Bangladeshi naiste tervisenäitajate mõju rasedusaegsele riskihinnangule

Projekti raport

Hannes Aaresild,
Natalja Beljajeva,
Mari Helen Štarkov

Kõik projekti failid asuvad GitHubis:
https://github.com/aaresild/maternal-health-risk#maternal-health-risk

Sisukord

1 Sissejuhatus

Uurime andmeteaduse projektis Bangladeshi rasedate naiste tervisenäitajate mõju rasedusaegsele terviseriski hinnangule. Projektis kasutatud andmed saime Kaggle keskkonnast ning need koguti Bangladeshi maapiirkondade haiglatest, kliinikutest ja emaduskeskustest läbi IoT riski monitoorimise süsteemi.

1.1 Taust ja motivatsioon

Bangladeshis on puudus oskuslikest sünnitusabistajatest, eriti maapiirkondades. Bangladesh kaotab ligi 7 660 naist igal aastal ennetatavate põhjuste tõttu, mis on seotud raseduse ja sünnitusega. Bangladeshi emade suremuse määr on 245 naist 100 000 elussünni kohta.(The Lancet 2015) Umbes 71% sünnitustest toimub kodus, millest vaid 4% juhtudest on kohal väljaõppe saanud tervisetöötaja. Maapiirkonnas on olukord veelgi drastilisem, sest kodusünnituste osakaal moodustab ligi 90%.(Bangladesh DHS 2014) Rohkem kui pooled naised ei pääse raseduse ajal tervishoiuasutustesse, vaid 11% emadest kasutavad kuue nädala jooksul pärast sünnitust tervishoiuteenuseid (HOPE Foundation).[1]

Projekti tulemusena soovime pakkuda Bangladeshi juhtivatele organisatsioonidele andmepõhist teavet, et tuvastada ja keskenduda kõige mõjukamatele rasedusaegsetele tervisenäitajatele, aidates seeläbi parandada ravimite ning muude sobivate meetmete kättesaadavust. See on eriti oluline arvestades, et paljud rasedad ja hiljuti sünnitanud emad ei saa piisavalt tervishoiuteenuseid. Meie ülim eesmärk on aidata vähendada Bangladeshi emade suremuse määra, viies abi otse emade kodudesse.

1.2 Eesmärk

Projekti eesmärk on leida Bangladeshi emade rasedusaegsete tervisenäitajate mõju rasedusaegsele riskihinnangule. See annaks juhtivatele organisatsioonidele kätte parameetrid, mis enam tegelemist vajavad.

Soovime leida vastused järgmistele küsimustele:

1. Kuidas mõjutavad andmestikus olevad tervisetunnused riskihinnangut?
2. Millised tunnused mõjutavad riskihinnangut olulisemalt?
3. Milliste tunnuste koosmõju on riskihinnangule olulisim?

[1] https://everymothercounts.org/grants/bangladesh-a-deeper-dive/?fbclid=IwAR0u9NEdMFeHH7YgG_rdjIJ3ogMvCJangMJGfTAu3zpsbCv20aIGfKS9BP0#:~:text=In%20Bangladesh%2C%20the%20rate%20of%20maternal%20mortality%20is,from%20preventable%20causes%20related%20to%20pregnancy%20and%20childbirth

Seeläbi saaksime teada, milliste tunnuste positiivne muutus võiks muuta ema rasedusaegset riskihinnangut positiivses suunas, see tähendab kõrgest keskmisse ja keskmisest madalasse riskiklassi.

2 Andmete allikad ja kirjeldus

Andmed on eelnevalt kogutud ja pärinevad Kaggle keskkonnast, Maternal Health Risk lehelt[2]. Kasutame andmestikku tervikuna, tunnuseid ära ei jäta ning valimit koguandmetest ei koosta. Andmevigade ja duplikaatide näol jätame endale õiguse neid korrigeerida.

Andmeid on kogutud 1014 naise kohta, kus iga ühe kohta kirjeldatakse tema vanus, ülemine ja alumine vererõhk, veresuhkur, kehatemperatuur, pulss ning riskihinnang. See tähendab, et andmestikus on 1014 rida ja 7 veergu, millest 2 on ujukomaarvude, 4 täisarvude ja 1 objekti tüüpi. Puuduvaid väärtusi pole. Tunnused on kirjeldatud tabelis 1.

Tabel 1 Tunnuste kirjeldus.

Column	Dtype	Attribute	Ühik
Age	int64	Continuous	Aastad
SystolicBP	int64	Continuous	mmHg (millimeetrit elavhõbedasammast)
DiastolicBP	int64	Continuous	
BS	float64	Continuous	mmol
BodyTemp	float64	Continuous	Fahrenheit
HeartRate	int64	Continuous	Löögid minutid
RiskLevel	object	Categorical (Low, Med, High)	Kategooria (Low, Med, High)

Tabelis 2 toome välja andmete statistilised väärtused. Täpsemalt käsitleme iga tunnust järgnevates alampeatükkides ning kirjeldame omakorda ka väärtusi täpsemalt.

[2] https://www.kaggle.com/datasets/joebeachcapital/maternal-health-risk/

Tabel 2 Andmete statistiline kirjeldus.

	Vanus	Ülemine vererõhu näitaja	Alumine vererõhu näitaja	Veresuhkur	Kehatemperatuur	Pulss
	Age	**SystolicBP**	**DiastolicBP**	**BS**	**BodyTemp**	**HeartRate**
count	1014,0000	1014,0000	1014,0000	1014,0000	1014,0000	1014,0000
mean	29,8717	113,1982	76,4606	8.725986	98,6651	74,3018
std	13,4744	18,4039	13,8858	3,2935	1,3714	8,0887
min	10,0000	70,0000	49,0000	6,0000	98,0000	7,0000
max	70,0000	160,0000	100,0000	19,0000	103,0000	90,0000

2.1 Vanus

Emad jäävad vanuselt vahemikku 10 kuni 70 aastat, keskmisena ~30 aastat. Standardhälve 13.47 näitab üsna suurt varieeruvust.

2.2 Vererõhk

Ülemine vererõhk (SystolicBP) näitab arterites olevat rõhku (mmHg) siis, kui süda tõmbub kokku ja pumpab verd arteritesse. Alumine vererõhk (DiastolicBP) näitab arterites olevat rõhku (mmHg) siis, kui süda on lõõgastunud ja täitub verega. Toome välja vererõhu näitajate tähendused:

- Normaalne: <120/<80 mmHg
- Kõrgenenud: 120-129/80-84 mmHg
- Kõrge (tase 1): 130-139/85-89 mmHg
- Kõrge (tase 2): 140-179/90-119 mmHg
- Kõrge (tase 3): >180/>120 mmHg[3]

Andmestikus on kõige kõrgem ülemise vererõhu näitaja 160 mmHg, mis viitab kõrgvererõhutõve 2. tasemele. Madalaim ülemise vererõhu näitaja on 70 mmHg, mis võib viidata samuti ebanormaalselt madalale näitajale. Keskmine on 113.20 mmHg, mis viitab tervele inimesele. Standardhälve on 18.40.

Kõrgeim alumise vererõhu näitaja on 100 mmHg, mida loetakse kõrgvererõhutõve 2. tasemeks, madalaim alumise vererõhu näitaja on 49 mmHg, mis tundub esmapilgul liiga madal

[3] https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

näitaja. Keskmine alumise vererõhu näitaja on 76,1 mmHg, mis viitab tervele inimesele. Standardhälve on 13.89.

2.3 Veresuhkur

Raseduse ajal on veresuhkru normiks vahemik 5,2–6,9 mmol. Kõrgem veresuhkur võib tähendada rasedusdiabeedi riski.[4]

Andmestikus on kõrgeim näitaja 19 mmol, madalaim 6 mmol ja keskmine 8.73 mmol. Standardhälve on 3.30. Kõrgeim veresuhkru näitaja on väga palju üle normi piiride, madalaim jääb rasedusaegse normi piiridesse. Keskmine tase (8,7 mmol) ületab normi ja viitab rasedusdiabeedile. Standardhälve on antud näitaja kohta võrdlemisi suur, mis viitab suurele varieeruvusele.

2.4 Kehatemperatuur

Raseduseaegne normaalne kehatemperatuur jääb vahemikku 99°F - 100°F. Andmestikus on kõrgeim näitaja 103°F, madalaim 98°F ja keskmine 98.67°F. Standardhälve on 1.37. Seega antud andmestiku naised on keskmiselt normist veidi madalama kehatemperatuuriga.

2.5 Pulss

Raseduse ajal jääb normaalne pulss vahemikku 70-90 lööki minutis (l/m). Kõrgeima pulsiga naisel on see 90 l/m, madalaimaga 7 l/m ning keskmine väärtus 74.30 l/m. Standardhälve on 8.09. Maksimaalne ja keskmine pulss jäävad normi piiresse, madalaim 7 l/m viitab andmeveale. Eemaldasime viimased väärtused andmestikust.

2.6 Riskihinnang

Rasedusaegne riskihinnang on jaotatud kolme kategooriasse: madal, keskmine, kõrge. Risk määratakse eelkirjeldatud tunnuste põhjal. Kaggle-s ei ole infot, kuidas see määrati.

3 Meetodid ja tulemused

Töö käigus jagunes andmetega töötamine üheksasse ettappi, mida kirjeldame järgnevates alampeatükkides.

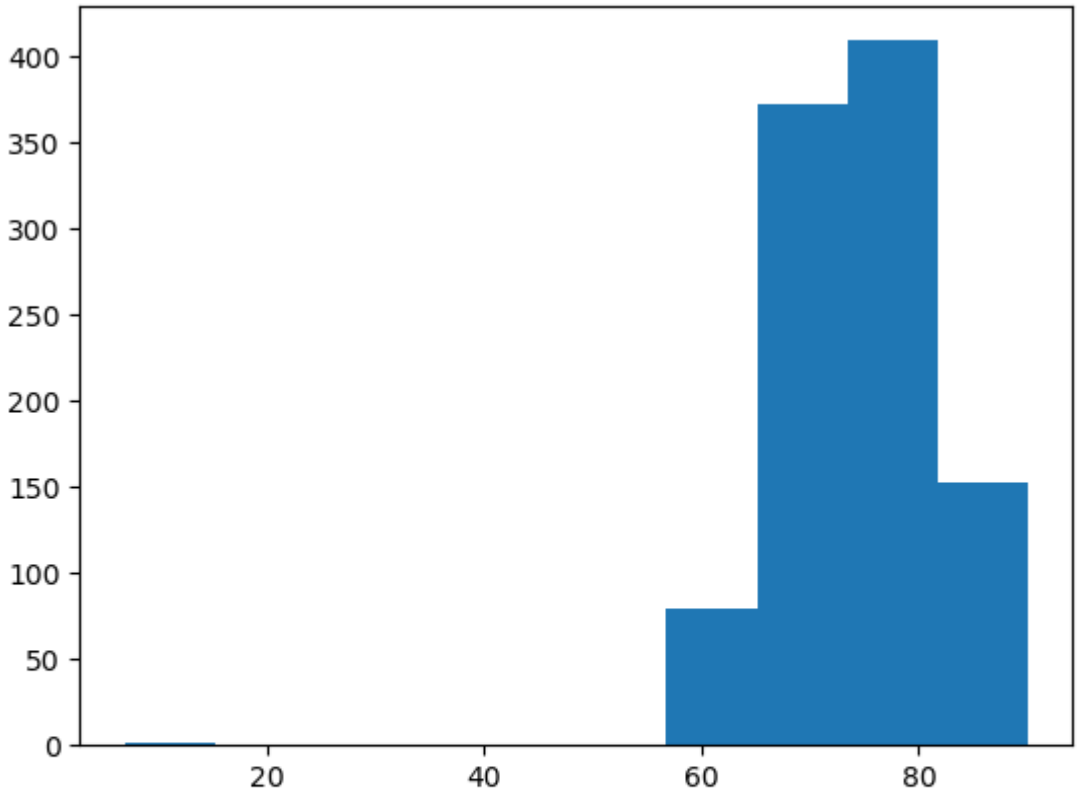
3.1 Andmete puhastamine

Puhastamise algfaasis proovisime üksikuid tunnuseid eemaldada, kuid loobusime peagi, sest see viis mudelite täpsuse oluliselt alla.

[4] https://www.kliinikum.ee/patsiendiinfo-andmebaas/rasedusdiabeet-ja-glukoosi-taluvuse-test-gtt/

Puhastamiseks analüüsisime esmalt kõigi tunnuste statistilisi näitajaid kasutades *Pandase* funktsiooni *describe*, millega tuvastasime andmetes olevad erindid-ekstreemsused. Lisaks analüüsisime tunnuseid visuaalselt karp- ja tulpdiagrammidega. Lõpuks eemaldasime vigaste tunnustega read.

Leidsime andmestikust kaks erindina tunduvat rida, kus kahel isikul on pulsi sageduseks 7 l/m, mis ei ole elujõulise inimese puhul võimalik (joonis 1). Otsustasime need kaks andmerida oma andmestikust eemaldada.



	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
499	16	120	75	7.9	98.0	7	low risk
908	16	120	75	7.9	98.0	7	low risk

Joonis 1 Uuringus osalejate pulss (l/m) ja kaks ekstreemse väärtusega rida.

Andmete süvitsi uurimisel märkasime, et esineb märkimisväärselt palju duplikaate, mida illustreerib tabel 3. Seda oli näha ka eelmises näites, kus täpselt samade näitajatega on andmestikus kahe naisterahva andmed.

Tabel 3 Näide sama väärtusega andmeridadest.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
127	55	140	95	19.0	98.0	77	high risk
137	55	140	95	19.0	98.0	77	high risk
182	55	140	95	19.0	98.0	77	high risk
278	55	140	95	19.0	98.0	77	high risk
373	55	140	95	19.0	98.0	77	high risk
554	55	140	95	19.0	98.0	77	high risk
599	55	140	95	19.0	98.0	77	high risk
603	55	140	95	19.0	98.0	77	high risk
968	55	140	95	19.0	98.0	77	high risk
1002	55	140	95	19.0	98.0	77	high risk

Duplikaate analüüsides selgus, et kõiki andmestiku veerge arvesse võttes on selliseid ridu kokku 562. See on rohkem kui pool kogu andmestikust. Eemaldades andmestikust riskihinnangu veeru (RiskLevel), on duplikaate veelgi rohkem – 598. See tähendab, et osadel juhtudel on naiste tervisenäitajad täpselt samad, aga hinnatud riskitase erinev.

Otsustasime edasist analüüsi jätkata nii duplikaatidega kui ilma, leidmaks tõesemat mudelit. Duplikaatide eemaldamiseks kasutasime *Pandase* funktsiooni *drop_duplicates*. Duplikaatide ja kahe erindi eemaldades jäävad alles 451 naise andmed. Järgnevates tabelites on esitatud duplikaatidega (tabel4) duplikaatideta (tabel 5) andmestike peamised statistilised näitajad. Kõige suurem erinevus andmestikes seisneb vererõhu näitajates, kus duplikaatidega andmestikus on vererõhu näitajad veidi kõrgemad kui duplikaatideta andmestikus.

Tabel 4 Duplikaatidega puhastatud andmete statistilised näitajad.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
count	1012.000000	1012.000000	1012.000000	1012.000000	1012.000000	1012.000000
mean	29.899209	113.184783	76.463439	8.727619	98.666403	74.434783
std	13.473560	18.419618	13.899372	3.296583	1.372421	7.521857
min	10.000000	70.000000	49.000000	6.000000	98.000000	60.000000
25%	19.000000	100.000000	65.000000	6.900000	98.000000	70.000000
50%	26.000000	120.000000	80.000000	7.500000	98.000000	76.000000
75%	39.000000	120.000000	90.000000	8.000000	98.000000	80.000000
max	70.000000	160.000000	100.000000	19.000000	103.000000	90.000000

Tabel 5 Duplikaatideta puhastatud andmete statistilised näitajad.

	Age	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate
count	451.000000	451.000000	451.000000	451.000000	451.000000	451.000000
mean	29.223947	110.532151	75.419069	8.347162	98.694013	74.097561
std	13.768594	17.886574	13.769838	2.832273	1.412086	7.530045
min	10.000000	70.000000	49.000000	6.000000	98.000000	60.000000
25%	19.000000	90.000000	65.000000	6.900000	98.000000	70.000000
50%	25.000000	120.000000	80.000000	7.500000	98.000000	76.000000
75%	35.000000	120.000000	87.000000	7.900000	98.000000	80.000000
max	70.000000	160.000000	100.000000	19.000000	103.000000	90.000000

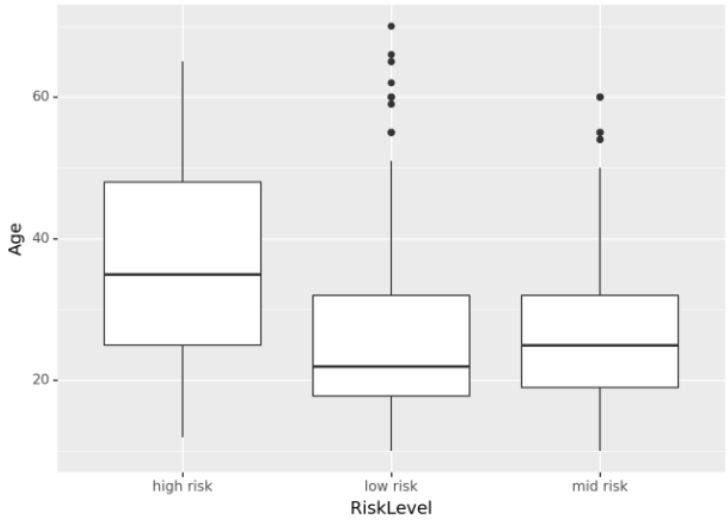
3.2 Riskihinnangu arvulise tunnuse lisamine

Kodeerimise eesmärgil lisasime andmestikku tunnuse RiskLevel_encoded, mis väljendab riskihinnangut arvulise väärtusena. Väärtustasime tunnuse järgmiselt:

- Low risk = 1
- Mid risk = 2
- High risk =3

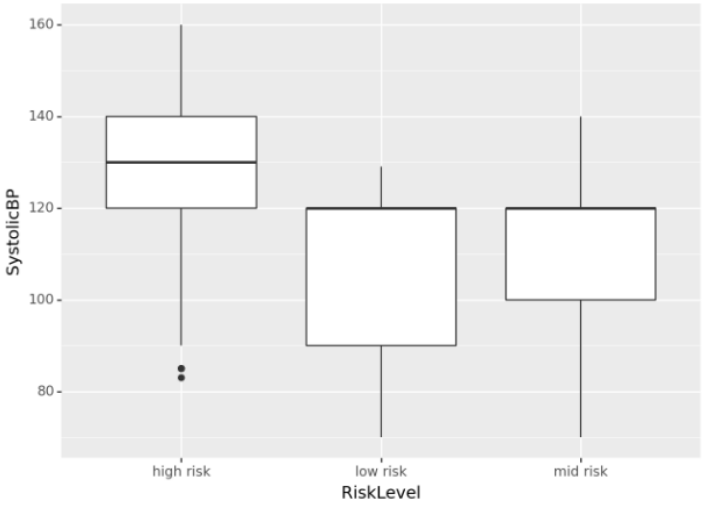
3.3 Karpdiagrammid

Esialgseks analüüsiks kasutasime karpdiagramme, mille abil võrdlesime tunnuste mõju riskihinnangule. Jooniselt 2 näeme, et vanus ja riskihinnang on võrdelises seoses.

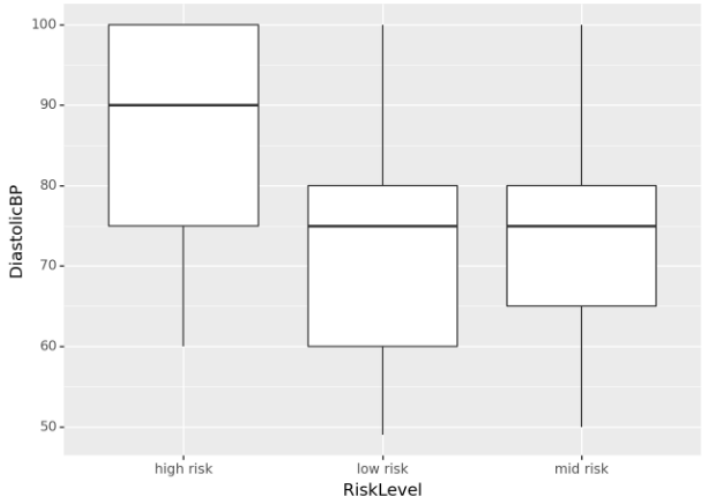


Joonis 2 Vanuse mõju riskihinnangule.

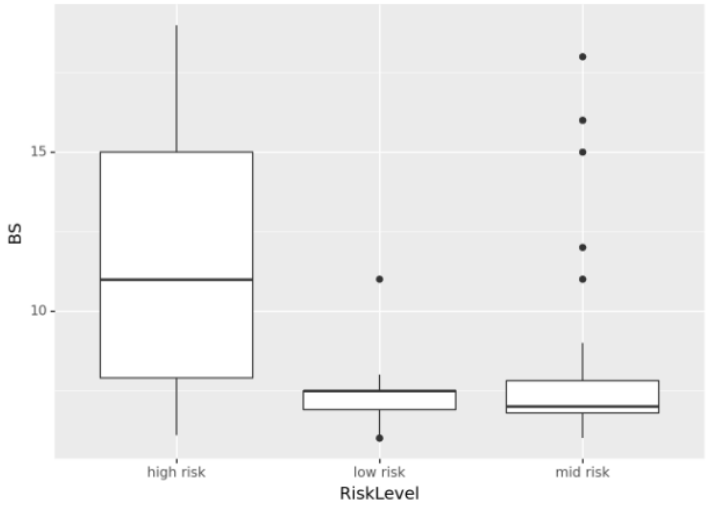
Joonistelt 3, 4 ja 5 on näha, et nii kõrge vererõhu kui kõrge veresuhkru taseme puhul on riskihinnang üldiselt samuti kõrge. Normaalse vererõhu ja veresuhkru taseme puhul ei saa lugeda otsest mõju riskihinnangule.



Joonis 3 Ülemise vererõhu mõju riskihinnangule.

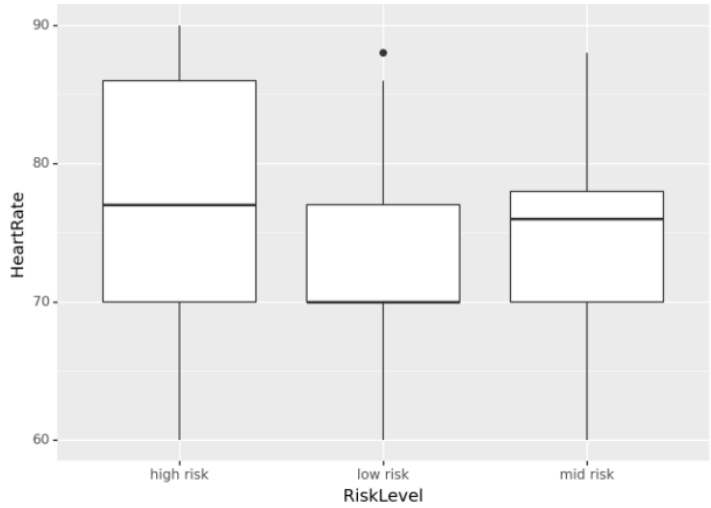


Joonis 4 Alumise vererõhu mõju riskihinnangule.



Joonis 5 Veresuhkru taseme mõju riskihinnangule.

Joonisel 6 näeme, et madalam pulss viitab pigem madalale riskitasemele. Kõrgem pulss, ehkki see on normi piires, viitab keskmisele või kõrgele riskitasemele.



Joonis 6 Pulsi mõju riskihinnangule.

Analüüsi käigus joonestasime diagramme nii duplikaatidega kui duplikaatideta andmestikest, kuid seos riskihinnangu ja tervisetunnuste vahel jäi samaks.

3.4 Kruskal-Wallise test

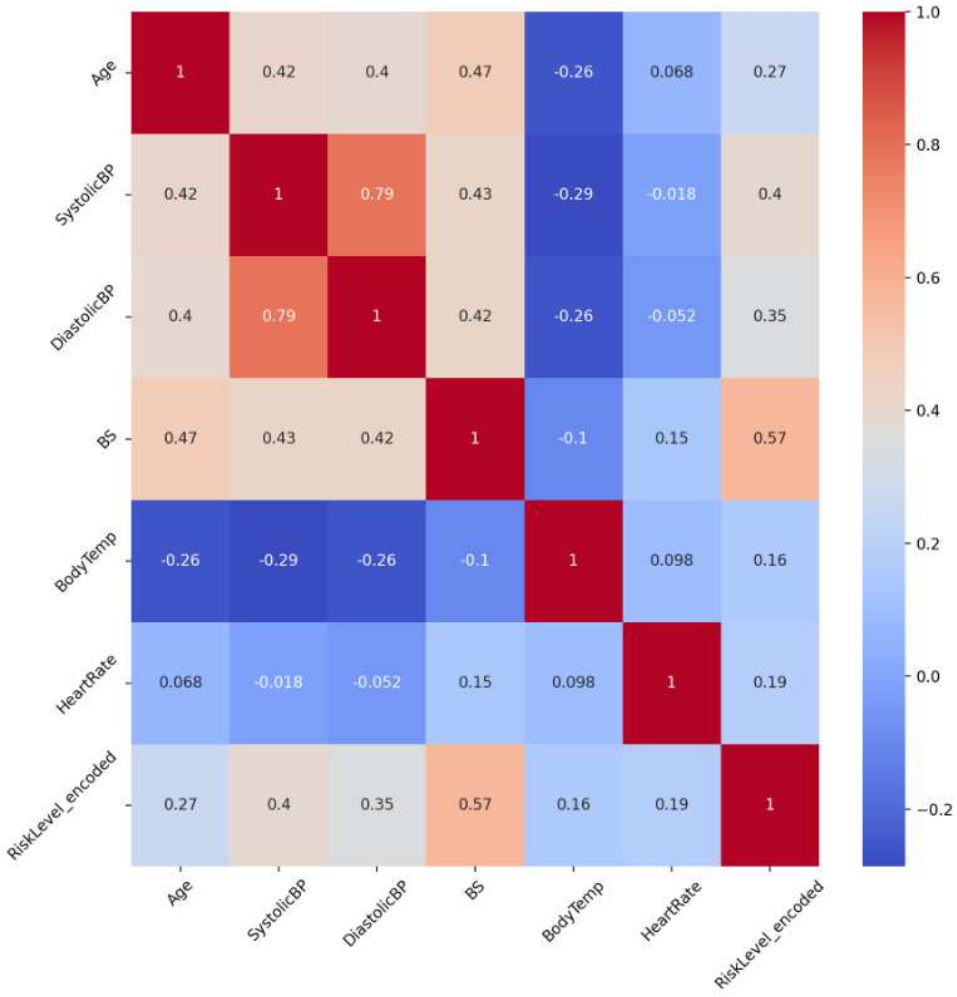
Tunnuste mõjuhinnangu olulisuse arvutamiseks kasutasime Kruskal-Wallise testi, milleks kasutasime teegi *scipy.stats* funktsiooni *kruskal*. Saime kinnituse, et tunnuste mõju riskihinnangule on statistiliselt oluline (tabel 6). Kaalusime ka ANOVA testi, kuid valisime Kruskal-Wallise, sest tunnuste värtused ei ole normaaljaotuses. Sarnase tulemuse saime nii duplikaatidega kui duplikaatideta andmestikuga.

Tabel 6 Kruskal-Wallise testi tulemused.

	Variable	P-Value	Significant
0	Age	6.556175e-22	True
1	SystolicBP	6.784520e-37	True
2	DiastolicBP	9.659143e-30	True
3	BS	9.685238e-67	True
4	BodyTemp	8.648539e-08	True
5	HeartRate	1.209704e-08	True
6	RiskLevel	2.911652e-220	True
7	RiskLevel_encoded	2.911652e-220	True

3.5 Korrelatsioonigraafik

Lisaks üksikute tervisenäitajate mõju analüüsimisele riskihinnangule soovisime näha, milline seos näitajate koosmõjul riskihinnangule, milleks koostasime korrelatsioonigraafiku (joonis 7). Graafikult näeme, et tugevas omavahelises korrelatsioonis on vaid ülemine ja alumine vererõhk. Keskmises omavahelises korrelatsioonis on vanus ja vererõhunäitajad ning veresuhkur ja vererõhunäitajad. Korrelatsioonigraafikult ilmneb kõige tugevam mõju riskihinnangule veresuhkrul.



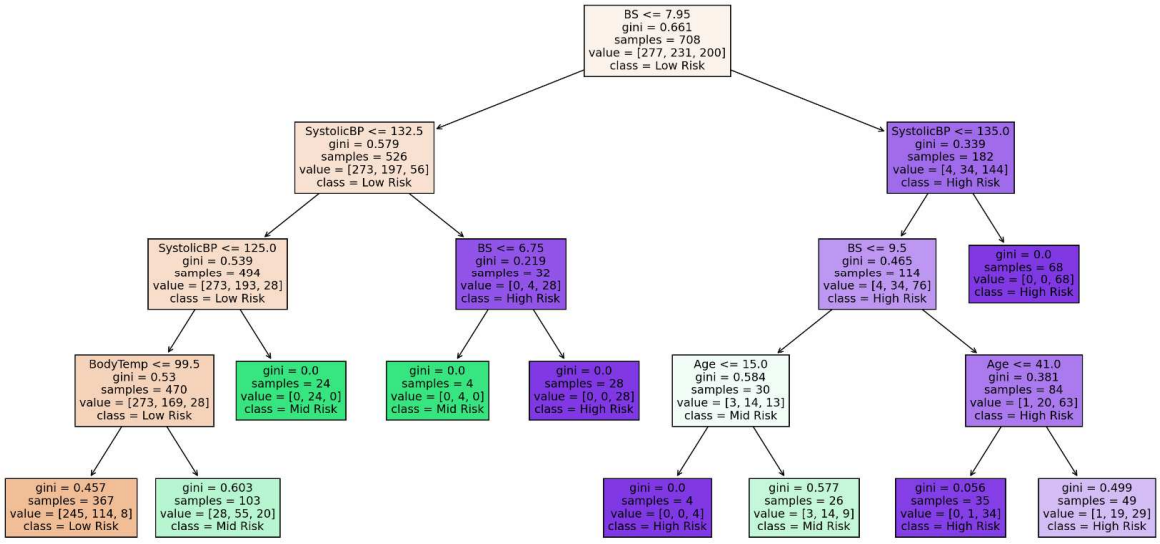
Joonis 7 Tunnuste koosmõju visualiseerimine korrelatsioonigraafikul.

3.6 Decision Tree (otsustuspuu)

Kasutasime mõjuhinnangu määramise interpreteerimiseks ka teegi *sklearn.tree* mudelit *DecisionTreeClassifier*. Saime mudeli täpsuseks 68.75%. Kõige täpsemini suudab mudel ennustada kõrget riskiklassi (80% täpsusega), lisaks on märkimisväärselt kõrge ka F1 skoor (0.82). Keskmise ja madala riskihinnangu puhul on mudeli hinnang ebatäpne.

Kuna otsustuspuu suudab kõige täpsemini ennustada kõrget riskiklassi, interpreteerime kõrge riskiklassini jõudmist neil juhtudel, kui otsustuspuu leht on tumedam ehk kõrge riskiklassiga tulemuste hulk suurem:

- Kui veresuhkur on vahemikus 6,75-7,95 mmol ja ülemine vererõhk >132,5 mmHg, määrab mudel riskihinnangu kõrgeks.
- Kui veresuhkur on >7,95 mmol ja ülemine vererõhk >135 mmHg, määrab mudel riskihinnangu kõrgeks.
- Kui veresuhkur on vahemikus 7,95-9,5 mmol, ülemine vererõhk <135 mmHg ja vanus on alla 15 aasta, määrab mudel riskihinnangu kõrgeks.
- Kui veresuhkur on >9,5 mmol, ülemine vererõhk <135 mmHg ja vanus <41 aasta , määrab mudel riskihinnangu kõrgeks.



Antud mudeli põhjal saab visuaalselt näha, milliste tulemuste muutmine võiks liigutada riskiklassi kõrgemast madalamaks:

- Oletame, et veresuhkur on <7,95 mmol ning ülemine vererõhk >132,5 mmHg. Juhul kui veresuhkur oleks <6,75 mmol, määrab mudel riskihinnanguks kõrge, vastupidisel juhul aga keskmise.
- Oletame, et veresuhkur on vahemikus 7,95-9,5 mmol ja ülemine vererõhk <135 mmHg. Sellisel juhul sõltub riskihinnang vanusest – <15 aastastel on riskiklass kõrge, teistel aga keskmine.

- Oletame, et veresuhkur on <7,95 mmol ja ülemine vererõhk <125 mmHg. Sellisel juhul määrab riskiklassi kehatemperatuur. Kehatemperatuuriga >99.5°F on naine keskmises, alla selle aga madalas riskiklassis.

3.7 K Nearest Neighbors (KNN)

Kasutasime analüüsiks *sklear.neighbors* teegi mudelit *KNeighborsClassifier*. KNN mudeliga saavutasime testandmetel täpsuse 63,7%.

3.8 Support Vector Classification (SVC)

Kasutasime analüüsiks *sklearn.svm* teegi mudelit SVC. SVC mudel saavutas testandmetel täpsuse 64,8%.

3.9 Random Forest (RF)

Kasutasime analüüsiks *sklear.ensemble* teegi mudelit *RandomForestClassifier*. Vaikimisi parameetritega saavutas mudel testandmetel täpsuse 67% (tabel 7). RF on efektiivsem madala ja kõrge riskiga juhtumite tuvastamisel. Madalam täpsus keskmise riskiga juhtumite tuvastamisel võib viidata asjaolule, et andmestik ei ole esinduslik või meil puudub oskus seda töödelda.

Tabel 7 *RandomForestClassifier* vaikimisi parameetritega.

	Klass	Täpsus	Taastuvus	F1-Score
0	High Risk	0,83	0,86	0,84
1	Low Risk	0,70	0,76	0,73
2	Mid Risk	0,33	0,27	0,30

Katsetasime mudelit erinevate parameetritega ning saavutasime optimaalseks täpsuseks 75,8% (tabel 8). Parameetreid muutsime järgmiselt:

n_estimators 50…1000, optimaalseks väärtuseks jäi 50.

max_depth 5...10, optimaalseks väärtuseks jäi 5.

max_features auto, 1, optimaalseks väärtuseks jäi 1.

min_samples_split 2…20, optimaalseks väärtuseks jäi 3.

min_samples_leaf 1...20, optimaalseks väärtuseks jäi 1.

random_state none, 1, optimaalseks väärtuseks jäi 1.

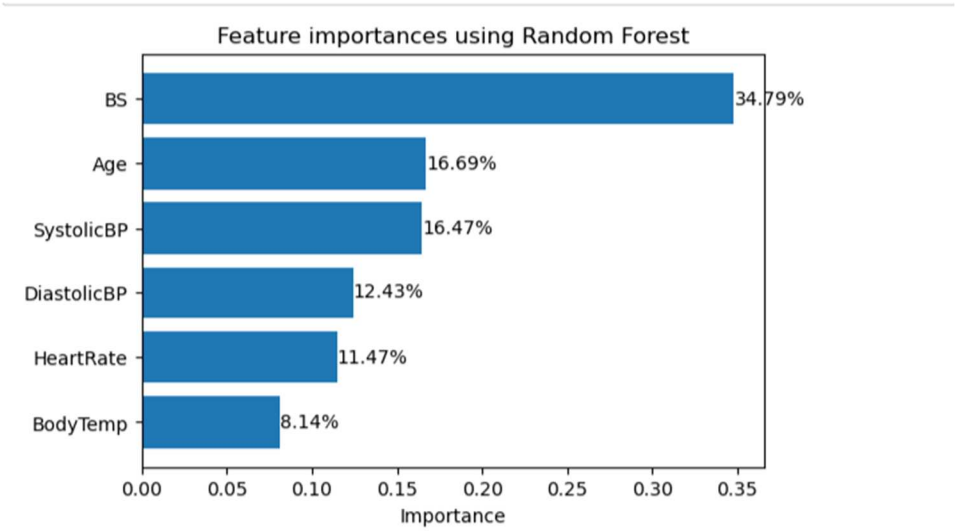
Tabel 8 *Random forest classifier* optimeeritud parameetritega.

	Klass	Täpsus	Taastuvus	F1-Score
0	High Risk	0,83	0,86	0,86
1	Low Risk	0,71	1,00	0,83
2	Mid Risk	0,80	0,18	0,30

RF on hea kõrge ja madala riskiga juhtumite tuvastamisel. Keskmise riski kategoorias on mudelil on kõrge täpsus, kuid madal taastuvus, mis viitab paljude vale-negatiivsetele tulemustele.

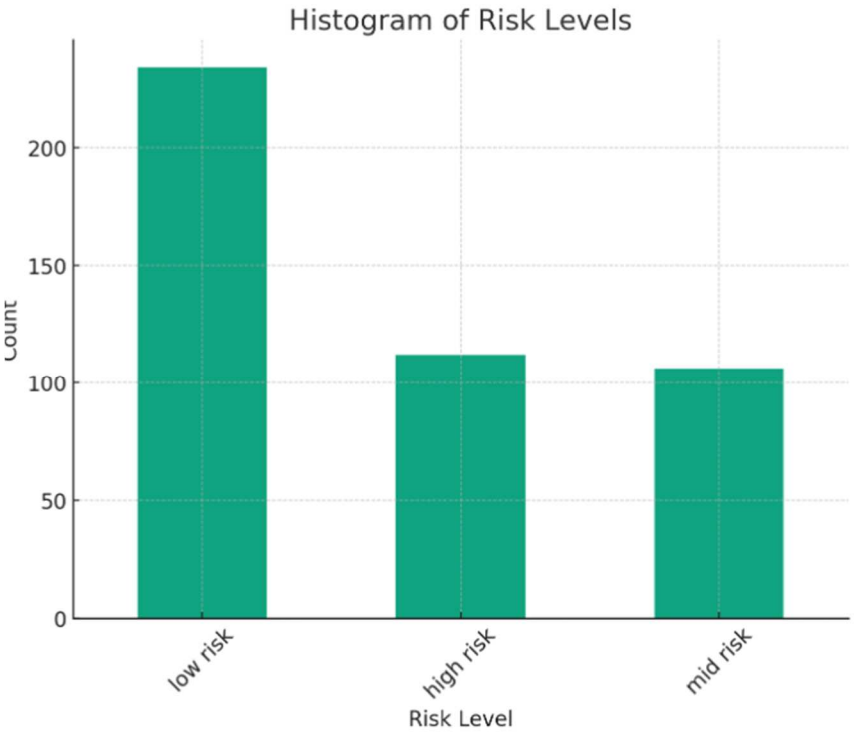
Kasutasime analüüsil kõiki omadusi, seega võib tunnuste tähtsuse analüüsimine anda teavet neist, mis on riskitaseme ennustamisel kõige mõjukamad.

Joonise 8 järgi on veresuhkur selgelt kõige suurema tähtsusega (35%) näitaja mudeli ennustuste tegemisel, seega olulisim tegur rasedate naiste terviseriski hindamisel. Veresuhkrule järgnevad vanus ja ülemine vererõhk 16% tähtsusega. Ülejäänud parameetrid omavad juba vähem tähtsust, kuid on siiski olulised, sest nende eemaldamisel väheneb mudeli täpsus drastiliselt.



Joonis 8 Tunnuste osatähtsus RF analüüsil.

Analüüsi käigus leidsime, et täpsus 75,6% on terviseriski hindamisel liiga madal, mistõttu otsustasime andmeid tasakaalustada. Puhastatud ja duplikaatideta andmete puhul näitab histogramm (joonis 9), et madala riskiga naisi on kaks korda enam kui keskmise või kõrge riskiga.



Joonis 9 Naiste jaotus riskirühmadesse puhastatud duplikaatideta andmete puhul.

Tasakaalustame andmed teegi *sklearn.utils* funktsiooniga *resample*, mis tähendab, et suurendame vähemusklassi esindavate ridade arvu. Tehes arvutuse uuesti, saame täpsuseks 77,4% (tabel 9). Kui teeme sama arvutuse algandmetega ehk nendega, mis sisaldavad duplikaate, saame täpsuseks suisa 89% (tabel 10).

Tabel 9 *resample*-ga tasakaalustatud andmete tulemus.

	Klass	Täpsus	Meenutus	F1-Score
0	High Risk	0,72	0,85	0,78
1	Low Risk	0,64	0,90	0,75
2	Mid Risk	0,67	0,29	0,40

Tabel 10 arvutustulemus algandmetega.

	Klass	Täpsus	Meenutus	F1-Score
0	High Risk	0,95	0,94	0,95
1	Low Risk	0,91	0,88	0,89
2	Mid Risk	0,80	0,84	0,82

4 Järeldused ja arutelu

Käesoleva uurimuse eesmärk oli analüüsida Bangladeshi naiste rasedusaegset tervist, keskendudes eriti riskirühmadele – madal, keskmine ja kõrge. Uurimuse käigus tehti kindlaks, kuidas erinevad tervisetunnused mõjutavad rasedusaegset terviseriski ning millised neist on olulisemad. Oluline oli ka välja selgitada, milliste tunnuste koosmõju on riskihinnangule kõige olulisem.

Analüüsist selgus, et suur osa andmestikust koosnes duplikaatidest, mis oli märkimisväärne avastus, kuna see tähendas, et keskmise ja kõrge riskiga naiste esinemissagedus oli kunstlikult tõusnud. See avastus rõhutab täpse ja puhastatud andmestiku olulisust usaldusväärsete järelduste tegemiseks. Karpdiagrammide ja Kruskal-Wallise testi abil leiti, et tervisenäitajate normist kõrvalekaldumine suurendab riskihinnangu tõenäosust. Sellest järeldame, et rasedusaegse terviseriski hindamisel on oluline arvestada mitmete tervisenäitajatega.

Kõige märgatavam oli veresuhkru taseme mõju riskihinnangule. See tulemus on kooskõlas meditsiiniliste uuringute ja teooriatega, mis näitavad, et rasedusaegne diabeet võib suurendada mitmesuguseid riske nii emale kui ka lapsele. Uurimus tõi välja ka vanuse ja vererõhu kui olulised riskitegurid, mis on samuti kooskõlas varasemate teadusuuringutega.

Huvitavaks leiuks oli see, et kuigi kehatemperatuur ja pulss olid vähem olulised kui teised tunnused, mängisid need siiski rolli riskihinnangute täpsuses. Nende eemaldamine andmestikust viis mudeli täpsuse olulisele langusele, mis rõhutab erinevate tervisenäitajate koosmõju olulisust riskihinnangute puhul.

Random Foresti mudeli abil saavutati tulemus, mis näitas 75,8% täpsust. See on oluline tulemus, kuid tuleb märkida, et mudel genereeris arvestatava hulga vale-negatiivseid tulemusi. Meditsiinilises kontekstis võib see olla suur probleem, kuna tähendab, et tõsised terviseriskid võivad jääda diagnoosimata. Antud tulemus rõhutab vajadust mudeli täiendavaks arendamiseks ja täpsustamiseks. Kindlasti aitaks mudeli täpsust tõsta ka esinduslikum valim.

Kokkuvõttes annab uurimus väärtuslikku teavet rasedusaegse terviseriski hindamise kohta. Näeme, et rasedusaegse terviseriski hinnangute tegemisel on oluline arvestada mitmeid tervisenäitajaid, eriti veresuhkru taset, vanust ja vererõhku. Samuti näeme tulemustest kui oluline on andmete puhastamise põhjalikkus enne analüüsi, et tagada usaldusväärsete järelduste tegemine. Loodetavasti aitavad need tulemused kaasa paremate ennetusstrateegiate väljatöötamisele ja rasedate naiste tervise parandamisele Bangladeshis ning laiemaltki.