

DUOMENŲ TYRYBOS METODŲ TAIKYMAS NESTRUKTŪRIZUOTIEMS DUOMENIMS TIRTI

Maksim Jakovčenko ir Kristupas Viščiūna, lekt. Dainius Savulionis

Vilniaus kolegija, J. Vilnius, Jasinskio g. 15

Išvadas

Spartus technologijų vystymasis, verslo ir kitų veiklos sričių skaitmenizavimas lemia IT specialistų poreikį. Verslo įmonės, valstybinės institucijos bei nevyriausybinės organizacijos vis dažniau remiasi technologiniais sprendimais, siekdamos efektyvumo, inovacijų bei konkurencinio pranašumo. IT darbo rinka nuolat keičiasi, reaguodama į ekonominius, technologinius ir socialinius pokyčius, todėl IT specialistų darbo skelbimų ypač gausu įvairiuose darbo skelbimų portaluose. IT specialistai bei analitikai atlieka esminį vaidmenį diegiant ir administruojant informacines sistemas, kuriant programinę įrangą, analizuojant duomenis bei formuojant strateginius sprendimus, pagrįstus gautomis žiniomis. Spartus technologijų progresas, duomenų analitikos galimybės, dirbtinio intelekto plėtra ir kiekvienos įmonės duomenų kiekio augimas tik dar labiau stiprina šių profesijų aktualumą.

Vienas iš efektyvių būdų analizuoti darbo rinką, tai tyrinėti darbo skelbimuose įvardijamas kompetencijas. Tokie tyrimai leidžia įvertinti, kokių įgūdžių, patirties ir kvalifikacijos reikalauja dabartiniai darbo pasiūlymai, kokios yra atlyginimų tendencijos ir kokios darbo sąlygos dominuoja IT sektoriuje.

Mokslinė problema – Kaip pritaikyti duomenų tyrybos metodus, norint iš nestruktūrizuotų IT darbo skelbimų tekstų automatiškai išgauti ir struktūrizuoti atvaizduoti aktualiausias darbuotojų kompetencijas?

Straipsnio tikslas – pritaikyti duomenų tyrybos metodus (klasterizavimą ir asociatyvumo taisykles) nestruktūrizuotų darbo skelbimų duomenų tyrybai ir analitikai, siekiant identifikuoti ir struktūrizuoti svarbiausias IT srities specialistų kompetencijas.

Straipsnio uždaviniai:

1. Surinkti darbo skelbimus iš darbo skelbimų portalų.
2. Sistemingai suvesti atrinktus darbo skelbimus į Excel skaičiuoklę, kad būtų lengviau tvarkyti duomenis tolimesnei analizei.
3. Apdoroti susistemintus duomenis duomenų tyrybos ir analitikos metodais
4. Vizualizuoti gautus rezultatus, identifikuojant darbo rinkos tendencijas, poreikius ir dažniausiai pasikartojančius reikalavimus IT specialistams.

Tyrimo metodai - teorinės literatūros ir kitų informacijos šaltinių analizė, apibendrinimas, duomenų tyrybos ir analitikos metodų taikymas, grafinis duomenų vaizdavimas.

Tyrimo objektas - darbo skelbimų portaluose pateikiama informacija.

Duomenų analitikos metodų teoriniai aspektai literatūroje

Išanalizavus mokslinius šaltinius galima teigti, kad duomenų analitika yra susijusi su pridėtinės vertės sukūrimu iš turimos informacijos, t.y. informacijos analizės metu yra sukuriamos papildomos naujos žinios, galima įvertinti vykstančius procesus, išvelgti įvairias galimybes ir problemas. Duomenų analitika labai svarbi verslo procesų analizėje, rinkodaroje ir pan. (Corte - Real, 2019). Duomenų analitikos metodų taikymas sukauptos informacijos, žinių ir kitų duomenų apdorojimui, gali padėti priimti reikšmingus ir naudingus veiklos sprendimus (Jin, 2022).

2022 m. Dehbi ir kiti autoriai išskyrė 7V modelį, kuriame išskiriamos šios duomenų savybės: dydis (Volume), greitis (Velocity), įvairumas (Variety), Teisingumas (Veracity), vertė (Value), kintamumas (Variability), vizualizacija (Vizualization). Tokius duomenis labai svarbu sugebėti paversti naudinga informacija ir tinkamai tai panaudoti.

Moksliniuose informacijos šaltiniuose duomenų analitika skirstoma į šešias pagrindines dalis (Saggi, Jain, 2018), t.y. duomenų generavimas; duomenų gavimas; duomenų saugojimas; duomenų analitika; duomenų vizualizacija; sprendimų priėmimas.

Duomenų analitikos metodų taikymas yra svarbus įvairiose srityse, nes šiais laikais susiduriama su dideliais informacijos kiekiais, įvairiais informacijos pateikimo būdais ir pan.

M. Bahrami (2022) savo darbuose išskiria duomenų vizualizavimo svarbą, t.y. duomenų vizualizavimas padeda suprasti sudėtingą informaciją gautą iš įvairių didelių duomenų bazių.

P. Canonico (2022) teigia, kad verslo duomenų analitikos įrankiai bei žinių vizualizacija skatina žinių dalijimąsi ir įsisavinimą, o tai padeda priimti sprendimus.

Duomenų tyrybos teoriniai aspektai literatūroje

Duomenų tyryba – tai tarpdisciplininė sritis, apjungianti informacines technologijas, matematinę statistiką, mašininių mokymąsi ir duomenų bazes. Dažniausiai mokslinėje literatūroje duomenų tyryba apibrėžiama kaip duomenų analizės procesas, kurio tikslas yra atrasti paslėptus, anksčiau nežinomus, tačiau prasmingus ir naudingus duomenų modelius, ryšius ar tendencijas dideliuose duomenų rinkiniuose. Duomenų tyryba automatizuoja didelio kiekio duomenų analizę, kurios tikslas atrasti naudingas žinias ir modelius iš heterogeninių duomenų šaltinių (Shu & Ye, 2023)

Pagrindiniai duomenų tyrybos etapai:

1. **Duomenų rinkimas.** Duomenys surenkami iš įvairių šaltinių: jutiklių rodmenų, transakcijų duomenų bazių, žurnalų įrašų, tekstų, vaizdų, socialinių tinklų pranešimų ir kt. Tikslas – gauti pakankamai platų ir reprezentatyvų duomenų rinkinį, leidžiantį atskleisti paslėptus ryšius. Tradiciniai statistikos metodai ne visada tinka didelės apimties ir įvairialypiams duomenims, todėl būtini pažangesni sprendimai (Shu & Ye, 2023).
2. **Duomenų paruošimas.** Werner de Vargas ir kiti autoriai (2023) apžvelgia pagrindinius duomenų paruošimo etapus:
 - a. Duomenų valymas, t.y. anomalijų aptikimas ir šalinimas.
 - b. Triukšmo mažinimas, t.y. pasikartojančių įrašų šalinimas duomenų rinkinyje ir pan.
 - c. Transformavimas, t.y. normalizavimas, dimensių mažinimą, kategorinių požymių kodavimas.
3. **Metodų taikymas:**
 - a. **Klasterizavimas**, t.y. duomenų suskirstymas į grupes pagal panašumus, be iš anksto žinomų klasių. Algoritmas K-means laikomas vienu iš populiariausių ir dažniausiai naudojamų klasterizavimo algoritmų. Šis algoritmas turi keletą modifikacijų (Ikotun ir kt., 2023).
 - b. **Asociatyvumo taisyklės**, t.y. taisyklių, kurios apibrėžia kaip dažnai tam tikri duomenų elementai pasitaiko kartu, nustatymas (Shahin et al., 2021). Asociatyvumo taisyklių nustatymui naudojami Apriori, FP-Growth ir įvairūs kiti metodai.
 - c. **Klasifikavimas**, t.y. duomenų suskirstymas į iš anksto žinomas kategorijas ar klases pagal tam tikras savybes. Klasifikavimui dažnai naudojami sprendimų medžiai, logistinė regresija, neuroniniai tinklai.
 - d. **Regresija**, t.y. skaitinės reikšmės prognozavimas remiantis kitais duomenų požymiais. Dažniausiai naudojami metodai: linijinė regresija, daugianarė regresija, regresiniai medžiai, ROC-AUC, MAE, MSE (Rainio, Teuho & Klén, 2024).
4. **Rezultatų interpretavimas ir vertinimas:**
 - a. Klasterizavimo vertinimui dažniausiai naudojami (Kossakov ir kt., 2024):
 - i. Silhouette koeficientas, kuris parodo, kaip gerai objektas atitinka savo klasterį, palyginus su artimiausiu kitu klasteriu. Vertė nuo –1 iki 1 (kuo arčiau 1 – tuo geresnis klasterizavimas).
 - ii. Davieso-Bouldin indeksas, kuris įvertina klasterių kokybę pagal jų tankumą ir atstumą vienas nuo kito. Kuo mažesnė šio indekso reikšmė, tuo geresnis rezultatas.
 - iii. Calinski–Harabasz indeksas yra vidinis klasterizavimo kokybės vertinimo rodiklis, kuris parodo, kaip gerai duomenys susiskirstyti į klasterius. Kuo didesnė Calinski–Harabasz indekso reikšmė, tuo geresnis klasterizavimas, t.y.

klasteriai yra labai aiškiai atskirti vienas nuo kito ir klasterių viduje taškai yra glaudžiai susitelkę.

- b. Asociatyvių taisyklių vertinimui dažniausiai naudojami (Shahin et al., 2021):
 - i. Taisyklės dažnumas, kuris parodo, kaip dažnai taisyklėje minimas elementų rinkinys pasikartoja visame duomenų rinkinyje.
 - ii. Taisyklės patikimumas, kuris parodo, kokia tikimybė, kad pasirinkus elementą A, bus pasirinktas ir elementas B.
- c. Klasifikavimo rezultatų vertinimui dažniausiai naudojami (Rainio, Teuho & Klén, 2024):
 - i. Tikslumas, kuris parodo teisingų prognozių dalį visų prognozių atžvilgiu arba kiek iš visų modelio pateiktų teigiamų prognozių buvo teisingos.
 - ii. Atpažinimo rodiklis, kuris parodo kiek iš visų teigiamų atvejų modelis sugebėjo atpažinti.
 - iii. F1-matas, t.y. vidurkis tarp tikslumo ir atpažinimo.

Darbo skelbimų duomenų analizė

Profesijos pasirinkimas, profesinės karjeros planavimas yra svarbu kiekvieno žmogaus gyvenime. Nuo profesijos pasirinkimo priklauso būsimas darbas, pasitenkinimas gyvenimu, savirealizacija, padėtis visuomenėje ir pan. Renkantis profesiją daug informacijos galima gauti iš darbo skelbimų analizės. Darbo skelbimuose įvardijamos aktualios tam tikros profesijos kompetencijos, darbo užmokesčio dydis, darbo aplinka ir pan.

Pradėjus darbo skelbimų analizę, pirmiausia buvo susitelkta į Lietuvos darbo skelbimų portalus, kuriuose dominavo IT srities profesijos. Analizė buvo atlikta naudojant raktinius žodžius „Duomenų analitika“ ir „Programuotojas“. Atrinkti darbo skelbimai buvo sukelti į Excel darbo knygą, duomenys struktūrizuoti ir pasiruošti vizualizacijai su Power BI.

Darbo skelbimų duomenys buvo suskirstyti pagal keletą pagrindinių atributų, kad būtų galima atlikti kokybišką analizę ir vizualizaciją. Buvo išskirti šie atributai:

1. Pozicijos pavadinimas: Skelbime nurodytas profesijos pavadinimas.
2. Įgūdžiai ir reikalavimai: Kiekviename darbo skelbime reikalaujami techniniai ir minkštieji įgūdžiai (pvz., programavimo kalbos, analitiniai įrankiai, bendravimo gebėjimai ir pan.).
3. Privalumai: Kiekvienos įmonės siūlomi darbo vietos privalumai.
4. Šalis ir miestas: Informacija apie darbo vietą, skirta atskirti darbo skelbimus pagal regionus (pvz., Lietuva, Vilnius).
5. Darbo užmokestis: Jei nurodyta, įrašomas darbo užmokestis arba atlyginimo intervalas.

Šie duomenys buvo sukelti į Excel skaičiuoklę, užtikrinant, kad kiekvienas stulpelis atitiktų konkrečią kategoriją, tokią kaip pozicijos pavadinimas, miestas, šalis, reikalavimai ir pan. Power BI veikia efektyviausiai su tvarkingai struktūrizuotais duomenimis, kiekviena kategorija buvo suformatuota taip, kad duomenų tipai atitiktų Power BI reikalavimus (pvz. tekstas, skaičiai, data).

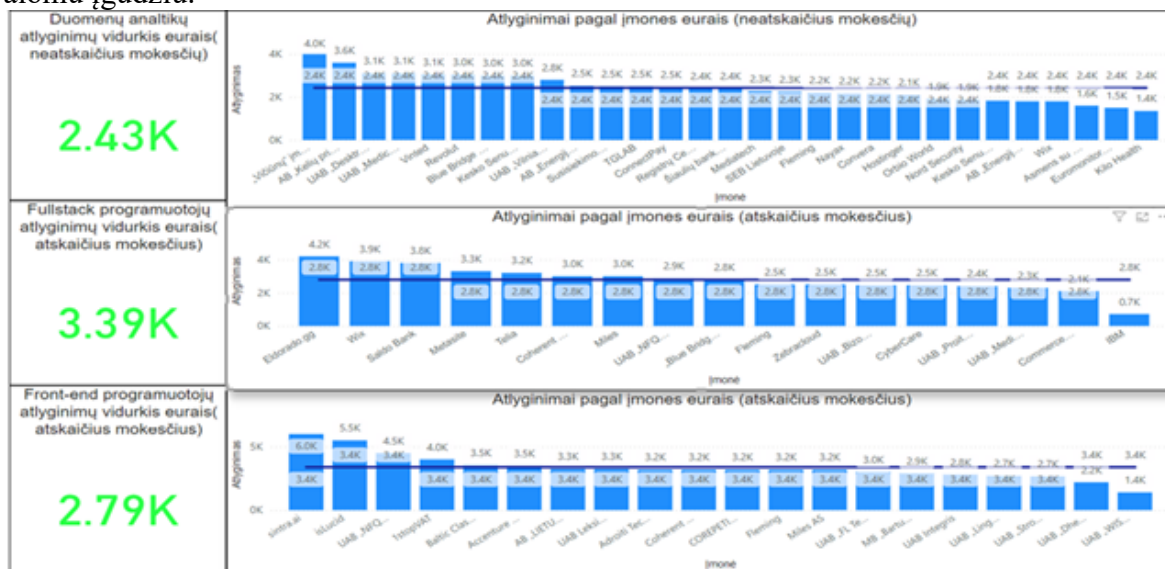
Duomenų vizualizacija leidžia aiškiai matyti darbo rinkos IT specialistų poreikį skirtingose šalyse, taip pat įžvelgti atlyginimų skirtumus ir keliamus reikalavimus įvairiose darbo pozicijose. Šios vizualizacijos padėjo išryškinti pagrindines darbo rinkos tendencijas bei atlikti tarptautinį lyginamąjį tyrimą.

Vizualizacijų ir analizės elementai:

1. Pozicijų reikalavimų vizualizacija: naudojant duomenų grafikus buvo palyginta, kokie įgūdžiai ir patirties lygiai yra reikalaujami darbo skelbimuose tam tikrose šalyse.
2. Reikalavimų dažnumo palyginimas: skirtingoms šalims buvo sukurti grafikai, kurie parodė kiek įmonių reikalauja tam tikrų specifinių įgūdžių. Ši vizualizacija leido matyti kiek dažnai pasikartodavo tam tikri reikalavimai ir kokie įgūdžiai yra laikomi esminiais.
3. Atlyginimų palyginimas tarp šalių: sukurti stulpeliniai grafikai, kuriuose buvo atvaizduotas IT specialistams siūlomas atlyginimas.

1 paveikslėlyje yra matomas vidutinis fullstack programuotojų atlyginimas siekia 3390 eurų, front-end programuotojų – 2790 eurų, o duomenų analitikų – 2430 eurų (neatskaičius mokesčių). Be to, iš grafikų galima matyti, kad konkrečių įmonių siūlomi atlyginimai ženkliai skiriasi.

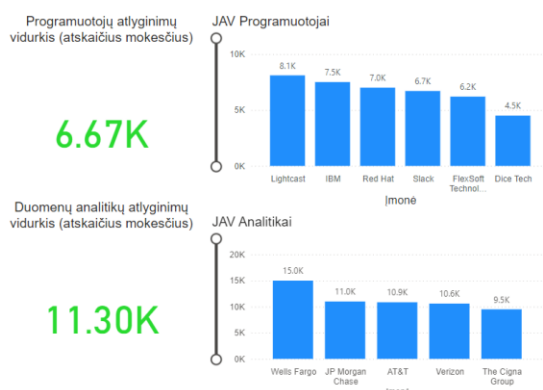
Tyrimo metu nustatyta, kad 90% įmonių reikalauja SQL, 58% – MS Power BI, 52% – Python, o 23% – R. Toks pasiskirstymas rodo, kad duomenų analitikams itin svarbu gebėti dirbti su duomenų bazėmis ir verslo analitikos įrankiais. Taip pat pastebėta, kad 100% front-end programuotojų reikalaujama JavaScript, o 88% – Agile metodikų, tuo tarpu Git (82%) ir OOP (71%) nurodo būtinybę išmanyti versijų kontrolę ir objektinio programavimo principus. Tokia reikalavimų struktūra atskleidžia, kad šiuolaikinis front-end specialistas turi derinti programavimo, komandinio darbo ir projektų valdymo kompetencijas. Stebima, kad 88% įmonių fullstack programuotojams kelia JavaScript reikalavimą, o Git ir OOP (po 71%) rodo būtinybę išmanyti versijų kontrolę ir objektinio programavimo principus. Be to, PHP (54%) išlieka aktualus, tačiau ne visose įmonėse laikomas privalomu įgūdžiu.



1 pav. Lietuvos rinkos atlyginimai
Šaltinis: Sudaryta autorių pagal surinktus duomenis

JAV darbo rinka ženkliai skiriasi nuo Europos darbo rinkos ne tik atlyginimo dydžiu, bet ir kompetencijomis. 2 paveikslėlyje yra matoma, kad JAV duomenų analitikų vidutinis atlyginimas 11300 yra didesnis nei programuotojo vidutinis atlyginimas 6670. Diagramose taip pat matyti, kaip šios specialybės vertinamos skirtingose įmonėse, o tai leidžia spręsti apie didėjančią duomenų analizės svarbą darbo rinkoje. JAV stebima, kad Git yra privalomas įgūdis visiems (100% reikalavimų), o OOP ir MySQL reikalauja net 83% įmonių. Be to, 67% darbdavių pabrėžia debesų technologijų (Azure/AWS) svarbą, rodančią, jog modernioje darbo rinkoje vertinama patikima versijų kontrolė, tvirtas objektinio programavimo supratimas ir gebėjimas dirbti su debesų sprendimais. JAV 83% įmonių reikalauja Python ir Tableau, 67% – R, o 50% – Power BI. Tai rodo, kad JAV darbo rinkoje ypač vertinamos universalios programavimo kalbos bei verslo analitikos platformos.

JAV darbo rinkos programuotojų ir duomenų analitikų atlyginimai



2 pav. JAV rinkos atlyginimai

Šaltinis: Sudaryta autorių pagal surinktus duomenis

DAX (angl. *Data Analysis Expressions*) formules yra specializuota kalba, leidžianti kurti dinamiškus skaičiavimus ir analizes duomenų modeliuose, pavyzdžiui, Power BI aplinkoje. Jos suteikia galimybę apjungti, filtruoti ir manipuliuoti duomenimis, kad būtų galima gauti įžvalgas, kurios padeda priimti informatyvius sprendimus. Naudodami DAX, analitikai gali kurti tiek paprastas metrikas, tiek sudėtingus skaičiavimus, kurie atskleidžia tendencijas ir anomalijas dideliuose duomenų kiekiuose. Pavyzdžiui, naudojant DAX formulę galima lengvai nustatyti, kiek įmonių savo skelbimuose išryškina tokį privalumą, kaip draugiškas kolektyvas. DAX padeda ne tik automatizuoti sudėtingus skaičiavimus, bet ir parodyti duomenų poveikį verslo procesams. Ši formulavimo kalba yra neatsiejama priemonė modernioje analitikoje, leidžianti paversti skaitmeninius duomenis vertingomis įžvalgomis, kurios padeda tobulinti organizacijų veiklą.

1 lentelė. DAX formulės pavyzdys

DAX formulė	Friendly Team Positions = COUNTROWS(FILTER(Sheet4, Sheet4[Draugiškas kolektyvas] = 1))
Formulės aprašymas	Apskaičiuojami skelbimai kurie siūlo draugišką kolektyvą.
Rezultatas	17 Friendly Team Positions

1 lentelėje pateikta DAX formulė, trumpas formulės aprašymas ir gaunamas apdorojus duomenis rezultatas.

Šaltinis: Sudaryta autorių pagal surinktus duomenis

Duomenų tyrybos metodų taikymas

Tyrimo duomenų rinkinys sudarytas iš binarinių požymių, todėl asociatyviųjų taisyklių analizė (Apriori algoritmas) buvo pasirinkta siekiant atskleisti, kurios požymių kombinacijos dažniausiai sutampa darbo skelbimuose bei kurios sąsajos tarp reikalavimų yra patikimiausios. Hierarchinė klasterizacija leido grupuoti skelbimus pagal visų penkių požymių derinius ir pamatyti, kiek įmonių reikalavimus formuoja kompleksiniu (visapusišku) arba labiau taikiniu (ribotu) modeliu. Abu metodai – asociatyviosios taisyklės ir klasterizavimas – puikiai papildė vienas kitą. Pirmasis

atskleidžia tiesiogines sąsajas tarp reikalavimų, antrasis – bendresnius duomenų grupavimo dėsningumus.

2 lentelė. Lietuvos darbo rinkos asociatyvumo taisyklės

Jei darbo skelbime reikalaujama Python=T, tada reikalaujama Anglų kalba=T	Jei SQL=T ir Anglų kalba=T, tada Patirtis duomenų bazėse=T	Jei Python=T, tada Patirtis duomenų bazėse=T ir SQL=T
Beveik pusė (42%) visų skelbimų reikalauja Python ir iš jų didžioji dalis (81%) kartu prašo anglų kalbos.	Darbdaviai, kurie reikalauja, kad kandidatas mokėtų SQL ir turėtų anglų kalbos žinių, taip pat visuomet (100% atvejų) prašo turėti patirties su duomenų bazėmis.	Darbdaviai, kurie ieško Python specialistų, beveik visuomet (apie 94%) prašo SQL žinių ir patirties su duomenų bazėmis.

Lentelėje pateikti trys pagrindiniai Lietuvos darbo skelbimų asociatyvumo ryšiai su jų palaikymo ir patikimumo vertėmis:

1. **Python ⇒ anglų kalba** – 42 % skelbimų reikalauja Python, iš jų 81 % kartu reikalauja anglų kalbos mokėjimo.
2. **SQL + anglų kalba ⇒ DB patirtis** – kai reikalaujama SQL ir anglų kalba, visuomet (100 %) reikalaujama ir patirties su duomenų bazėmis.
3. **Python ⇒ SQL + DB patirtis** – 35 % skelbimų reikalauja Python, iš jų 94 % kartu reikalaujama SQL žinių ir duomenų bazių patirties.

3 lentelė. Vokietijos darbo rinkos asociatyvumo taisyklės

Jei darbo skelbime reikalaujama patirties duomenų bazėse, tada taip pat reikalaujama aukštojo išsilavinimo, SQL žinių, anglų kalbos gebėjimų ir biuras nurodytas kaip darbo privalumas	Jei darbo skelbime reikalaujama Python=T, tada reikalaujama Anglų kalbos=T	Jei skelbime reikalaujama Python = T, tada reikalaujama Anglų kalbos, aukštojo išsilavinimo ir patirties duomenų bazėse.
Maždaug du trečdaliai skelbimų, kuriuose ieškoma duomenų bazių specialistų, vienu metu pabrėžia darbą iš biuro, aukštąjį išsilavinimą, SQL, ir anglų kalbos žinias.	Beveik pusė (42%) visų skelbimų reikalauja Python, ir iš jų didžioji dalis (81%) kartu prašo anglų kalbos.	Daugiau nei puse skelbimų, kur reikalingas Python, visuomet reikalaujama aukštojo išsilavinimo, duomenų bazių patirties, ir anglų kalbos žinių.

Lentelėje pateikiami trys pagrindiniai Vokietijos IT-darbo skelbimų asociatyvumo ryšiai su jų palaikymo ir patikimumo rodikliais:

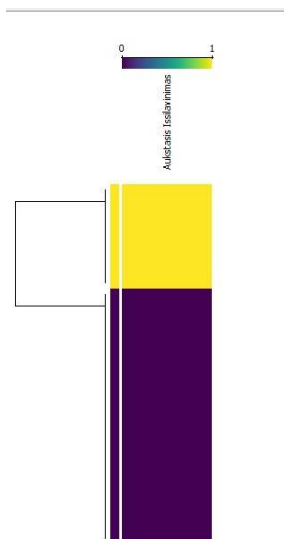
1. **Patirtis su duomenų bazėmis ⇒ aukštasis išsilavinimas + SQL žinios + anglų kalba + darbas biure** – apie du trečdaliai (≈ 66 %) skelbimų, kuriuose reikalaujama patirties su duomenų bazėmis, visuomet (100 %) pabrėžia aukštąjį išsilavinimą, SQL žinias, anglų kalbos gebėjimus ir darbą biure kaip privalumą.
2. **Python ⇒ anglų kalba** – beveik pusė (42 %) visų skelbimų reikalauja Python, ir iš jų didžioji dalis (81 %) kartu reikalauja anglų kalbos gebėjimų.
3. **Python ⇒ aukštasis išsilavinimas + duomenų bazių patirtis + anglų kalba** – daugiau nei pusė skelbimų, kuriuose reikalaujama Python, visuomet (100 %) nurodo reikalingą aukštąjį išsilavinimą, patirtį su duomenų bazėmis ir anglų kalbos žinias.

Klasterizavimas

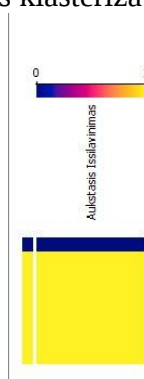
3 ir 4 pav. pateikti hierarchinio klasterizavimo pasiskirstymai pagal aukštąjį išsilavinimą Lietuvoje ir Vokietijoje, kurie parodo dvi skelbimų grupes pagal aukštojo išsilavinimo reikalavimą:

- **Geltonoji juosta (~ 90–95 %)** – didžioji dalis (apie 90–95 %) Vokietijos skelbimų **reikalauja** aukštojo išsilavinimo.
- **Violetinė juosta (~ 5–10 %)** – tik nedidelė dalis (~ 5–10 %) šios šalies skelbimų **jo nereikalauja**.

Lietuvoje aukštasis išsilavinimas dažniausiai **nėra** būtinas, o Vokietijoje – beveik **visada** reikalaujama aukštojo išsilavinimo.



3 pav. Lietuvos klasterizavimo pasiskirstymas



4 pav. Vokietijos klasterizavimo pasiskirstymas

Darbo skelbimų duomenų vizualizacija

Duomenų vizualizacija – tai esminė priemonė, leidžianti aiškiai ir suprantamai perteikti sudėtingą informaciją bei atrasti svarbias tendencijas. Šis metodas padeda analitikams ir tyrėjams greitai identifikuoti paslėptus duomenų ryšius, kas yra ypač svarbu sprendimų priėmimo. Python kalba išlieka itin populiari dėl savo lankstumo ir plačios specializuotų bibliotekų, tokių kaip Matplotlib, Seaborn ir Plotly. Naudojant šias priemones, galima greitai kurti tiek statines, tiek interaktyvias vizualizacijas, kurios padeda perteikti analizės rezultatus. Pavyzdžiui, Matplotlib leidžia lengvai pavaizduoti stulpelines, linijines ir skritulines diagramas, o Seaborn kurti estetiškai patrauklius grafikus su minimaliu kodo kiekiu. Nepaisant Python privalumų, rinkoje egzistuoja ir kiti modernūs įrankiai, tokie kaip Power BI, siūlantis interaktyvias ataskaitas bei realaus laiko duomenų analizę. Tableau ir QlikView išsiskiria savo intuityviomis naudotojo sąsajomis bei gebėjimu integruotis su įvairiais duomenų šaltiniais. Integruotas Python ir kitų įrankių naudojimas leidžia pasiekti hibridinius sprendimus, maksimaliai išnaudojant kiekvieno įrankio stipriąsias puses. Tokiu būdu, duomenų vizualizacija tampa ne tik informacijos perteikimo, bet ir analizės bei inovatyvių sprendimų priėmimo priemone. Šie įrankiai suteikia galimybę ne tik efektyviai apdoroti duomenis, bet ir kurti moksliskai pagrįstas ataskaitas, kurios yra lengvai interpretuojamos tiek specialistų, tiek plačiosios visuomenės.

Išvados

1. Darbo skelbimų surinkimas. Buvo atliktas darbo skelbimų surinkimas iš darbo skelbimų portalų, naudojant raktinius žodžius „duomenų analitikas“ ir „programuotojas“. Šis procesas leido surinkti tikslus darbo skelbimus, atitinkančius IT specialistų paieškos kriterijus, ir užtikrinti, kad jie būtų aktualūs pagal šias profesijas.

2. Atrinkimas ir duomenų struktūrizavimas. Iš Lietuvos ir Vokietijos darbo skelbimų portalų buvo atrinkti darbo skelbimai, kurie buvo suvesti į Excel lenteles. Šis žingsnis buvo svarbus, kad duomenys būtų lengvai tvarkomi, struktūrizuojami ir paruošti tolimesnei analizei, užtikrinant duomenų tikslumą ir atitiktį analizės tikslams.

3. Sisteminė duomenų suvestinė. Visi atrinkti darbo skelbimai buvo sistemingai suvesti į Excel skaičiuoklę, kas padėjo struktūrizuoti informaciją ir supaprastino duomenų analizę. Šis metodas leido lengvai atlikti duomenų valymą ir organizavimą, kas buvo būtina tolimesniems žingsniams – analizės atlikimui ir vizualizacijoms.

4. Duomenų klasterizavimas buvo atliktas naudojant atvirojo kodo programinę įrangą Orange, kuri skirta duomenų analizei, mašininio mokymosi metodų taikymui ir vizualizacijai.

5. Power BI vizualizacijos. Naudojant Power BI, surinkti ir suvesti duomenys buvo vizualizuoti, kas leido aiškiai matyti darbo rinkos tendencijas ir dažniausiai pasikartojančius reikalavimus IT specialistams įvairiose šalyse. Vizualizacijos padėjo identifikuoti pagrindinius darbo rinkos poreikius ir geriau suprasti, kokie įgūdžiai ir kompetencijos yra reikalingiausios darbo rinkoje. Svarbiausios programuotojo kompetencijos yra PHP, MySQL, JS, CSS, HTML, OOP, Unit testai, Python, Git, Agile ir kt., o duomenų analitiko pagrindinės kompetencijos yra Python, SQL, R, SAS, MS Power BI, IBM Cognos, Tableau, QlikView. Pastebėta, kad Python ir SQL kompetencijos svarbios tiek programuotojui, tiek duomenų analitikui.

Informacijos šaltiniai

Bahrami, M., Shokouhyar, S. (2022). The role of big data analytics capabilities in bolstering supply chain resilience and firm performance: A dynamic capability view. *Information Technology & People*, 35(5), 1621–1651. <https://doi.org/10.1108/ITP-01-2021-0048>

Canonico, P., De Nito, E., Esposito, V., Fattoruso, G., Pezzillo Iacono, M., Mangia, G. (2022). Visualizing knowledge for decision-making in Lean Production Development settings. Insights from the automotive industry. *Management Decision*, 60(4), 1076–1094. <https://doi.org/10.1108/MD-01-2021-0144>

Chen, M., Pu, X., Zhang, M., Cai, Z., Chong, A. Y.-L., Tan, K. H. (2022). Data analytics capability and servitization: The moderated mediation role of bricolage and innovation orientation. *International Journal of Operations & Production Management*, 42(4), 440–470. <https://doi.org/10.1108/IJOPM-10-2021-0663>

Chua, K.C., Henderon, C., Grey, B., Holland, M., Sevdalis, N. (2022). Evaluating quality improvement at scale: A pilot study on routine reporting for executive board governance in a UK National Health Service organisation. *Volume* 98. <https://doi.org/10.1016/j.evalprogplan.2022.102222>

Corte-Real, N.; Ruivo, P.; Oliveira, T.; Popovic A. Unlocking the drivers of big data analytics value in firms, *J. Bus. Res.* 2019. 97, 160–173

Dehbi, S., Lamrani, H. C., Belgnaoui, T., Lafou, T. (2022). Big Data Analytics and Management control. *Procedia Computer Science*, 203, 438–443. <https://doi.org/10.1016/j.procs.2022.07.058>

Jin, C., Li, F., Ma S., Wang, Y. (2022). Sampling scheme-based classification rule mining method using decision tree in big data environment. *Volume* 244. <https://doi.org/10.1016/j.knosys.2022.108522>

MySQL Documentation (2024). Prieiga per internetą: <https://dev.mysql.com/doc/>

Power BI documentation (2024). Prieiga per internetą: <https://learn.microsoft.com/en-us/power-bi/>

Python documentation (2024). Prieiga per internetą: <https://docs.python.org/3/>

Saggi, M. K., Jain, S. (2018). A survey towards an integration of big data analytics to big insights for value-creation. *Information Processing and Management*, 54(5), 758–790. <https://doi.org/10.1016/j.ipm.2018.01.010>

Shu, X., & Ye, Y. (2023). *Knowledge discovery: Methods from data mining and machine learning*. *Social Science Research*, 110, Article 102817. <https://www.sciencedirect.com/science/article/pii/S0049089X22001284?via%3Dihub>

Werner de Vargas, V., Schneider Aranda, J. A., dos Santos Costa, R., et al. (2023). *Imbalanced data preprocessing techniques for machine learning: A systematic mapping study*. Knowledge and Information Systems, 65, 31–57. <https://link.springer.com/article/10.1007/s10115-022-01772-8>

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). *K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data*. Information Sciences, 622, 178–210. <https://www.sciencedirect.com/science/article/abs/pii/S0020025522014633?via%3Dihub>

Shahin, M., Arakkal Peious, S., Sharma, R., Kaushik, M., Ben Yahia, S., Shah, S. A., & Draheim, D. (2021). Big data analytics in association rule mining: A systematic literature review. In *Proceedings of the 3rd International Conference on Big Data Engineering and Technology (BDET 2021)* (pp. 40–49). ACM. <https://dl.acm.org/doi/10.1145/3474944.3474951>

Rainio, O., Teuvo, J., & Klén, R. (2024). *Evaluation metrics and statistical tests for machine learning*. Scientific Reports, 14, 6086. <https://www.nature.com/articles/s41598-024-56706-x>

Kossakov, M., Mukasheva, A., Balbayev, G., Seidazimov, S., Mukammejanova, D., & Sydybayeva, M. (2024). Quantitative comparison of machine learning clustering methods for tuberculosis data analysis. *Engineering Proceedings*, 60(1), Article 20. <https://www.mdpi.com/2673-4591/60/1/20>