaq mağazası nümunəsini nəzərdən keçirək. Tutaq ki, hər bir əməliyyat müştəri tərəfindən birlikdə satın alınan əşyaları təmsil edən bir əməliyyat məlumat dəstiniz var. Budur bəzi əməliyyatlar:

Transaction I: Bread, Milk, Eggs

Transaction 2: Bread, Diapers(uşaq bezi), Beer

Transaction 3: Milk, Diapers, Beer, Coke

Transaction 4: Bread, Milk, Diapers, Beer

Transaction 5: Bread, Milk, Diapers, Coke

Nümunəyə əsasən association rule üçün lift, antecedent, consequent, və support hesablayaq: {Bread} -> {Milk}.

Əvvəlki nümunədən aşağıdakı məlumatları əldə edirik:

- Support({Bread}): 4 out of 5 transactions

- Support({Milk}): 4 out of 5 transactions

- Support({Bread, Milk}): 3 out of 5 transactions

***Metrics***

*I. Support:*

*Support – bütün mümkün tranzaksiyların neçəsində çörək olub.*

- Support({Bread}) = 4 / 5 = 0.8 (80%)

- Support({Milk}) = 4 / 5 = 0.8 (80%)

- Support({Bread, Milk}) = 3 / 5 = 0.6 (60%)

*2. Confidence*:

*Confidence – çörək alan tranzaksiyların neçəsində süd alıb. Nə qədər confidencedir? 75%*

- Confidence({Bread} -> {Milk}) = Support({Bread, Milk}) / Support({Bread}) = 0.6 / 0.8 = 0.75 (75%)

*3. Lift:*

*Lift – bu ikilinin birlikdə alınma ehtimalı tək-tək alınma ehtimalından nə qədər çoxdur?*

- Lift({Bread} -> {Milk}) = Support({Bread, Milk}) / (Support({Bread}) \* Support({Milk})) = 0.6 / (0.8 \* 0.8) = 0.9375

- Antecedent: {Bread} (If) 1ci bu, bunu alan nəyi alıb

- Consequent: {Milk} (then) 2ci bu, sonra bunu alıb

- Support({Bread]) = 0.8 (80%)

- Support({Milk}) = 0.8 (80%)

- Support({Bread, Milk}) = 0.6 (60%)

- Confidence({Bread} -> {Milk}) = 0.75 (75%)

- Lift({Bread} -> {Milk}) = 0.9375

*Interpretation:* The interpretation of the metrics is as follows:

* + Support({Bread}) Çörəyin tranzaksiyalarda nə qədər tez-tez göründüyünü ifadə edir.
  + Support({Milk}) Südün tranzaksiyalarda nə qədər tez-tez göründüyünü göstərir.
  + Support({Bread, Milk}) Çörək və Südün nə qədər tez-tez birlikdə göründüyünü ifadə edir.
  + Confidence({Bread] -> {Milk}) göstərir ki, Çörək alındıqda 75% hallarda Süd də alınır.
  + Lift({Bread} -> {Milk}) Çörəyin və Südün birlikdə alınma ehtimalı, çörəyin tək alınma ehtimalından 0,9375 qat daha çoxdur.

{Bread} -> {Milk} association rule üçün qaldırma dəyəri təxminən 0,93-dür, bu da Çörəyin alınması ilə Südün alınması arasında müsbət əlaqə olduğunu göstərir. Çörək alan müştərilər Südü müstəqil olaraq aldıqları zamanla müqayisədə təxminən 0,93 dəfə daha çox Süd alırlar. Bu onu göstərir ki, Çörək satışı Süd satışına orta dərəcədə müsbət təsir göstərir.

***Lift = 1: The iterms X and Y are independent of each other. X və Y ayrılıqda da yaxşı alınırlar, birlikdə də yaxşı alınırlar.***

***Lift > 1: The items X and Y are positively correlated. X-in baş verməsi Y-nin ehtimalını artırır və bunun tərsidə ola bilər. 1-dən çox olması X və Y arasında "müsbət" və ya "ilgili" əlaqəni göstərir.***

***Lift < I: The items X and Y are negatively correlated. X-in meydana gəlməsi Y-nin ehtimalını azaldır və bunun tərsidə ola bilər. 1-dən az olması X və Y arasında "mənfi" və ya "daha az ilgi çekici" əlaqəni göstərir.***

***Sequence*** ***algorithms*** ardıcıl datalar daxilində patternləri təhlil etmək və aşkar etmək üçün istifadə olunur. Ardıcıl məlumatlar, elementlərin sırasının vacib olduğu xüsusi bir sıra və ya hadisələr ardıcıllığı ilə xarakterizə olunur. Bu alqoritmlər xüsusilə vaxt seriyası məlumatları, mətn təhlili və məlumatların müvəqqəti sırasının vacib olduğu istənilən ssenari ilə bağlı tapşırıqlar üçün faydalıdır. Budur bəzi ümumi ardıcıllıq alqoritmləri:

***Retail and Market Basket Analysis:*** Sequence mining algorithms, müəyyən bir ardıcıllıqla birlikdə tez-tez satın alınan əşyaların nümunələrini tapmaq üçün alış-veriş tranzaksiya məlumatlarını təhlil etmək üçün istifadə olunur. Bu məlumat pərakəndə satıcılara müştərilərin seçimlərini başa düşməyə və mağaza planlarını və məhsul yerləşdirmələrini optimallaşdırmağa kömək edir. Məsələn, müştərilərin çörək almadan əvvəl tez-tez süd aldığını aşkar etmək bu əşyaları bir-birinə daha yaxın yerləşdirməyə səbəb ola bilər.

***Web Clickstream Analysis:*** Onlayn istifadəçi davranışının təhlilində sequence mining algorithms istifadəçilərin vebsaytda gəzdiyi sifarişdəki nümunələri müəyyən etməyə kömək edir. Bu məlumat vebsayt dizaynını optimallaşdırmaq, istifadəçi təcrübəsini təkmilləşdirmək və məzmunun yerləşdirilməsi ilə bağlı məlumatlara əsaslanan qərarlar qəbul etmək üçün çox vacibdir.

***Biological Sequence Analysis:*** DNT, RNT və zülal ardıcıllıqlarını təhlil etmək üçün bioinformatikada sequence algorithms tətbiq edilir. Məsələn, ümumi gen tənzimləmə ardıcıllığını müəyyən etmək, genlərin xüsusi şəraitdə necə aktivləşdiyini anlamağa kömək edir.

***Customer Journey Analysis:*** Marketinqdə sequence algorithms müştərilərin reklamlar, sosial media və alışlar kimi müxtəlif əlaqə nöqtələrində marka ilə ardıcıl qarşılıqlı əlaqəsini təhlil etməyə kömək edir. Bu fikir marketinq strategiyalarını müştəri davranışına uyğunlaşdırmağa kömək edir.

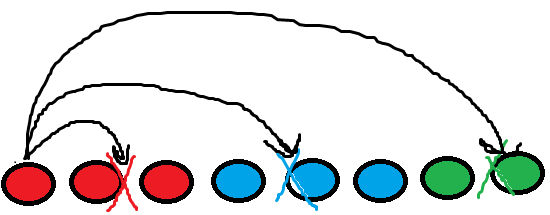
**Aprioridə *hansı birinci satılılan hansısdır fərqi yoxdu. Sequence da zaman yönümlüdü, ardıcıldı birinci satılan önəmlidir(if birinci gəlir).***

**CLUSTERING ALGORITHMS**

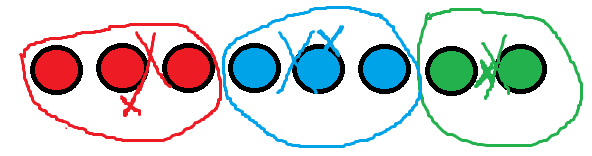
Clustering algorithms oxşar məlumat nöqtələrini oxşarlıqlarına əsasən qruplar şəklində qruplaşdırmaq üçün istifadə edilən unsupervised learning texnikasının bir növüdür. Məşhur clustering algorithmlərdən bəzi nümunələri:

***K-Means:*** K-Means ən çox istifadə edilən clustering algorithmlərindən biridir. O, məlumatları K klasterlərinə bölür, burada K istifadəçi tərəfindən müəyyən edilmiş parametrdir. O, iterativ olaraq məlumat nöqtələrini ən yaxın klaster mərkəzinə təyin edir və sonra konvergensiyaya qədər mərkəzləri yenidən hesablayır. K-Means effektivdir və dəqiq müəyyən edilmiş, sferik klasterləri olan məlumatlar üçün yaxşı işləyir. *K-Means adına əsasən means – ortalamaya görə, K-Median isə mediana görə baxacaq.*

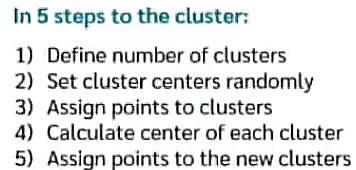
*Example:* 3 dənə random point atır: qırmızı, göy və yaşıl. Bir-bir, hər bir data point üçün random pointlər arası məsafəyə baxır. Ən yaxın məsafədə olan random pointə assign(təyin) edir.

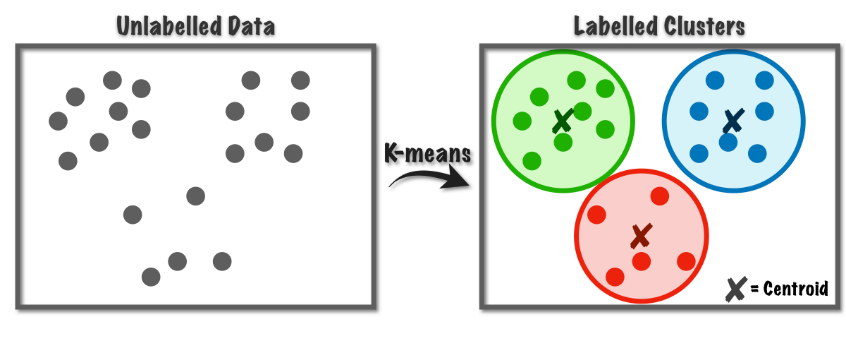


Hər birini bir clustering bölür. Hər bir clustering mərkəzin (ədədi ortasın) tapır. Sonra yeni data point təyin edir (mərkəzdə).

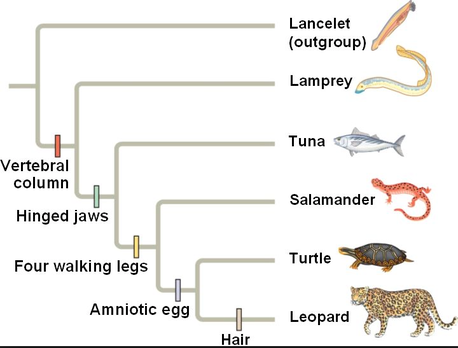


Təkrar bu proses davam edir. New update olunmuş mərkəzi data pointlə digər pointlərin məsafəsinə baxır. Proses o vaxta kimi davam edir ki, heç bir data point rəngini, yerini dəyişmir.





***Hierarchical Clustering:*** Hierarchical clustering dendroqram kimi tanınan klasterlərin tree-like structurunu, oxşarlıqlarına əsaslanaraq, clusterləri təkrar-təkrar merging və ya splitting qurur. Agglomerative (aşağıdan yuxarı) və ya divisive (yuxarıdan aşağı) ola bilər. Klasterlərin sayı əvvəlcədən müəyyən edilməmişdir, bu, hierarchical clustering daha çevik olmasına imkan verir.



*Example:* Turtle və Leopardı yumurta formasın görə, daha sonra Salamender əlavə edib 4 ayaqlığına görə, Tuna əlavə edib hinged jaws görə, Lamprey əlavə edib onurğa sütunu görə qruplaşdırıb. Minimum cluster size 2 olur.

***The Two Step*** Cluster algorithm böyük datasetlərini səmərəli idarə etmək üçün hierarchical clustering and non-hierarchical clustering (k-means) birləşməsindən istifadə edən hybrid clustering methodur. ***K-Means fərqili olaraq Two stepdə deyə bilmirəm bir neçə clustering qur, özü ən optimalın qurur.***

First Step (Hierarchical Clustering)

Second Step (K-Means Clustering)

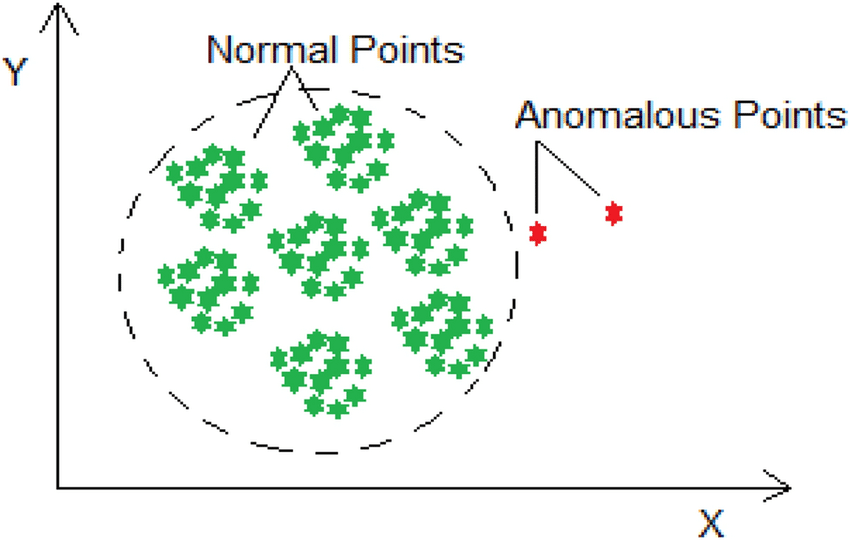
The hierarchical clustering nəticələri k-means algorithm üçün ilkin mərkəzləri yaratmaq üçün istifadə olunur. Hierarchial ilə iki hissəyə bölür. İkili sistemdi deyə minimum 2 olur. Sonra K-means tədbiq edir.

K-means clustering dataset də tətbiq edilir və məlumat nöqtələri ən yaxın klaster mərkəzinə təyin edilir.

***Kohonen networks*** klasterləşdirməni həyata keçirən bir növ neyron şəbəkələridir, həmçinin knet və ya self-organizing map kimi də tanınır. Başlanğıcda bu qrupların nə olduğunu bilmədiyiniz zaman bu network növü dataset də fərqli qruplara cluster üçün istifadə edilə bilər. Qeydlər elə qruplaşdırılıb ki, qrup və ya klaster daxilindəki qeydlər bir-birinə bənzəyir və müxtəlif qruplardakı qeydlər bir-birinə bənzəmir. Oxşar kombinasiyalar çox olduqda onları bir cluster add edir.

*Example:* Heyvanlar arasında kohonen network baxsaq. Goose, duck və dove birlikdə, hawk, eagle və owl birlikdə, wolf,dog və fox birlikdə oxşar kombinasiyalara görə bir qruplaşdıracaq.

***Anomaly algorithm:*** Anomaliyaların aşkarlanması alqoritmləri datanın əksəriyyətindən əhəmiyyətli dərəcədə kənara çıxan rare və ya unusual data points müəyyən etmək üçün istifadə olunur. Anomaliyaların aşkarlanması üçün statistik metodlar, sıxlığa əsaslanan metodlar və machine learning üsulları daxil olmaqla müxtəlif yanaşmalar mövcuddur. ***Anomaly algorithm də ardıcılıq önəmlidir. Məsələn, fraud detection da birinci baxmalıyıq nəyə click edib sonra nəyi click edib. Biz analizmizə sorted formada baxmalıyıq.***



*Example:* Tutalım bir sayt var. İnsanlar normal da tez-tez o sayta daxil olmurlar. Tez-tez click etmirlər. Normal bir halda insanlar girib bir məqalə oxuyur, çıxır və digər məqaləni oxuyur. Fraud detection da anomaliya alqoritminə görə kiberhücum olanda orda clicklərin sayı artır. Hər dəfə siqnal gedir və pointlər sayı çoxalır yəni normal haldakı pointlərin sayından kənara çıxır, bu isə anomalıya sayılır. Normal halla müqayisə edəndə məlum olur ki bu normal bir hadisə deyil. Onu artıq anomaliya kimi detect edəcək. Onu artıq fraud kimi dəyərləndirəcək.

**Deployment**

CRISP-DM methodologysın da son stagedir. Bütün stagelər tamamlanır sonda data sayta əlavə olunur(deployment). Məsələn evlə bağlı dataları add etmisən və hər hansı bir şəxs sayta daxil olub məlumatlarını əlavə etsə, istəyinə uyğun evin məlumatlarını əldə edəcək.