

Studia podyplomowe SAS

Syllabus

Semestr 1.

1. Wprowadzenie do środowiska SAS

7 godzin **mgr inż. Wojciech Uss**

- | | |
|--|-----|
| 1) Architektura platformy SAS: | 2 h |
| a) podział platformy na warstwy, | |
| b) serwery. | |
| 2) SAS Management Console – narzędzie do administracji. | 2 h |
| 3) Podstawowe zadania administracyjne: | 2 h |
| a) serwery (uruchomienie, zatrzymanie, sprawdzenie stanu), | |
| b) użytkownicy (dodawanie, usuwanie, grupy). | |
| 4) Kontrola dostępu. | 1 h |

2. Język 4GL – przetwarzanie danych w SAS

14 godzin **dr inż. Anna Szafrńska**

- | | |
|--|-------|
| 1) Wprowadzenie – środowisko pracy, biblioteki | 0.5 h |
| 2) Programowanie w SAS – PROC STEP, DATA STEP | 0.5 h |
| 3) Zbiory danych w SAS | 4 h |
| a) czytanie zbioru – instrukcja SET, MERGE, instrukcja GO TO, LINK | |
| b) filtrowanie danych – wierszy, kolumn, zmiana nazw zmiennych | |
| c) łączenie zbiorów danych – instrukcje SET, MERGE | |
| d) transpozycja zbiorów danych | |
| 4) Obliczenia na danych | 5 h |
| e) wyrażenia SAS – stałe numeryczne, znakowe, operatory | |
| f) funkcje SAS | |
| g) instrukcja warunkowa IF, pętla DO | |
| h) instrukcja warunkowa IF, pętla DO | |
| i) przechowywanie informacji w pętli głównej – RETAIN, LAG, DIF | |
| j) sortowanie zbiorów i przetwarzanie w grupach | |
| k) tablice i tablice tymczasowe | |
| 5) Formaty | 2 h |
| 6) Tworzenie i zastosowanie indeksów | 1 h |
| 7) Procedury usługowe | 1 h |

3. Makroprogramowanie w SAS

7 godzin **dr inż. Anna Szafrńska**

- | | | |
|--|-----|-----|
| 1) Makrozmiennne | | 1 h |
| 2) Systemowe makrozmiennne | 1 h | |
| 3) Makrofunkcje | 1 h | |
| 4) Tworzenie i zastosowanie makroprogramów | 4 h | |
| a) użycie makroprogramów | | |
| b) globalne i lokalne zmienne | | |
| c) instrukcja warunkowa IF, pętla DO | | |

- d) przechowywanie makroprogramów

4. Język SQL

7 godzin

mgr inż. Dorota Kuziak

Celem zajęć jest zapoznanie się ze składnią i strukturą języka SQL (Structured Query Language).

Ponadto poruszane również będą tematy związane z definiowaniem zagnieżdżonych zapytań, wykonywaniem złożonych łączy zbiorów oraz nakładaniem więzów integralności. Studenci nauczą się także, jak wykorzystać język SQL do tworzenia zbiorów, perspektyw oraz raportów na podstawie danych dostępnych w SAS.

- | | |
|---|-----|
| 1) Wprowadzenie do języka SQL | 1 h |
| 2) Podstawowe zapytania: wybór zmiennych, agregacja danych i podzapytania | 2 h |
| 3) Łączenie zbiorów pionowe i poziome | 1 h |
| 4) Tworzenie i modyfikowanie zbiorów i perspektyw | 1 h |
| 5) Tworzenie indeksów | 1 h |
| 6) Dodatkowe możliwości języka SQL: | 1 h |
| a) opcje procedury SQL, | |
| b) tablice słownikowe, | |
| c) wykorzystanie zapytań SQL w makroprogramowaniu, | |
| d) zagadnienia wydajnościowe. | |

Podczas tych zajęć zostanie wykorzystane oprogramowanie:

- SAS Foundation,
- SAS Enterprise Guide.

5. Język macierzowy IML

14 godzin

mgr inż. Wojciech Uss

- | | |
|--|-----|
| 1) Wprowadzenie do języka IML | 9 h |
| a) Praca z macierzami | |
| b) Konstrukcje programistyczne, moduły | |
| c) Praca ze zbiorami danych SAS | |
| d) Praca z plikami zewnętrznymi | |
| e) Tworzenie wykresów | |
| 2) Implementacja przykładowych algorytmów statystycznych | 5 h |

6. Integracja danych z SAS Data Integration Server

14 godzin

mgr inż. Dorota Kuziak

Celem zajęć jest zapoznanie się z kompleksowym środowiskiem SAS Data Integration Studio pozwalającym na realizację procesów integracji danych.

- | | |
|---|-----|
| 1) Interfejs SAS Data Integration Studio oraz tematy związane z przygotowaniem i organizacją środowiska pracy | 6 h |
| a) profile użytkowników, | |
| b) funkcjonalność środowiska SAS Data Integration Studio, | |
| c) dostęp i praca z danymi: projektowanie struktur źródłowych – Source Designer, | |
| d) projektowanie struktur docelowych – Target Designer oraz praca z plikami płaskimi, | |
| e) synchronizacja warstwy metadanych z warstwą modelu fizycznego danych, | |
| f) architektura procesów ETL w architekturze SAS 9: Top-Down i Bottom-Up, | |

- g) usługi obliczeniowe wykorzystywane przez SAS Data Integration Studio: Workspace Server i Remote processing,
 - h) konstrukcja i funkcjonalność transformacji: reguły mapowania, pozycje wyliczane,
 - i) struktury danych (tabela, widok), debugging, definiowanie mechanizmu Alert Notification, parametryzacja transformacji oraz mechanizm Pre- i Post Processing,
 - j) transformacje użytkownika: Extract, Transform i Load,
 - k) Data Loader.
- 2) Praktyczne wykorzystanie środowiska SAS Data Integration Studio w procesie projektowania zadań integracji danych 6 h
- a) parametryzacja procesów,
 - b) przetwarzanie synchroniczne i asynchroniczne,
 - c) zbiory słownikowe,
 - d) zarządzanie zmianami w zbiorach słownikowych,
 - e) walidacja danych,
 - f) obsługa wyjątków przetwarzania,
 - g) promocja i replikacja obiektów,
 - h) praca grupowa i zarządzanie zmianami: typy repozytoriów, operacja check-in/check out, fetch, oraz historia zmian,
 - i) harmonogram zadań,
 - j) zależności między zdaniami,
 - k) autoryzacja dostępu.
- 3) Własne transformacje używane w SAS Data Integration Studio (trzy sposoby realizacji nowych transformacji) 2 h
- a) wprowadzenie,
 - b) projektowanie transformacji,
 - c) mechanizm importu/eksportu transformacji,
 - d) kreator transformacji.

Podczas tych zajęć zostanie wykorzystane oprogramowanie:

- SAS AppDev Studio,
- SAS Data Integration Studio,
- SAS Foundation,
- SAS Management Console.

7. Jakość danych

14 godzin

mgr Magdalena Godlewska

Celem zajęć jest zapoznanie z problematyką jakości danych, a dokładniej – poszczególnymi etapami w procesie poprawy jakości danych.

- 1) Profilowanie – ocena jakości danych jako pierwszy etap poprawy jakości danych 5 h
 - a) typy danych w obszarze Data Quality,
 - b) predefiniowane metryki oceniające jakość danych: miary statystyczne, miary techniczne, analiza wzorców, analiza częstości i wartości skrajne,
 - c) efektywność modelu danych,
 - d) zależności i związki w modelu danych – analiza kompletności kluczy głównych i obcych oraz analiza redundancji,
 - e) tworzenie własnych metryk,
 - f) budowa procesu profilowania danych,
 - g) analiza dużych wolumenów danych.
- 2) Standaryzacja jako kolejny etap w procesie poprawy jakości danych 5 h

- a) SAS Quality Knowledge Base: architektura i funkcjonalność oraz wykorzystanie bazy wiedzy o budowie języka polskiego,
 - b) czynniki wpływające na efektywność standaryzacji – parsing wg typów danych,
 - c) schematy standaryzacyjne,
 - d) budowa schematów: łączenie schematów i schematy użytkownika,
 - e) wzbogacanie danych – analiza płci i identyfikacja,
 - f) miary oceny podobieństwa danych,
 - g) budowa procesów standaryzacyjnych,
 - h) rozbudowa SAS Quality Knowledge Base.
- 3) Reduplikacja rekordów – końcowy etap procesu poprawy jakości danych 4 h
- a) metody grupowania (clustering) danych,
 - b) wykorzystanie SAS Quality Knowledge Base,
 - c) tworzenie kodów dopasowań (match codes),
 - d) identyfikacja tzw. surviving rekord – zasady „najlepszego rekordu” i „złotego rekordu”,
 - e) analiza skuteczności reduplikacji.
- Podczas tych zajęć zostanie wykorzystane oprogramowanie:
- SAS dfPowerStudio,
 - SAS Foundation.

8. Wizualizacja i raportowanie w środowisku SAS Enterprise Guide

14 godzin dr inż. Magdalena Lemańska

- 1) Środowisko pracy SAS Enterprise Guide 1 h
- 2) Praca z danymi zewnętrznymi – import i export danych (m.in. do i z Excela) 2 h
- 3) Techniki agregacji danych i przetwarzanie danych w grupach 4 h
- 4) Raportowanie: 7 h
 - a) przygotowanie danych do raportu,
 - b) listowanie danych,
 - c) tabele zagregowane,
 - d) wykresy słupkowe i kołowe,
 - e) wykresy mapowe – mapa geograficzna,
 - f) wykres radarowy,
 - g) analiza kontyngencji,
 - h) formaty i opcje raportów,
 - i) budowa dokumentów.

9. Struktury wielowymiarowe – OLAP

7 godzin dr inż. Joanna Raczek

Zajęcia ze struktur wielowymiarowych OLAP będą dotyczyły projektowania i budowy struktury OLAP w środowisku SAS OLAP Cube Studio. Prezentowane zagadnienia mają na celu umożliwienie użytkownikom swobodnego posługiwania się efektywnym narzędziem analizy i prezentacji danych, jakim są kostki OLAP.

- 1) Omówienie struktur wielowymiarowych. 1 h
- 2) Modele danych dla struktur OLAP. 1 h
- 3) Nawigacja po strukturze OLAP. 1 h
- 4) Konfiguracja i wykorzystanie mechanizmu Drill-Through. 1 h
- 5) Wymiary, hierarchie, agregaty, wykorzystanie pozycji wyliczanych w OLAP. 2 h
- 6) Przenoszenie struktur OLAP między środowiskami. 1 h

10. SAS Business Intelligence Server

14 godzin mgr inż. Magdalena Dettlaff

SAS Enterprise BI Server jest kompleksowym rozwiązaniem z dziedziny analityki biznesowej. BI ogólnie można przedstawić jako proces przekształcania danych w informacje, a informacji w wiedzę, która może być wykorzystana do zwiększenia konkurencyjności przedsiębiorstwa.

Funkcjonalność środowiska SAS Enterprise BI Server pozwala na tworzenie analiz i raportów, ale również na odpowiednie przygotowanie danych. Praca z danymi i raportami możliwa jest zarówno z poziomu aplikacji biurowych (np. MS Excel), przeglądarki WWW, jak i dedykowanego interfejsu SAS. Dostęp do tak przygotowanych danych możliwy jest poprzez, wspólną warstwę metadanych.

- 1) Integracja z Microsoft Office: 4 h
 - a) możliwości dostępu do danych,
 - b) możliwości raportowania i wykonywania analiz dostępnych w Systemie SAS, pracując w środowisku narzędzi Microsoft Office.
 - c) wykorzystywanie zaawansowanych możliwości analitycznych
- 2) Budowa map informacyjnych: 4 h
 - a) generowanie raportów i analiz operujących pojęciami biznesowymi
- 3) Wykorzystanie webowego środowiska budowy i dystrybucji raportów 6 h
 - a) praca z repozytorium danych wielowymiarowych, które udostępnia uprzednio zagregowane dane, generowane na podstawie wielu różnorodnych źródeł,
 - b) obsługa interfejsów webowych pozwalających na przeglądanie danych pod różnymi kątami, z możliwością dojścia do szczegółowej informacji,
 - c) dołączanie wykresów lub map.

Semestr 2.

11. Probabilistyka i wprowadzenie do statystyki

35 godzin dr hab. Karol Dziedziul mgr inż. Agnieszka Wałachowska

Opis ogólny: Do zajęć z probabilistyki będą przygotowane „programiki” wizualizujące zagadnienia probabilistyczne. Będziemy rozważać zagadnienia probabilistyczne na przykładzie problematyki finansowej i ubezpieczeniowej. Przykładowe procedury: proc rand; proc kde; proc univariate. Do statystyki wykorzystana zostanie interface SAS Enterprise Guide.

- 1) Generowanie zmiennych losowych. 4 h
 - a) Teoretyczne miary ryzyka (rezerwy finansowe): Var, czyli kwantyle.
 - b) Własności Var dla portfela, czyli zmienne losowe dwuwymiarowe.
- 2) Problemy ubezpieczeń majątkowych: problem bankructwa, czyli sumy wielu zmiennych losowych. 3 h
- 3) Model cen akcji, aktywów. Model wyceny obligacji Mertona, czyli zagadnienia modelowania procesu Wienera oraz równań Blacka Scholesa. 4 h
- 4) Kryteria statystyczne do oceny uzyskanej generacji, czyli testy nieparametryczne. 2 h
- 5) Statystyka, czyli sztuka wnioskowania z danych na przykładzie miar ryzyka. Empiryczne miary ryzyka. 4 h
- 6) Testy statystyczne. 3 h
- 7) Analiza kontyngencji. 3 h

- 8) Statystyka dla danych skategoryzowanych. Paradoks Simpsona. 4 h
- 9) Statystyka dla danych ciągłych, proste testy parametryczne. 4 h
- 10) Statystyka dla wielu zmiennych (badanych cech), macierze korelacji, dane nadmiarowe. 4 h

12. Analiza regresji, współzależności i wariancji z punktu widzenia Data Mining

28 godzin

dr inż. Krzysztof Świetlik

mgr inż. Magdalena Kucharska

Strona | 6

- 1) Analiza regresji liniowej 11 h
 - a) Model jednorównaniowy regresji liniowej.
 - b) Założenia klasycznej metody najmniejszych kwadratów.
 - c) Metoda najmniejszych kwadratów.
 - d) Metoda największej wiarygodności.
 - e) Estymacja parametrów modelu z jedną zmienną egzogeniczną. Współczynnik korelacji liniowej Pearsona.
 - f) Estymacja parametrów modelu z wieloma zmiennymi egzogenicznymi.
 - g) Testowanie istotności parametrów modelu (test t-studenta, test mnożnika Lagrange'a).
 - h) Analiza dobroci dopasowania (współczynnik korelacji wielorakiej, współczynnik determinacji).
 - i) Dobór zmiennych objaśniających.
 - j) Badanie współliniowości zmiennych objaśniających (test Farrara-Glaubera).
 - k) Testowanie własności stochastycznych modelu (test normalności składnika losowego Jarque-Bera i Shapiro-Wilka, test heteroskedastyczności White'a i Harrisona-McCabe'a, test autokorelacji składnika losowego Durbin-Watsona).
- 2) Analiza regresji nieliniowej 7 h
 - a) Nieliniowy jednorównaniowy model regresji.
 - b) Liniowość i nieliniowość modelu względem parametrów strukturalnych.
 - c) Linearyzacja modeli nieliniowych i ich szacowanie metodą najmniejszych kwadratów.
 - d) Modele ściśle nieliniowe i ich estymacja metodami numerycznymi.
- 3) Wielorównaniowe modele regresji liniowej 5 h
 - a) Identyfikowalność modeli wielorównaniowych.
 - b) Szacowanie parametrów strukturalnych pośrednią i podwójną metodą najmniejszych kwadratów.
- 4) Analiza korelacji: 5 h
 - a) Analiza skupień
 - b) Analiza czynnikowa
 - c) Analiza dyskryminacyjna.

13. Analiza i prognozowanie szeregów czasowych w oparciu o metody Data Mining

28 godzin

dr Krystyna Ambroch

dr inż. Krzysztof Świetlik

- 1) Składowe szeregu czasowego. 2 h
- 2) Estymacja tendencji rozwojowej liniowej i nieliniowej. 3 h
- 3) Analiza wahań cyklicznych addytywnych i multiplikatywnych. 3 h
- 4) Wyrównywanie wykładnicze. 2 h
- 5) Modele z rozkładem opóźnień i modele autoregresyjne. 2 h
- 6) Estymacja modeli z nieskończonym rozkładem opóźnień. 2 h
- 7) Prognozowanie z modeli adaptacyjnych. 3 h

- | | |
|---|-----|
| 8) Prognozowanie z modelu ekonometrycznego. | 2 h |
| 9) Wprowadzenie do modeli ARMA. | 2h |
| 10) Modele SARIMA. | 2h |
| 11) Prognozowanie z modeli SARIMA. | 3h |
| 12) Model GARCH. | 2h |

14. Analiza przeżycia

7 godzin **dr Krystyna Ambroch**

- | | |
|---|----|
| 1) Analiza tablic trwania życia | 1h |
| 2) Rozkłady funkcji przeżycia | 1h |
| 3) Estymator Kaplana-Meiera | 1h |
| 4) Porównanie grup z obserwacjami uciętymi | 2h |
| 5) Modele regresji dla danych dotyczących czasu przeżycia | 2h |

15. Data Mining

14 godzin **dr hab. Karol Dzedziul**
mgr inż. Wojciech Uss

- | | |
|---|------|
| 1) Case study (Karol Dzedziul, Wojciech Uss, Magdalena Kucharska) | 14 h |
|---|------|
- Zaprezentowane zostaną całościowe analizy danych biznesowych (Business Case Studies) do wyboru przez grupę:
- opis klienta internetowego,
 - analiza marketingowa,
 - model zadowolenia klienta,
 - opis ryzyka dla small businesses,
 - ryzyko operacyjne.