UNIVERSITATEA DE MEDICINĂ,
FARMACIE, ȘTIINȚE ȘI TEHNOLOGIE
„GEORGE EMIL PALADE"
DIN TÂRGU MUREȘ

FACULTATEA DE INGINERIE ŞI TEHNOLOGIA INFORMAȚIEI

PROGRAMUL DE STUDII UNIVERSITARE DE MASTERAT

INTELIGENȚĂ ARTIFICIALĂ, ANUL I

Proiect individual de analiză statistică

Disciplina: Probabilităţi şi statistică în sisteme medicale

Profesor:
Manuela Rozalia Gabor

Student:
Finucă Andrei-Cosmin

Târgu Mureş,
Noiembrie 2022

CUPRINS

1. Descrierea bazei de date

Baza de date utilizată în prezenta lucrare este parte integrată a cercetării publicate în articolului *„Predicting seminal quality with artificial intelligence methods”* [1], și a fost obținută de la University of California, Irvine „Machine Learning Repository” [2].

Eșantionul studiului este format din 100 de studenții voluntari ai Universității Alicante, cu vârste cuprinse între 18 și 36 de ani, ce nu prezintă alterării reproductive cunoscute precedent (ex. varicocel).

Setul de date este format din 100 de observații statistice, cu 10 atribute descrise în *Tabel 1 Descrierea atributelor setului de date*.

Tabel 1 Descrierea atributelor setului de date

Atribut	Valoare (codificare) / Interval	Tip de variabilă	Unitate de măsură
Sezon recoltare	Winter (1), Spring (2), Summer (3), Fall (4)	Calitativă	-
Vârsta la recoltare	interval de valori întregi	Cantitativă	ani
Boli ale copilăriei	No (0), Yes (1)	Calitativă	-
Accidente sau traume serioase	No (0), Yes (1)	Calitativă	-
Intervenții chirurghicale	No (0), Yes (1)	Calitativă	-
Febre ridicate în ultimul an	No (1), More than 3 months ago (2), Less than 3 months ago (3)	Calitativă	-
Consum de alcool	hardly ever or never (1), once a week (2), several times a week (3), every day (4), several times a day (5)	Calitativă	-
Fumat	Never (1), Ocassional (2), Daily (3)	Calitativă	-
Număr de ore de stat pe scaun	interval de valori întregi	Cantitativă	ore
Rezultat analiză material seminal	N-Normal (0), O-Altered (1)	Calitativă	-

Variable	Name	Type	Width	Decimal	Label	Value Labels	Missing Values	Columns	Align	Measure	Role
1	season	Numeric	2	0	Sezon recol	{1, Winter}...	None	8	Right	Nominal	Input
2	age	Numeric	3	0	Varsta la re	None	None	8	Center	Scale	Input
3	diseases	Numeric	1	0	Boli ale cop	{0, No}...	None	8	Center	Nominal	Input
4	accidents	Numeric	1	0	Accidente	{0, No}...	None	8	Center	Nominal	Input
5	surgery	Numeric	1	0	Interventii c	{0, No}...	None	8	Center	Nominal	Input
6	fevers	Numeric	1	0	Febra ridica	{1, no}...	None	17	Right	Nominal	Input
7	alcohol	Numeric	2	0	Consum de	{1, hardly or never}...	None	15	Right	Nominal	Input
8	smoking	Numeric	1	0	Fumator	{1, never}...	None	14	Right	Nominal	Input
9	sitting	Numeric	2	0	Ore de inac	None	None	8	Center	Scale	Input
10	diagnosis	Numeric	2	0	Rezultat	{0, Normal}...	None	8	Center	Nominal	Input

Fig. 1 Structura variabilelor în PSPP

2. Statistica descriptiva

2.1. Indicatorii tendinței centrale

Variabilele calitative sunt reprezentate prin tabele de frecvență și grafice de structură:

Sezon recoltare

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Winter	28	28,0%	28,0%	28,0%
	Spring	37	37,0%	37,0%	65,0%
	Summer	4	4,0%	4,0%	69,0%
	Fall	31	31,0%	31,0%	100,0%
Total		100	100,0%		

Fig. 2 Tabel de frecvență "Sezon recoltare"

Pentru cei 100 de subiecții, conform *Fig. 2 Tabel de frecvență "Sezon recoltare"*, mostrele de material seminal pentru analizare au fost recoltate după cum urmează:

- pentru 28, respective 28%, iarna
- pentru 37, respectiv 37%, primăvara
- pentru 4, respectiv 4%, vara
- iar pentru 31, respectiv 31%, toamna

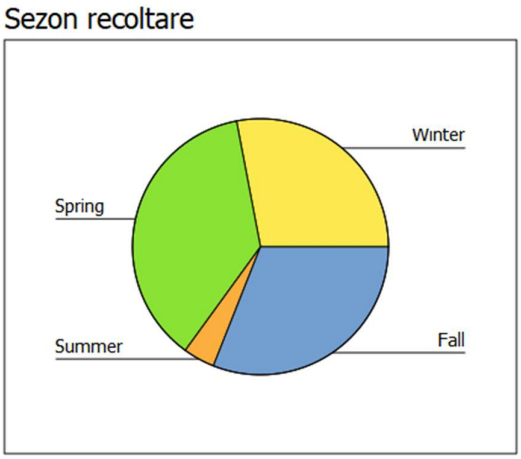


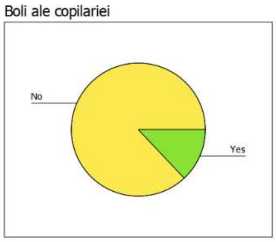
Fig. 3 Grafic "Sezon recoltare"

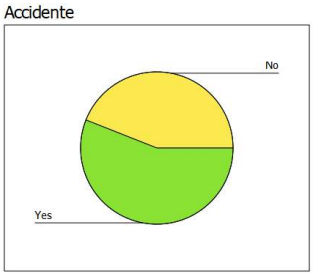
Se observă că din totalul observațiilor statistice, ponderea cea mai scăzută o reprezintă subiecții ale căror mostre au fost prelevate vara, celelalte categorii reprezentând procente relativ asemănătoare din totalul observațiilor.

Alte exemple de grafice de structură și tabele de frecvență pentru variabilele bazei de date:

Boli ale copilariei

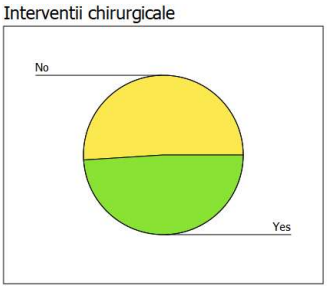
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	87	87,0%	87,0%	87,0%
	Yes	13	13,0%	13,0%	100,0%
Total		100	100,0%		





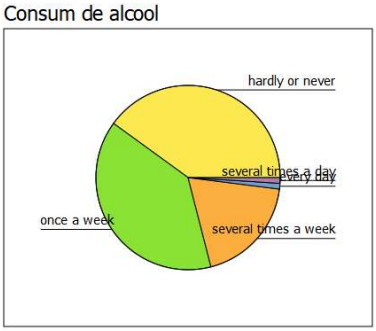
Accidente

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	44	44,0%	44,0%	44,0%
	Yes	56	56,0%	56,0%	100,0%
Total		100	100,0%		



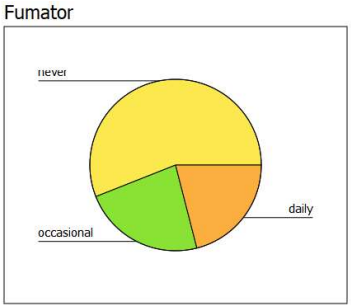
Interventii chirurgicale

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	51	51,0%	51,0%	51,0%
	Yes	49	49,0%	49,0%	100,0%
Total		100	100,0%		



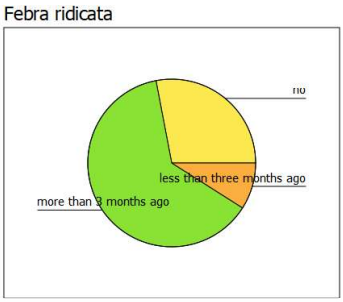
Consum de alcool

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	hardly or never	40	40,0%	40,0%	40,0%
	once a week	39	39,0%	39,0%	79,0%
	several times a week	19	19,0%	19,0%	98,0%
	every day	1	1,0%	1,0%	99,0%
	several times a day	1	1,0%	1,0%	100,0%
Total		100	100,0%		



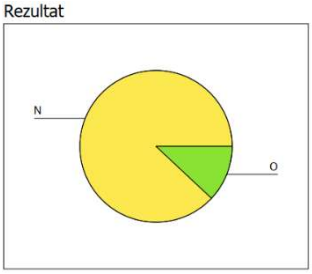
Fumator

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	never	56	56,0%	56,0%	56,0%
	occasional	23	23,0%	23,0%	79,0%
	daily	21	21,0%	21,0%	100,0%
Total		100	100,0%		



Febra ridicata

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	28	28,0%	28,0%	28,0%
	more than 3 months ago	63	63,0%	63,0%	91,0%
	less than three months ago	9	9,0%	9,0%	100,0%
Total		100	100,0%		



Rezultat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	N	88	88,0%	88,0%	88,0%
	O	12	12,0%	12,0%	100,0%
Total		100	100,0%		

Pentru variabilele cantitative s-au calculat indicatorii tendinței centrale: media, mediana și modul.

Statistics

		Varsta la recoltare	Ore de inactivitate fizica
N	Valid	100	100
	Missing	0	0
Mean		24,05	6,49
Median		24,00	6,00
Mode		24	.

Fig. 6 Indicatorii tendinței centrale

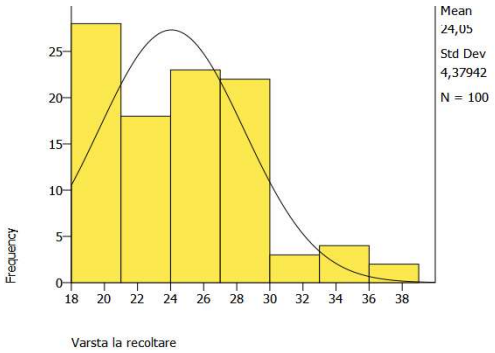


Fig. 5 Histogramă vârstă

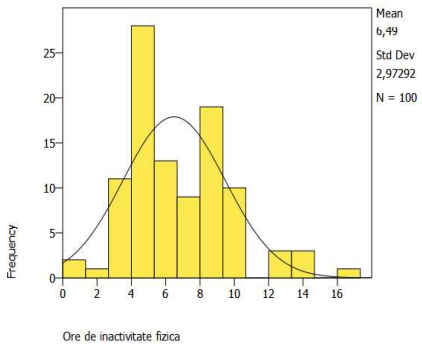


Fig. 4 Histogramă ore de stat pe scaun

Pentru *„ore de inactivitate fizică”,* sau numărul de ore petrecute pe scaun, media este de 6 ore și 30 de minute, 6,49 ore, iar mediana este de 6 ore, rezultând că jumătate, 50%, dintre subiecți petrec mai puțin de 6 ore stând pe scaun, iar 50% stau peste 6 ore pe scaun.

Modul pentru „ore de inactivitate fizică”, este reprezentant prin *„ . ”* întrucât această variabilă are o distribuție bimodală, altfel spus, există două valori cu aceeași frecvență, frecvența maximă de 17%, valorile 4 și 8 conform *Fig. 7 Tabel de frecvență "ore de inactivitate"*.

Ore de inactivitate fizica

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	2	2,0%	2,0%	2,0%
	2	1	1,0%	1,0%	3,0%
	3	11	11,0%	11,0%	14,0%
	4	17	17,0%	17,0%	31,0%
	5	11	11,0%	11,0%	42,0%
	6	13	13,0%	13,0%	55,0%
	7	9	9,0%	9,0%	64,0%
	8	17	17,0%	17,0%	81,0%
	9	2	2,0%	2,0%	83,0%
	10	10	10,0%	10,0%	93,0%
	12	3	3,0%	3,0%	96,0%
	14	3	3,0%	3,0%	99,0%
	16	1	1,0%	1,0%	100,0%
Total		100	100,0%		

Fig. 7 Tabel de frecvență "ore de inactivitate"

În cazul variabilei *„vârsta la recoltare”* modul este 24, și conform histogramei și curbei Gauss de normalitate, se observă o curbă normală.

2.2. Indicatorii împrăștierii

Descriptive Statistics

	N	Mean	Std Dev	Kurtosis	S.E. Kurt	Skewness	S.E. Skew	Range	Minimum	Maximum
Varsta la recoltare	100	24,05	4,38	-,02	,48	,67	,24	18,00	18	36
Ore de inactivitate fizica	100	6,49	2,97	,58	,48	,77	,24	15,00	1	16
Valid N (listwise)	100									
Missing N (listwise)	0									

Fig. 8 Indicatorii împrăștierii

Din *Fig. 8 Indicatorii împrăștierii* putem deduce că pentru variabila *„ore de inactivitate fizica”*:

- abaterea standard este de ~ 3 ore, respectiv fiecare dintre cei 100 de subiecți se abate în medie de la numărul mediu de ore de stat pe scaun cu ± 3 ore,
- valoarea cea mai mică este de 1 oră,
- valoarea ce mai mare este de 16 ore,
- amplitudinea (16 - 1 ore) = 15 ore
- aplatizarea este (+0,58) valoarea pozitivă indicând o distribuție cu un vârf mai înalt decât distribuția normală
- asimetria este (+0,77) valoarea pozitivă indicând o „coadă” a distribuției de frecvență în
- zona valorilor mai mari decât valoarea medie.

Scrierea pentru aceste date se face astfel: media ± std. dev. (min ÷ max), respectiv

$$6{,}49 \pm 3\ (1 \div 16)$$

3. Corelații

Correlations

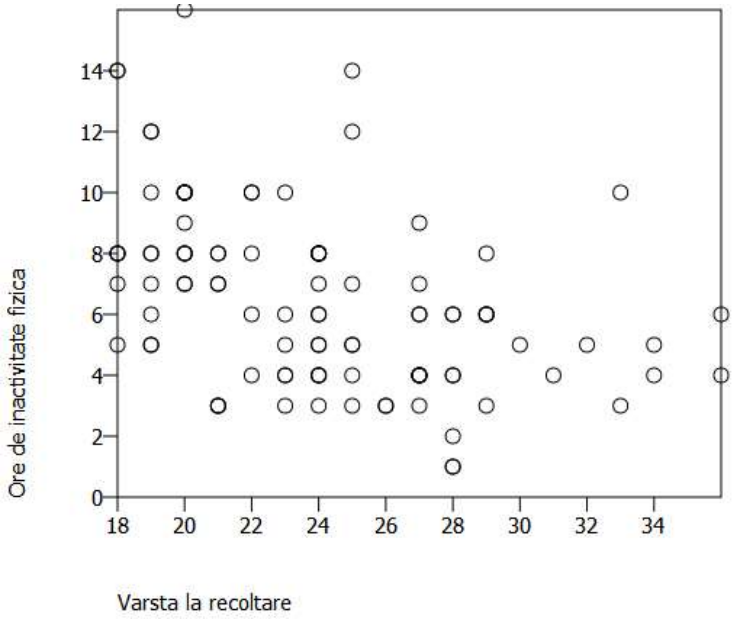
		Varsta la recoltare	Ore de inactivitate fizica
Varsta la recoltare	Pearson Correlation	1,000	-,448[a]
	Sig. (2-tailed)		,000
	N	100	100
Ore de inactivitate fizica	Pearson Correlation	-,448[a]	1,000
	Sig. (2-tailed)	,000	
	N	100	100

a. Significant at .05 level

Fig. 9 Tabel de corelații

Între variabilele „vârsta la recoltare” și „ore de inactivitate” există o corelație INVERSĂ (semn algebric negativ), de intensitate MEDIE (|0,40 < r < 0,59|), semnificativă pentru toți subiecții studiului.

Scatter plotul celor două variabile este:



4. Odds ratio & Risk ratio

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Interventii chirurgicale (No / Yes)	1,52	1,52	1,52
For cohort Accidente = No	,81	,81	,81
For cohort Accidente = Yes	1,18	1,18	1,18
N of Valid Cases	100,00		

Fig. 10 Odds ratio pentru „accidente sau traume" x „intervenții chirurgicale"

Întrucât valoarea odds ratio este >1, putem concluziona că există o asociere pozitivă între prezența factorului de risc „accidente sau traume" și apariția efectului „intervenții chirurgicale". Însă, întrucât p-value (0.302) > 0.05, această asociere este nesemnificativă statistic.

Chi-Square Tests

	Value	df	Asymptotic Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-tailed)
Pearson Chi-Square	1,06	1	,302		
Likelihood Ratio	1,07	1	,302		
Fisher's Exact Test				,321	,203
Continuity Correction	,69	1	,406		
Linear-by-Linear Association	1,05	1	,305		
N of Valid Cases	100				

Fig. 11 Semnificație statistică

5. Chi Square bivariat

Întrucât p-value (0.019) pentru testul Chi Square *„Febra ridicata x Intervenții chirurgicale”* este < 0.05, rezultatele acestuia vor fi interpretate.

Chi-Square Tests

	Value	df	Asymptotic Sig. (2-tailed)
Pearson Chi-Square	7,90	2	,019
Likelihood Ratio	8,09	2	,018
Linear-by-Linear Association	5,31	1	,021
N of Valid Cases	100		

Fig. 12 Test Chi Square „Febra ridicata x Intervenții chirurgicale”

Ipoteza nulă, H_0, este: Nu există diferențe semnificative statistic în funcție de modificarea *„febra ridicata”* referitoare la modificarea *„ intervenții chirurgicale”*.

Întrucât valoarea calculată pentru Chi Square este mai mare decât valoarea teoretică pentru 2 grade de libertate, se respinge ipoteză nula și se acceptă ipoteza alternativă, și anume că există diferențe semnificative statistic în funcție de modificarea *„febra ridicata”* referitoare la modificarea *„ intervenții chirurgicale”*.

Alte exemple de teste Chi Square, nesemnificative statistic:

Chi-Square Tests

	Value	df	Asymptotic Sig. (2-tailed)
Pearson Chi-Square	,34	2	,845
Likelihood Ratio	,34	2	,845
Linear-by-Linear Association	,28	1	,595
N of Valid Cases	100		

Fig. 13 Test Chi Square „Fumator x Intervenții chirurgicale”

Chi-Square Tests

	Value	df	Asymptotic Sig. (2-tailed)	Exact Sig. (2-tailed)	Exact Sig. (1-tailed)
Pearson Chi-Square	1,06	1	,302		
Likelihood Ratio	1,07	1	,302		
Fisher's Exact Test				,321	,203
Continuity Correction	,69	1	,406		
Linear-by-Linear Association	1,05	1	,305		
N of Valid Cases	100				

Fig. 14 Test Chi Square „Accidente x Intervenții chirurgicale”

6. Regresie liniară simplă

ANOVA (Interventii chirurgicale)

	Sum of Squares	df	Mean Square	F	Sig.
Regression	5,21	8	,65	2,99	,005
Residual	19,78	91	,22		
Total	24,99	99			

Fig. 15 Anova

Model Summary (Interventii chirurgicale)

R	R Square	Adjusted R Square	Std. Error of the Estimate
,46	,21	,14	,47

Fig. 16 Sumar model regresie

Conform *Fig. 16 Sumar model regresie,* modelul este semnificativ statistic (p-value = 0.005), și este valid pentru 21% dintre subiecții studiului.

Coefficients (Interventii chirurgicale)

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	1,27	,44	,00	2,90	,005
Sezon recoltare	,05	,04	,12	1,24	,219
Varsta la recoltare	-,03	,01	-,26	-2,33	,022
Boli ale copilariei	-,28	,15	-,19	-1,91	,060
Accidente	,05	,10	,05	,54	,593
Febra ridicata	-,25	,08	-,29	-3,01	,003
Consum de alcool	,00	,06	,01	,07	,945
Fumator	,05	,06	,08	,87	,389
Ore de inactivitate fizica	,03	,02	,16	1,54	,127

Fig. 17 Coeficienți regresie

Observând valoarea p-value ,din *Fig. 17 Coeficienți regresie*, putem deduce că următoarele variabile au o contribuție semnificativă în modelul de regresie:

- Vârsta la recoltare
- Febra ridicata
- Boli ale copilăriei, dacă acceptăm un nivel de 0.1 pentru semnificația statistică

Ecuația modelului de regresie, pentru p-value < 0.1, este:

$$intervenţii\ chirurgicale = 1.27 - 0.03\ varsta - 0.25\ febra\ r. - 0.28\ boli\ a.c.$$

7. ANOVA

```
ONEWAY /VARIABLES= age sitting BY diagnosis
       /STATISTICS=DESCRIPTIVES .
```

Descriptives

	Rezultat	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Varsta la recoltare	Normal	88	23,86	4,44	,47	22,92	24,81	18,00	36,00
	Altered	12	25,42	3,75	1,08	23,03	27,80	18,00	34,00
	Total	100	24,05	4,38	,44	23,18	24,92	18,00	36,00
Ore de inactivitate fizica	Normal	88	6,47	2,98	,32	5,83	7,10	1,00	16,00
	Altered	12	6,67	3,06	,88	4,73	8,61	1,00	14,00
	Total	100	6,49	2,97	,30	5,90	7,08	1,00	16,00

Fig. 19 Tabel Anova Descriptives

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Varsta la recoltare	Between Groups	25,47	1	25,47	1,33	,251
	Within Groups	1873,28	98	19,12		
	Total	1898,75	99			
Ore de inactivitate fizica	Between Groups	,43	1	,43	,05	,828
	Within Groups	874,56	98	8,92		
	Total	874,99	99			

Fig. 18 Tabel Anova

În *Fig. 19 Tabel Anova Descriptives* sunt prezentate informații referitoare la indicatorii tendinței centrale și a împrăștierii pentru fiecare grup aferent factorului de grupare „*rezultat*”.

Conform *Fig. 18 Tabel Anova*, variabilele „*vârsta*” și „*ore de inactivitate*” au medii cu diferențe nesemnificative statistic, pentru fiecare dintre grupurile considerate.

8. Testul t student

Group Statistics

	Group	N	Mean	Std. Deviation	S.E. Mean
Varsta la recoltare	Normal	88	23,86	4,44	,47
	Altered	12	25,42	3,75	1,08
Ore de inactivitate fizica	Normal	88	6,47	2,98	,32
	Altered	12	6,67	3,06	,88

Independent Samples Test

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Varsta la recoltare	Equal variances assumed	1,90	,171	-1,15	98,00	,251	-1,55	1,35	-4,22	1,12
	Equal variances not assumed			-1,31	15,54	,208	-1,55	1,18	-4,07	,96
Ore de inactivitate fizica	Equal variances assumed	,62	,435	-,22	98,00	,828	-,20	,92	-2,03	1,62
	Equal variances not assumed			-,21	14,01	,833	-,20	,94	-2,21	1,81

Atât prin probabilitatea *p* cât și prin compararea mediilor, concluzia testului t Student este că nu se observă diferențe semnificative între medii, sau altfel spus nu există diferențe între media vârstei la recoltare, sau a numărului de ore de stat pe scaun, între grupul cu rezultat modificat și cel cu valori normale, prin urmare nu influențează modificarea valorilor analizelor (rezultatul).

9. Test neparametric

```
NPAR TEST
        /KRUSKAL-WALLIS =  season accidents surgery fevers alcohol smoking diagnosis BY diseases (0, 1)
        .
```

Ranks

		N	Mean Rank
Sezon recoltare	No	87	48,71
	Yes	13	62,46
	Total	100	
Accidente	No	87	48,94
	Yes	13	60,96
	Total	100	
Interventii chirurgicale	No	87	51,86
	Yes	13	41,38
	Total	100	
Febra ridicata	No	87	49,71
	Yes	13	55,77
	Total	100	
Consum de alcool	No	87	49,90
	Yes	13	54,54
	Total	100	
Fumator	No	87	51,26
	Yes	13	45,38
	Total	100	
Rezultat	No	87	50,25
	Yes	13	52,19
	Total	100	

Fig. 21 Ranks

Test Statistics

	Sezon recoltare	Accidente	Interventii chirurgicale	Febra ridicata	Consum de alcool	Fumator	Rezultat
Chi-Square	2,83	2,63	1,97	,68	,33	,58	,16
df	1	1	1	1	1	1	1
Asymp. Sig.	,093	,105	,161	,410	,564	,447	,689

Fig. 20 Test Kruskal-Wallis

Din *Fig. 20 Test Kruskal-Wallis* rezultă că dacă considerăm subiecții studiului ca provenind din două eșantioane independente, grupate în funcție de prezența bolilor specifice în perioada copilăriei, pentru nici una dintre variabilele analizate nu există diferențe semnificative statistic.

10. Concluzii

12. Bibliografie

[1] D. Gil, J. L. Girela, J. D. Juan, M. J. Gomez-Torres și M. Johnsson, „Predicting seminal quality with artificial intelligence methods," *Expert Systems with Applications,* vol. 39, pp. 12564-12573, 2012.

[2] D. Dua și C. Graff, „UCI Machine Learning Repository: Fertility Data Set," University of California, Irvine, School of Information and Computer Sciences, 2017. [Interactiv]. Available: https://archive.ics.uci.edu/ml/datasets/Fertility. [Accesat 10 2022].