Politechnika Wrocławska
Wydział Informatyki i Telekomunikacji

Kierunek: **Informatyka Techniczna (IMT)**

Przedmiot: **Statystyczna Analiza Danych Medycznych (Projekt)**

Statystyczna Analiza Danych Medycznych – Dokumentacja Projektu

Jakub Krupiński

Prowadzący:

prof. dr hab. inż. Robert Burduk

WROCŁAW 2025

SPIS TREŚCI

1. STATYSTYKA OPISOWA

Projekt został przeprowadzony na podstawie zbioru danych *Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)*. Składa się on z 2509 rekordów i 21 kolumn. Zbiór danych został zmodyfikowany w celu spełnienia wymagań projektu. Pierwszym krokiem było wygenerowanie syntetycznych rekordów pomiarów guzów nowotworowych w różnych momentach czasu (pierwsze badanie, 6 miesięcy i 12 miesięcy), co zostało udokumentowane w nowej syntetycznej kolumnie *timepoint*. Został on następnie zmniejszony w taki sposób, aby zapewnić równoliczność elementów w grupach, a następnie iteracyjnie wprowadzono szum z rozkładu normalnego do kolumny *age_at_diagnosis*, aby upewnić się, że wartości skalowane są do wspólnej wariancji (dzięki czemu spełniają założenia testu *Levene'a* oraz są z rozkładu normalnego, spełniając tym samym założenia testu *Shapiro-Wilka*. Modyfikacje te zostały wprowadzone świadomie, jako że celem projektu jest zaprezentowanie wykorzystania wybranych metod statystycznej analizy danych – rzetelność przeprowadzonych badań nie jest w tym wypadku istotna.

Otrzymany zbiór danych składa się z **2820** rekordów i **22** kolumn.

Tabela 1.1. Quantitative Statistics

Column name	Type	Count	Mean	Std	Min	25%	50%	75%	Max	Skewness	Kurtosis
age_at_diagnosis	Quant	874	60.26	17.13	11.42	48.54	60.28	72.30	121.27	0.02	−0.06
overall_survival_months	Quant	874	127.96	76.39	1.23	63.75	122.12	191.82	351.00	0.27	−0.86
tumor_size	Quant	868	25.82	15.98	0.95	16.43	21.26	30.00	139.95	2.61	11.17
nottingham_prognostic_index	Quant	874	3.94	1.18	1.00	3.04	4.04	5.04	6.30	−0.11	−0.33
relapse_free_status_months	Quant	874	114.78	77.35	0.70	46.13	104.03	174.75	351.00	0.43	−0.72
tumor_stage	Quant	669	1.71	0.65	0.00	1.00	2.00	2.00	4.00	0.25	0.23
lymph_nodes_examined_positive	Quant	835	1.85	3.72	0.00	0.00	0.00	2.00	31.00	3.33	13.39
cohort	Quant	874	2.56	1.24	1.00	1.00	3.00	3.00	5.00	0.18	−0.90
mutation_count	Quant	817	5.48	3.34	1.00	3.00	5.00	7.00	24.00	1.38	3.48

Tabela 1.2. Categorical Statistics

Column name	Type	Count	Unique	Top	Count of most frequent
patient_id	Categorical	874	694	MB-0226	4
overall_survival_status	Categorical	874	2	0:LIVING	437
pam50_claudin_low_subtype	Categorical	874	7	LumA	290
er_status	Categorical	874	2	Positive	643
pr_status	Categorical	874	2	Positive	471
her2_status	Categorical	874	2	Negative	753
patients_vital_status	Categorical	873	3	Living	437
relapse_free_status	Categorical	873	2	0:Not Recurred	544
cancer_type_detailed	Categorical	874	7	Breast Invasive Ductal Carcinoma	662
cellularity	Categorical	874	3	Moderate	292
type_of_breast_surgery	Categorical	860	2	MASTECTOMY	517
inferred_menopausal_state	Categorical	874	2	Post	656
timepoint	Categorical	341	3	baseline	117

2. TESTY STATYSTYCZNE

2.1. TESTY PARAMETRYCZNE

W celu sprawdzenia, czy zbiór danych spełnia wymagania do przeprowadzenia testów parametrycznych, przeprowadzono testy: *Shapiro-Wilka*, *Levene'a* oraz *Chi-kwadrat*.

2.1.1. Test t-Studenta

Przed przeprowadzeniem testu t-Studenta przeprowadzono testy w celu sprawdzenia zgodności z założeniami.

Tabela 2.1. Shapiro-Wilk test results for *age_at_diagnosis* by *overall_survival_status*

Group	p-value	Normal distribution?
1: DECEASED	0.5564	Yes
0: Living	0.0606	Yes

Test *Levene'a* wykazał, że wariancje są jednorodne ($p = 0.8803$). Test *Chi-kwadrat* wykazał, że badane grupy są idealnie zrównoważone ($p = 1.0000$).

Celem przeprowadzenia testu t-Studenta o przyjętym poziomie istotności *0.05* było wykazanie, że istnieje statystyczna różnica w średnim wieku przy diagnozie pomiędzy grupami pacjentów, którzy przeżyli bądź zmarli. Przyjęta hipoteza zerowa H_0 mówi, że: *średni wiek przy diagnozie jest taki sam w obu badanych grupach*, a hipoteza alternatywna H_1 mówi, że: *średni wiek przy diagnozie różni się między grupami*.

Tabela 2.2. t-Student test results for *age_at_diagnosis* by *overall_survival_status*

t-statistic	p-value	Conclusion
-4.3002	$1.899e-05$	Significant difference

Pacjenci którzy zmarli byli starsi w momencie diagnozy niż ci, którzy przeżyli, a różnica jest wysoce istotna ($p = 1.899e-05$). Ponieważ $p < 0.05$, odrzucono H_0 i przyjęto H_1.

2.1.2. Test ANOVA

Przed przeprowadzeniem testu ANOVA przeprowadzono testy w celu sprawdzenia zgodności z założeniami.

Tabela 2.3. Shapiro-Wilk test results for *age_at_diagnosis* by *cellularity*

Group	p-value	Normal distribution?
High	0.0.4242	Yes
Moderate	0.6631	Yes
Low	0.0640	Yes

Test *Levene'a* wykazał, że wariancje są jednorodne ($p = 0.7254$).

Celem przeprowadzenia testu ANOVA o przyjętym poziomie istotności *0.05* była analiza różnic w komórkowości komórek nowotworowych, czyli odsetka komórek nowotworowych w próbce w stosunku do komórek nienowotworowych (np. komórek układu odpornościowego, podścieliska i innych) w kontekście wieku przy którym pacjentom postawiono pozytywną diagnozę. Przyjęta hipoteza zerowa H_0 mówi, że: *średni wiek przy diagnozie jest taki sam dla wszystkich poziomów komórkowości*, a hipoteza alternatywna H_1 mówi, że: *przynajmniej jedna para grup różni się średnim wiekiem przy diagnozie*.

Tabela 2.4. ANOVA test results for *age_at_diagnosis* by *cellularity*

f-statistic	p-value	Conclusion
12.6639	$3.7897e-06$	Significant differences

Ponieważ $p < 0.05$, odrzucono H_0 i przyjęto H_1 - istnieją istotne różnice pomiędzy grupami. Komórkowość guza jest więc związana z wiekiem przy diagnozie.

2.2. TESTY NIEPARAMETRYCZNE

2.2.1. Test Wilcoxona

Celem przeprowadzenia testu Wilcoxona było sprawdzenie, czy wielkość guza ulegała statystycznie istotnym zmianom dla powtarzających się co pół roku badań dla tych samych pacjentów. Przyjęta hipoteza zerowa H_0 mówi, że: *nie ma istotnej zmiany wielkości guza po 6 miesiącach*, a hipoteza alternatywna H_1 mówi, że: *wielkość guza istotnie zmienia się po 6 miesiącach*.

Tabela 2.5. Wilcoxon signed rank test results for *tumor_size* by *timepoint*

Statistic	n-pairs	p-value	Conclusion
352	48	0.0148	Significant difference

Na podstawie testu Wilcoxona udowodniono statystyczną różnicę w rozmiarze guza po 6 miesiącach. W celu wykrycia kierunku zmiany obliczono mediany różnic pomiędzy pierwszymi i drugimi badaniami – wynik -0.5806 wskazuje na to, że guzy zmniejszają się.

2.3. TEST FRIEDMANA RANK I ANALIZA POST-HOC NEMENYI

Celem przeprowadzenia testu Friedmana było zbadanie, czy średnie rang wielkości guza różnią się istotnie pomiędzy trzema kolejnymi badaniami dla analizowanych pacjentów. Przyjęta hipoteza zerowa H_0 mówi, że: *rozkład (mediana rang) jest taka sama we wszystkich punktach czasowych (brak istotnych zmian guzów w czasie)*, a hipoteza alternatywna H_1 mówi, że: *co najmniej jedna para punktów czasowych różni się istotnie pod względem rang (wielkości guzów)*.

Tabela 2.6. Friedman's rank test results for *tumor_size* by *timepoint*

Statistic	p-value	Conclusion
15.5	0.00043	Significant difference

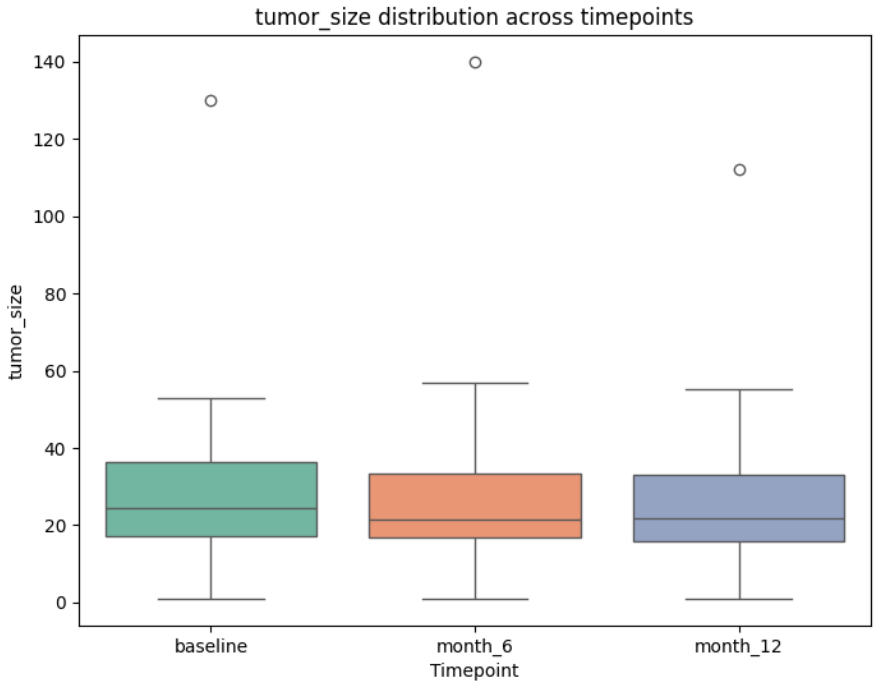
Na podstawie testu Friedmana rank wykazano, że przynajmniej jedna para punktów czasowych różni się istotnie pod względem rang, jako że $p < 0.05$ – z tego względu odrzucono hipotezę zerową H_0 i przyjęto hipotezę alternatywną H_1.

Następnym krokiem było przeprowadzenie testu post-hoc Nemenyi w celu sprawdzenia, które pary punktów czasowych różnią się między sobą w sposób statystycznie istotny.

Tabela 2.7. Nemenyi post-hoc test results for Friedman's rank test

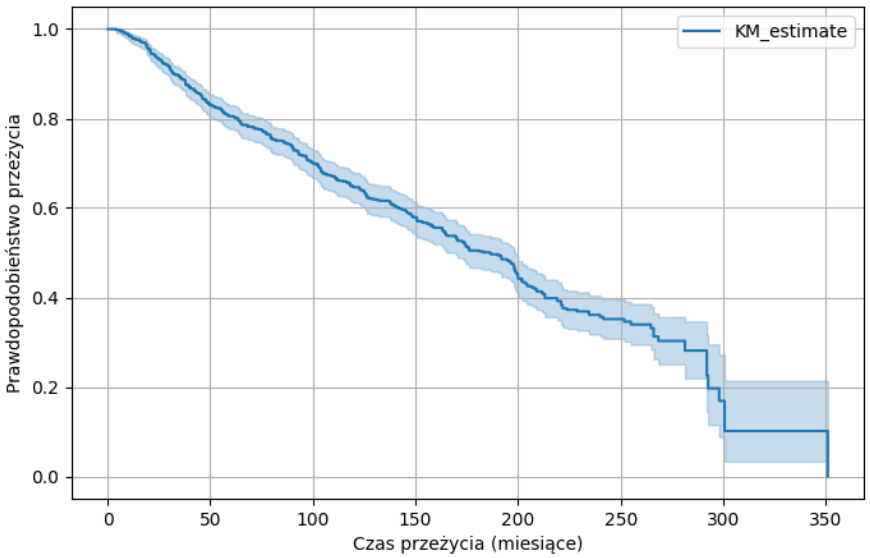
	Baseline	month_6	month_12
Baseline	1.0	0.3750	0.0003
month_6	0.3750	1.0	0.0299
month_12	0.0003	0.0299	1.0

Wyniki testu post-hoc wskazują na to, że różnice pomiędzy badaniem pierwszym, a badaniem przeprowadzonym po 6 miesiącach nie są statystycznie istotne, ponieważ $p > 0.05$ – pozostałe pary (badanie podstawowe a badanie po 12 miesiącach oraz badania kolejno po 6 i 12 miesiącach) różnią się od siebie znacząco. Rozkład zmiennej ilościowej zaprezentowany został na rysunku 2.1.



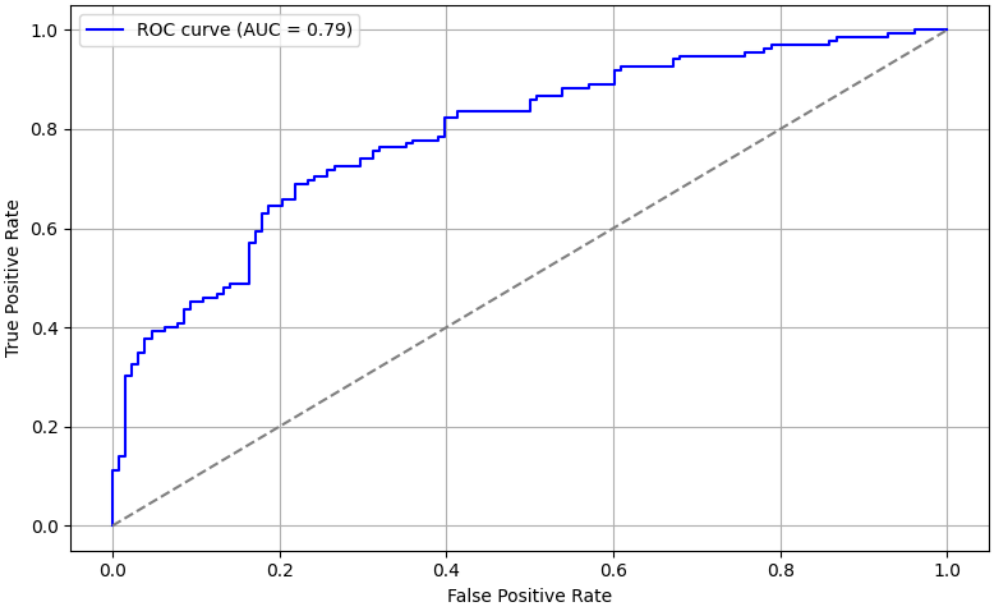
Rys. 2.1. Wykres przedstawiający rozkład zmiennej ilościowej *tumor_size* na przestrzeni kroków czasowych.

3. KRZYWA PRZEŻYCIA



Rys. 3.1. Wykres Kaplana-Mayera przedstawiający prawdopodobieństwo przeżycia w kolejnych punktach czasu dla czasu przeżycia *overall_survival_months* i zdarzenia *overall_survival_status*.

4. REGRESJA LOGISTYCZNA



Rys. 4.1. Krzywa ROC dla regresji logistycznej przewidującej status przeżycia na podstawie zebranych cech ilościowych.