Raport do projektu z przedmiotu Wybrane algorytmy i struktury danych

Anna Wawrzyńczak

Grudzień 2022

Spis treści

Rozdział 1

Wstęp

1.1 Opis zadania

Na zadanie projektowe składało się znalezienie zbioru danych, jego analiza i stworzenie modelu o charakterze predykcyjnym dla wybranej zmiennej za pomocą programów SAS Enterprise Guide i SAS Enterprise Miner (przy wykorzystaniu sesji WLATIN1).

Projekt opiera się na zbiorze danych "Absenteeism in work" dostępnej na stronie UCI Machine Learning Repository (adres do strony podany w bibliografii). Model projektowy na podstawie podanych danych o pracownikach przewiduje, do której grupy pod wgzlędem nieobecności w pracy pracownik należy.

1.2 Wprowadzenie do tematyki projektu

Na efekty pracy pracowników i zyski firmy ma wpływ wiele czynników. Przeanalizowanie tych zmiennych oraz wprowadzenie na podstawie tej analizy odpowiednich zmian w firmie sprawia, że możliwe będzie poprawienie wydajności osób pracujących oraz zwiększenie zysków firmy.
Jedną z najważniejszych cech charakteryzujących pracownika, które mają pośredni wpływ na działalność firmy, jest liczba godzin nieobecności w pracy. Istotne jest znalezienie występujących trendów pod względem liczby osób zatrudnionych biorących godziny lub dzień wolny w danym okresie czasu, aby firma mogła podjąc odpowiednie działania z powodu przewidywalnego deficytu pracowników w danym czasie skutkującym gorszymi efektami pracy. Ponadto, na podstawie liczby godzin nieobecności w pracy można opisać stosunek osoby pracującej do swojej funkcji, co ma znacząco wpływ na decyzję w zakresie promocji lub zwolnienia danego pracownika.

W ramach projektu został stworzony model, który przyporządkowuje pracownika do odpowiedniej grupy pod względem częstości nieobecności w pracy od momentu zatrudnienia. Model może być używany m.in. przez praconików zajmujących się ewaluacją lub monitorowaniem pracy osób zatrudnionych.

Rozdział 2

Opis zbioru danych

2.1 Ogólne informacje

Analizowanym zbiorem danych jest zbiór "Absenteeism in work", który jest zapisany w formacie CSV. Zbiór składa się z 21 kolumn i 740 wierszy. Dane opisują pracowników firmy kurierskiej w Brazyli w latach 2007-2010. Zbiór został przygotowany przez Andrea Martiniano, Ricardo Pinto Ferreira oraz Renato Jose Sassi.
Każdy z rekordów przedstawia informacje na temat liczby godzin nieobecności oraz cechy charakteryzujące danego pracownika w trakcie wybranego dnia pracy. Zbiór składa sie z następujących kolumn:

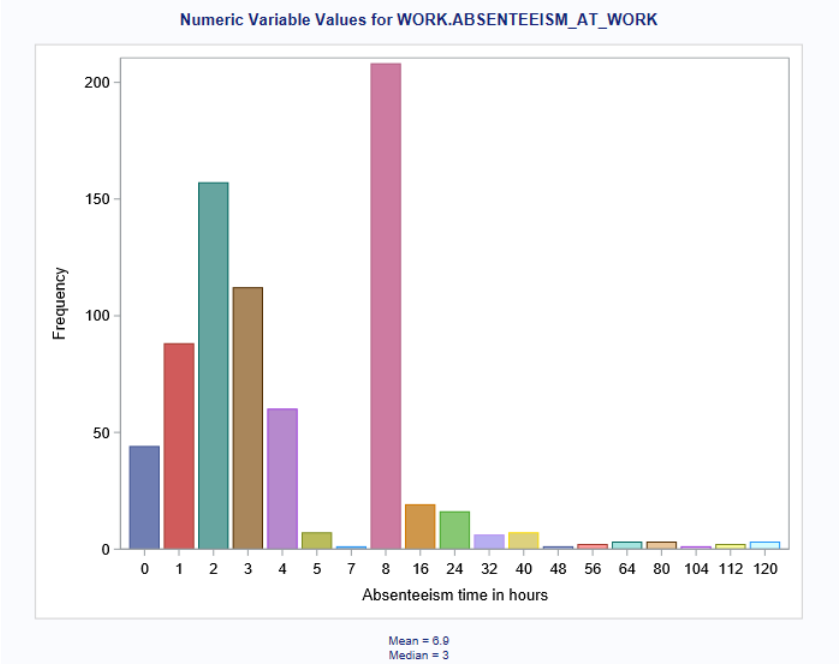
- *ID* - kolumna odpowiadająca numeru identyfikacyjnemu pracownika;
- *Month of absence* - kolumna przedstawiająca miesiąc nieobecności;
- *Day of the week* - kolumna przedstawiająca dzień tygodnia nieobecności;
- *Seasons* - kolumna opisująca porę roku nieobecności;
- *Distance from Residence to Work* - kolumna przedstawiająca dystans pomiędzy domem pracownika a miejscem pracy podany w kilometrach;
- *Transportation expense* - kolumna dotycząca kosztu transportu;
- *Service time* - kolumna opisująca czas zatrudnienia;
- *Age* - kolumna przedstawiająca wiek pracownika;
- *Work load Average/day* - kolumna przedstawiająca średnie obciążenie pracą na dzień;
- *Hit target* - kolumna dotycząca liczby spodziewanych dostaw na dzień;
- *Disciplinary failure* - kolumna dotycząca upomnień dyscyplinarnych;
- *Education* - kolumna odpowiadająca poziomowi edukacji pracownika;
- *Son* - kolumna przedstawiająca liczbę dzieci pracownika;
- *Pet* - kolumna odpowiadająca liczbie zwierząt domowych posiadanych przez pracownika;
- *Weight* - kolumna przedstawiająca wagę pracownika;
- *Height* - kolumna przedstawiająca wzrost pracownika;
- *Social drinker* - kolumna opisująca czy pracownik pije alkohol;
- *Social smoker* - kolumna opisująca czy pracownik pali;
- *BMI* - kolumna odpowiadająca wartości BMI pracownika;
- *Absenteeism time in hours* - kolumna przedstawiająca liczbę godzin nieobecności danego dnia;
- *Reason for absence* - kolumna opisująca powód nieobecności.

Powody nieobecności zostały podzielone na 28 kategorii, z czego 22 dotyczą konkretnych przypadków chorób lub urazów medycznych, zaś inne dotyczą kosnultacji z lekarzem, z dentystą, donacji krwi, badania laboratoryjnego, fizjoterpaii oraz nieusprawiedliwionych powodów nieobecności.
Niektóre dane takie jak miesiąc, dzień tygodnia zostały przedstawione jako liczby odpowiadające chronologicznej kolejności miesięcy w roku lub dni w tygodniu.

Rozdział 3

Analiza zbioru danych i preprocessing

Analiza zbioru została przeprowadzona w SAS Enterprise Guide. Na początku stworzono i przypisano do projektu bibliotekę pod nazwą *WORK*. Następnie za pomocą dostępnych w SAS Enterprise Guide funkcji *Import data* zimportowano zbiór danych "Absenteeism in work", który był zapisany na lokalnym dysku komputera.
W kolejnym kroku przeprowadzono analizę zbioru danych pod względem rozkładów i wartości odstających zmiennych oraz korelacji pomiędzy zmiennymi przy wykorzystaniu kafelków *Characterize Data*, *Histogram* oraz *Summary Statistics*. Szukano również błędnych danych, tzn. danych, które zgodnie z posiadaną wiedzą przedstawiają nierealistyczne wartości.
Zauważono, że zbiór składa się tylko z danych numerycznych, z których niektóre opisują dane nominalne.
Najbardziej zróżnicowane są dane dotyczące liczby godzin nieobecności. Zawierają się w przedziale od 0 do 120, co pokazuje Rysunek 3.1.
Ponadto, zbiór nie zawiera żadnych braków danych.



Rysunek 3.1: Histogram przedstawia rozkład liczby nieobecności pod względem godzin.

3.1 Błędne dane

Niektóre dane z zbioru są nierealistyczne lub nielogiczne. W zbiorze jest 36 różnych numerów identyfikacyjnych, co sugeruje, że w firmie pracowało 36 osób. Jednakże, wiek pracownika o ID równym 29 zmienił się o 13 lat w przedziale 2007-2010, co pokazuje Rysunek 3.2.

Stąd można wywnioskować, że jako pracownik o ID równym 29 pracowały 2 różne osoby, z czego po odejściu pierwszej osoby nowemu pracownikowi nadali taki sam numer identyfikacyjny jak poprzednikowi. Aby utrzymać unikatowość ID pracowników nadano pracownikowi w wieku 28 lat nowy numer identyfikacyjny 37. Zatem zbiór

	ID	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age
1	29	0	9	2	4	225	26	9	28
2	29	28	2	6	2	225	15	15	41
3	29	19	5	4	3	225	15	15	41
4	29	14	5	5	3	225	15	15	41
5	29	22	5	6	3	225	15	15	41

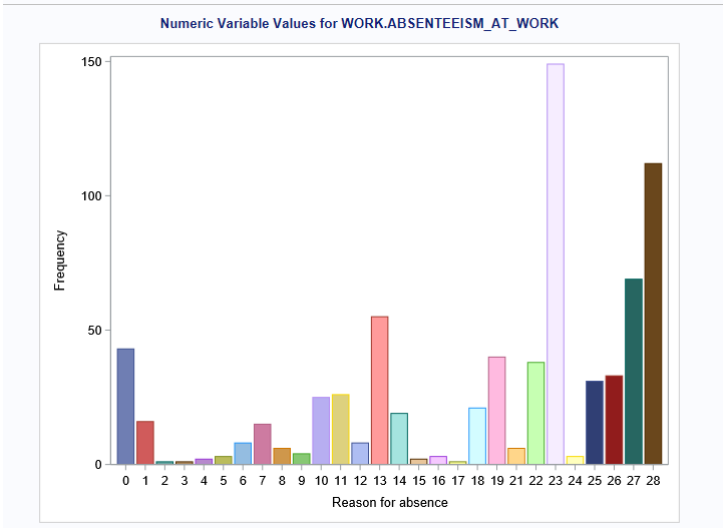
Rysunek 3.2: Dane o pracowniku z ID równym 29.

opisuje 37 pracowników.
Kolejnym przykładem błędnych danych są rekordy o numerze miesiąca równym 0 przedstawione na Rysunku 3.3. Wtedy miesięcy w ciągu roku jest 13, co wskazuje na błąd. Rekordów tych jest 3 i każdy z nich ma 0 godzin nieobecności, więc zostają usunięte ze zbioru.

ID	Reason for absence	Month of absence
4	0	0
35	0	0
8	0	0
15	3	1
32	10	1

Rysunek 3.3: Dane o numerze miesiąca równym 0.

Jak widać na Rysunku 3.4 w kolumnie opisującej powody nieobecności wartości znajdują się w przedziale od 0 do 28 mimo, że w opisie zostało sprecyzowane, że kategorii jest 28. Rekordów o powodzie nr 0 jest 43. Zatem te rekordy będą interpretowane jako dotyczące nieobecności z nieznanych powodów.



Rysunek 3.4: Histogram przedstawia rozkład danych ze względu na powód nieobecności

Ponadto, błędne dane zostały znalezione dla rekordu o numerze powodu nieobecności równym 27, który opisuje fizjoterapię. Jeden z takich rekordów ma podane 0 godzin nieobecności co pokazuje Rysunek 3.5, co by oznaczało, że pracownik o ID 34 był nieobecny przez 0 godzin ze względu na fizjoterapię. Jest to niemożliwe. Zatem w tym wierszu za liczbę godzin nieobecności została podstawiona średnia godzin nieobecności z powodu fizjoterapii innych pracowników (równa 2).

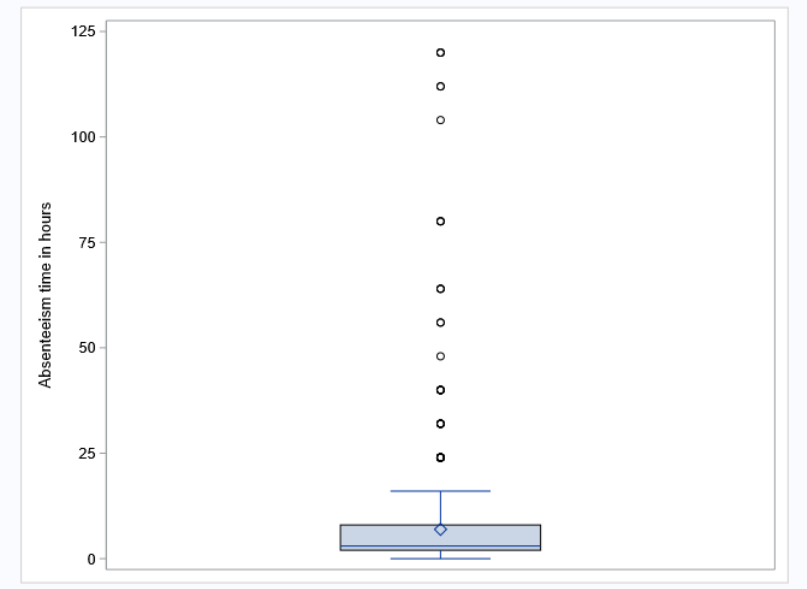
3.2 Przekształcenia danych

Podczas sprawdzania wartości odstających dla zmiennych zauważono, że wartości odstające dla danych z kolumny dotyczącej liczby godzin nieobecnych oraz dla kolumny dotyczącej obciążenia roboczego najbardziej wpływają na

	ID	Reason for absence	Absenteeism time in hours
1	34	27	0
2	28	27	1
3	34	27	1
4	22	27	2
5	28	27	2

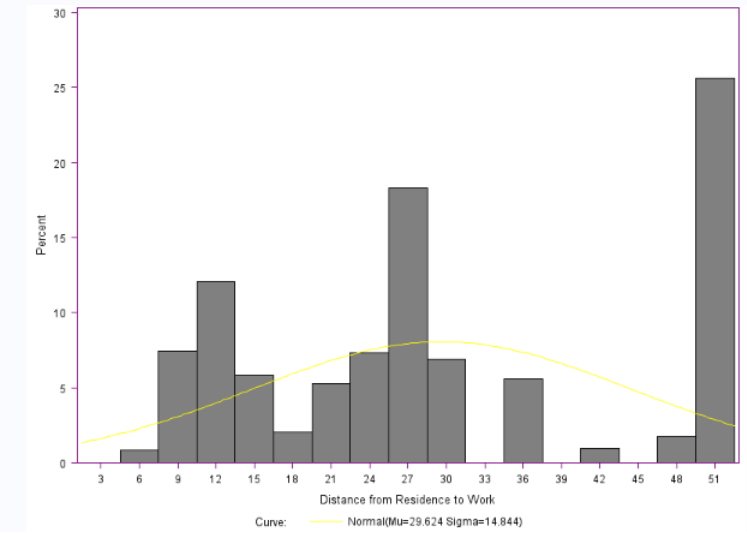
Rysunek 3.5: Dane przedstawiające rekordy dotyczące nieobecności z powodu fizjoterapii.

średnią tych cech (Rysunek 3.6). Zatem wartości odstające zostały zamienione na wartości odpowiednio kwantyla 97 i 95 .



Rysunek 3.6: Boxplot przedstawiający godziny nieobecności.

Jak można zauważyć na podstawie wykresu w Rysunek 3.7 dane opisujące dystans z domu pracownika do pracy można podzielić na 3 zbiory. Pierwsza grupa nazwana *close* w przedziale od 0 do 15 opisuje pracowników, którzy mieszkają blisko miejsca pracy, druga nazwana *far* w przedziale od 15 do 35 oraz trzecia nazwana *very far* opisująca rekordy o dystansie większym od 35.
Analogicznie podzielono dane z kolumny opisującej koszt transportu przedstawione na Rysunek 3.8. Pierwsza grupa rekordów dotycząca kosztów mniejszych od 180 została nazwana *low*, druga nazwana *average* obejmuje koszty w przedziale 180-248 oraz trzecia o kosztach większych niż 248 została nazwana *high*.



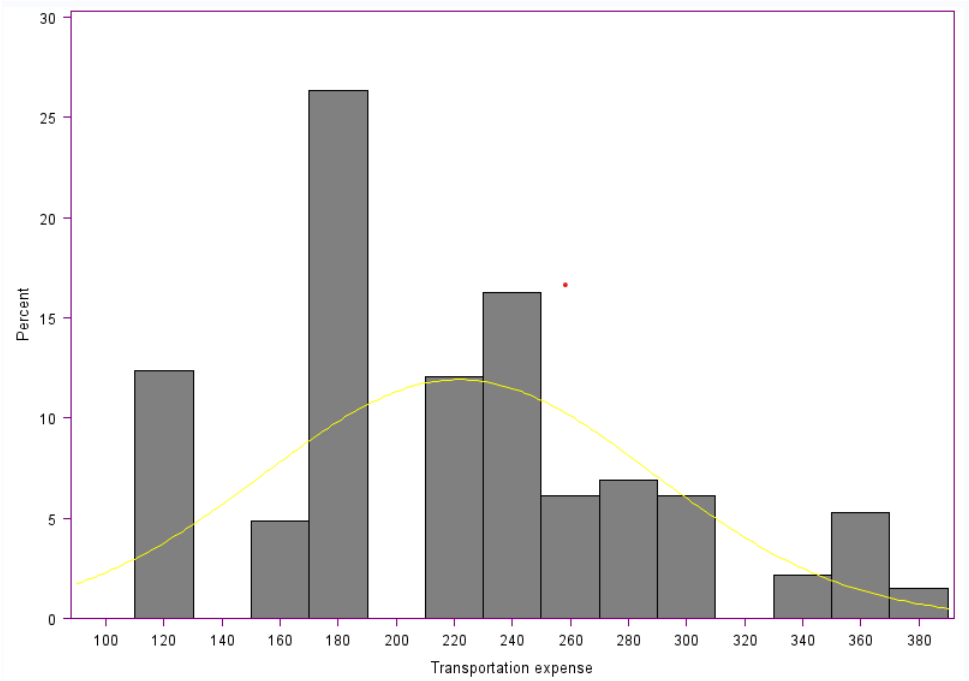
Rysunek 3.7: Histogram pokazuje rozkład danych opisujących dystans z domu pracownika do pracy.

Następnie obliczono całkowitą liczbę godzin nieobecności z danego miesiąca dla każdego pracownika. Zauważono, że dane były zebrane z okresu 3 lat, lecz nie ma kolumny, która opisywałaby, z którego roku pochodzą dane. Zatem zsumowanie liczby godzin nieobecnych w danym miesiącu dla danego pracownika daje pełną liczbę godzin nieobecności w ciągu danego miesiąca w ciągu 3 lat, więc aby uzyskać szukaną wartość podzielono sumę na 3. Na podstawie obliczonej nowej zmiennej *absent_hours_per_month1* (której rozkład przedstawiony jest na Rysunek 3.9) można podzielić pracowników na 3 grupy:

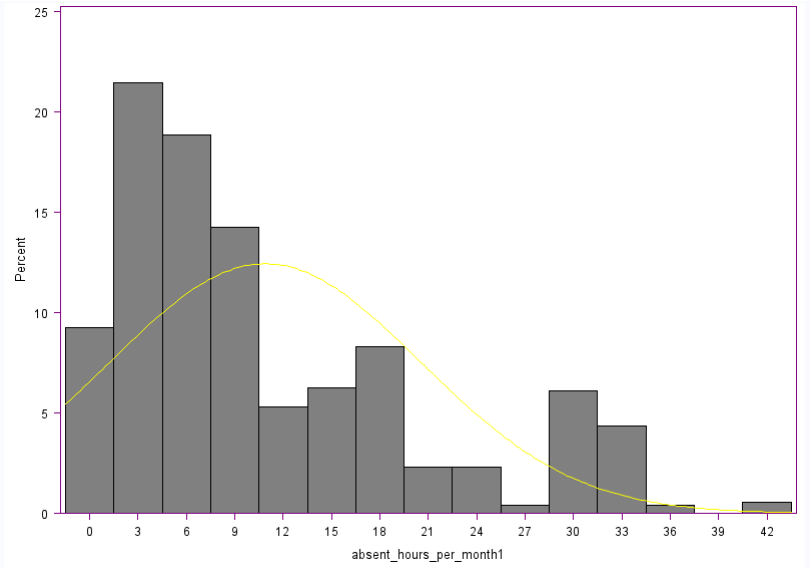
- *non-absent* - pracownicy, którzy średnio na miesiąc są nieobecni przez mniej niż 5 godzin;
- *moderate* - pracownicy, których miesięczna średnie liczba nieobecności mieści się w przedziale 5-20 godzin
- *excessive* - pracownicy, których średnia liczba godzin nieobecności na miesiąc przekracza 20

Na podstawie wyżej wymienionych grup model będzie przeprowadzać klasyfikację pracowników. Warto zauważyć, że każdy z pracowników ma przypisany inny typ nieobecności na każdy miesiąc pracy.
Po analizie scatter plots (Rysunek 3.10) i korelacji pomiędzy różnymi zmiennymi zauważono, że zmienne waga i BMI są mocno ze sobą skorelowane, więc można usunąc ze zbioru kolumnę BMI.

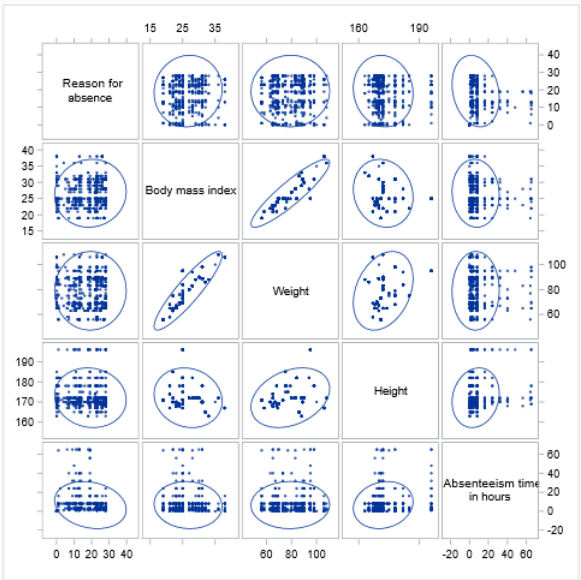
Ponadto, rozkład zmiennej opisującej czy pracownik pali papierosy w każdej z klas jest podobny (Rysunek 3.11), zatem ta kolumna również została usunięta.
Na koniec preprocessingu i analizy zbioru nazwy kolumn zbioru musiały zostać zmienione na nazwy z małymi literami i bez spacji, aby zbiór mógł zostać wczytany w SAS Minerze oraz została wyeksportowana tabela sasowa *db_to_extract*, która następnie została wykorzystana w kolejnej części projektu.



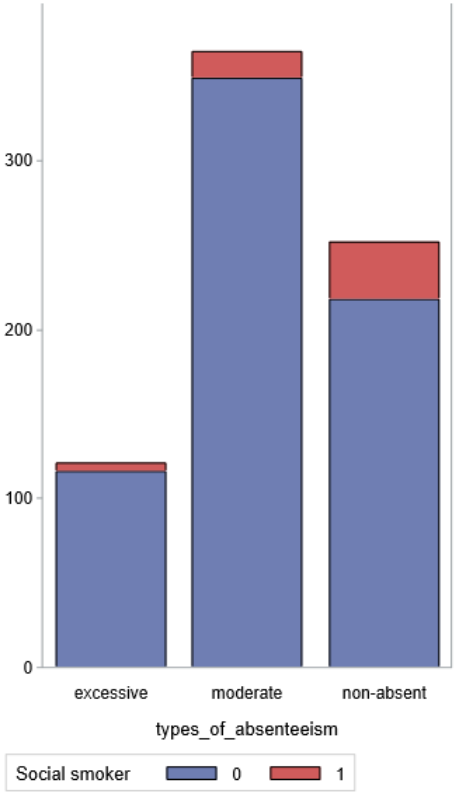
Rysunek 3.8: Histogram przedstawia rozkład danych opisujących koszty transportu.



Rysunek 3.9: Histogram przedstawia rozkład nowej zmiennej określającej całkowitą liczbę godzin nieobecnych dla danego pracownika w ciągu miesiąca.



Rysunek 3.10: Macierz scatter plots dla zmiennych BMI, waga, wzrost, powód nieobecności i godziny nieobecne.



Rysunek 3.11: Wykres słupkowy pokazuje rozkład zmiennej *Social smoker* w każdym z typów nieobecności.

Rozdział 4

Tworzenie modelu

4.1 Testowanie różnych modeli

Ta część projektu została wykonana w SAS Enterprise Miner.
Po zimportowaniu pliku z SAS Enterprise Guide i oraz zdefiniowaniu odpowiednich typów dla zmiennych (Rysunek 4.1) stworzono diagram o nazwie *fulldiagram*. Przeprowadzono podział danych na zbiór uczący (50%), walidacyjny (30%) oraz testowy (20%). Na tych zbiorach najpierw przeprowadzono testy z modelem drzewa decyzyjnego oraz lasu losowego. Zauważono, że duży wpływ na klasyfikację w obu modelach miała zmienna *ID* określająca numer identyfikacyjny pracownika (co jest pokazane na Rysunek 4.2 oraz Rysunek 4.3). ID pracownika nie powinno mieć wpływu na klasyfikację względem jego nieobecności zatem stworzono nową tabelę bez tej zmiennej i na niej przeprowadzono kolejne testy.

Name	Role	Level
Absenteeism_time_in_hours	Input	Interval
Age	Input	Interval
Day_of_the_week	Input	Ordinal
Disciplinary_failure	Input	Binary
Education	Input	Ordinal
Height	Input	Interval
Hit_target	Input	Interval
ID_worker	Input	Nominal
Month_of_absence	Input	Ordinal
Pet	Input	Ordinal
Reason_for_absence	Input	Ordinal
Seasons	Input	Ordinal
Service_time	Input	Interval
Social_drinker	Input	Binary
Son	Input	Ordinal
Weight	Input	Interval
Work_load	Input	Interval
distance_transformed	Input	Nominal
transport_transformed	Input	Nominal
types_of_absenteeism	Target	Nominal

Rysunek 4.1: Tabela przedstawia zmienne z tabeli sasowej i ich zdefiniowane typy.

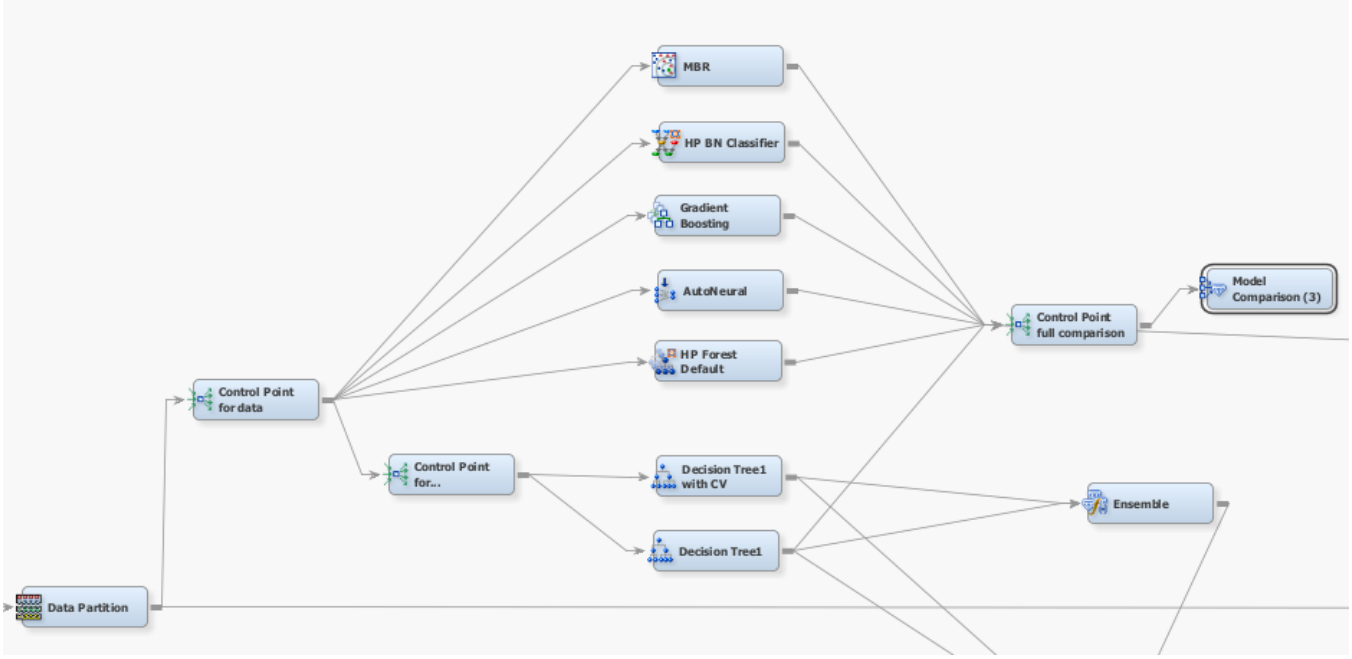
Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
Month_of_a...		5	1.0000	1.0000	1.0000
ID_worker		2	0.9815	0.7743	0.7889
Reason_for...		2	0.5331	0.0000	0.0000

Rysunek 4.2: Tabela przedstawia wpływ zmiennych przed redukacją na klasyfikację w drzewie decyzyjnym.

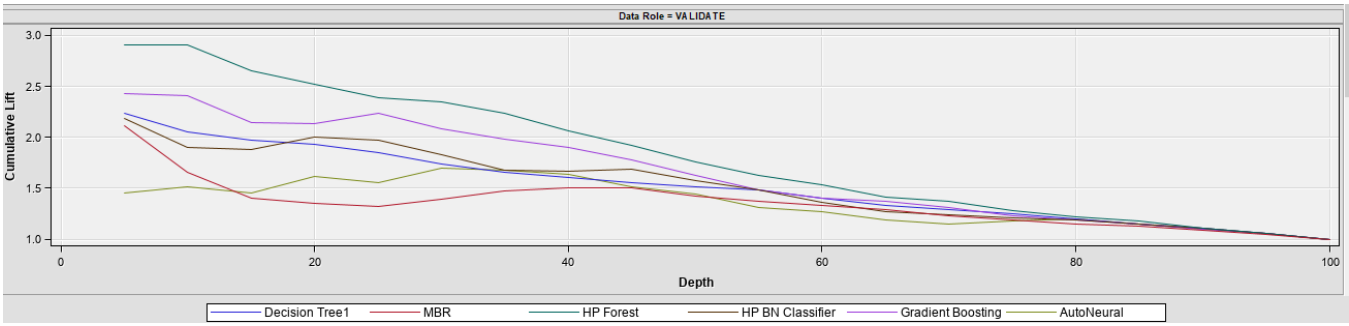
Variable Name	Number of Splitting Rules	Train: Gini Reduction	Train: Margin Reduction	OOB: Gini Reduction	OOB: Margin Reduction	Valid: Gini Reduction	Valid: Margin Reduction	Label
Age	76	0.014540	0.020766	0.00469	0.00831	0.01280	0.015978	
Month_of_a...	58	0.012178	0.015186	-0.00490	0.00502	-0.00309	0.007824	
Height	57	0.009275	0.010753	0.00309	0.00303	0.00167	0.002024	
ID_worker	54	0.011668	0.008527	-0.00993	-0.00144	-0.00728	0.000687	
Absenteeis...	51	0.009673	0.013097	0.00114	0.00301	0.00133	0.002305	

Rysunek 4.3: Tabela przedstawia wpływ zmiennych przed redukacją na klasyfikację w lasie losowym.

Następnie na nowym zbiorze przeprowadzono testowanie modelu lasu losowego (*HDMForest*), drzewa decyzyjnego z uwzględnieniem i bez uwzględnienia korswalidacji (*DecisionTree*), modelu naiwnego Bayesa (*HPBNC*), Gradient-Boosting (*Boost*), Memory-based reasoning (*MBR*) i sieci neuronowych (*AutoNeural*) z domyślnymi parametrami (co pokazuje Rysunek 4.4). Zgodnie z porównaniem modeli przeprowadzonym za pomocą węzła *Model comparison* (co przedstawiają Rysunek 4.5 oraz Rysunek 4.6) najlepszym modelem okazał się model lasu losowego. Model ten z domyślnymi parametrami jest przeuczony, więc przeprowadzono testy różnych parametrów w celu zmniejszenia przeuczenia.



Rysunek 4.4: Początkowy process flow w SAS Miner.



Rysunek 4.5: Krzywa łącznego wzrostu dla porównywanych modeli.

Fit Statistics	Statistics Label	HPDMForest	Boost	Tree4	HPBNC	MBR	AutoNeural
TKS	Test: Kolmogorov-Smir...	0.617	0.512	0.394	0.439	0.298	0.243
TASE	Test: Average Squared...	0.125533	0.144454	0.16674	0.180294	0.190526	0.273788
TAUR	Test: Roc Index	0.877	0.808	0.713	0.757	0.674	0.651
TAVERR	Test: Average Error Fu...	.	.	.	.	0.615935	1.061009
_TBINNED_KS_PRO...	Test: Bin-Based Two-...	0.383	0.393	0.403	0.393	0.25	0.747
TCAPC	Test: Cumulative Perce...	27.45098	23.52941	15.68627	19.60784	15.98793	21.56863
TCAP	Test: Percent Captured...	11.76471	11.76471	7.843137	3.921569	7.75264	9.803922
TDISF	Test: Frequency of Cla...	148	.	.	148	.	.
TDIV	Test: Divisor for ASE	444	444	444	444	444	444
TERR	Test: Error Function	.	.	.	.	273.4754	471.088
TGAIN	Test: Gain	170.8497	132.1569	54.77124	93.46405	57.74761	112.8105
TGINI	Test: Gini Coefficient	0.754	0.617	0.425	0.513	0.349	0.302
_TKS_BIN_	Test: Bin-Based Two-...	0.608	0.488	0.389	0.429	0.287	0.242
_TKS_PROB_CUTOF...	Test: Kolmogorov-Smir...	0.336	0.418	0.301	0.349	0.188	0.747
TLIFTC	Test: Cumulative Lift	2.708497	2.321569	1.547712	1.934641	1.577476	2.128105
TLIFT	Test: Lift	2.487395	2.487395	1.658263	0.829132	1.639129	2.072829
TMAX	Test: Maximum Absolut...	0.946667	0.933486	1	0.999879	0.9375	0.998841
TMISC	Test: Misclassification ...	0.25	0.27027	0.358108	0.418919	0.486486	0.581081

Rysunek 4.6: Wykres przedstawia porównanie działania rozważanych modeli na zbiorze testowym pod względem przedstawionych statystyk.

4.2 Wybór optymalnych parametrów

Przeprowadzono liczne testy dla modelu lasu losowego z różnymi parametrami. Na podstawie tych testów z wywynioskowano, że najlepsze wyniki pod względem zmniejszonego przeuczenia (pokazane na Rysunek 4.7) wystąpiły, gdy zostanie wybrany *Proportions* jako typ próbkowania, maksymalna liczba drzew i maksymalna głębokość drzewa zostaną zmniejszone, zostanie zwiększony poziom istotności oraz zostanie zmniejszona maksymalnej liczba kategorii w teście asocjacji. Optymlane parametry znalezione za pomocą modelu *HP Forest(4)* są przedstawione na Rysunek 4.8, zaś statystyki charakteryzujące model z tymi parametrami zostały pokazane na Rysunek 4.9.

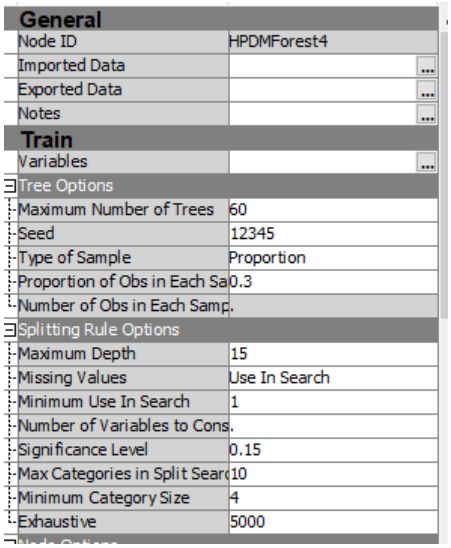
Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
HP Forest with Count	types_of_a...		0.217195
HP Forest (4)	types_of_a...		0.257919
HP Forest (5)	types_of_a...		0.294118

(a) Statystyka dotycząca zbioru walidacyjnego dla modeli lasu losowego z różnymi paramterami.

Train: Misclassification Rate
0.013587
0.100543
0.105978

(b) Statystyka dotycząca zbioru treningowego.

Rysunek 4.7: Powyższe tabele pokazują przeuczenie testowanych modelów, które wynika z dużej różnicy między statystykami dla zbioru treningowego i walidacyjnego.



General	
Node ID	HPDMForest4
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Tree Options	
Maximum Number of Trees	60
Seed	12345
Type of Sample	Proportion
Proportion of Obs in Each Sa	0.3
Number of Obs in Each Samp	.
Splitting Rule Options	
Maximum Depth	15
Missing Values	Use In Search
Minimum Use In Search	1
Number of Variables to Cons	.
Significance Level	0.15
Max Categories in Split Sear	10
Minimum Category Size	4
Exhaustive	5000

Rysunek 4.8: Tabela pokazuje znalezione optymalne parametry.

Statistics Label	Validation	Train	Test
Average Squared Error	0.128219	0.083132	0.139785
Divisor for ASE	663	1104	444
Maximum Absolute Error	0.895833	0.802698	0.9175
Sum of Frequencies	221	368	148
Root Average Squared Error	0.358077	0.288327	0.373878
Sum of Squared Errors	85.00914	91.77808	62.06452
Frequency of Classified Cases	221	368	148
Misclassification Rate	0.257919	0.100543	0.283784
Number of Wrong Classifications	57	37	42

Rysunek 4.9: Tbaela przedstawia statystyki modelu z optymalnymi parametrami.

Rozdział 5

Podsumowanie i wnioski

Podsumowując, zbiór danych *Absenteeism in work* nawet w wersji nieprzekształconej jest dobrze uzupełnionym zbiorem danych nadającym się do analizy i modelowania. Błędne dane obejmowały jedynie kilka rekordów, co nie wpłynęłoby w większym stopniu na wyniki. Krokiem, który zdecydowania ułatwia analizę zbioru, jest zamiana niektórych zmiennych numerycznych na dane kategoryczne, co zostało wykonane w przypadku kolumn *Transportation expense* i *Distance from Residence to work*.
Praca nad projektem w programie SAS Enterprise Guide okazała się dosyć czasochłonna. Ciągłe klikanie w *Query Builder*, by zmodyfikować zbiór, było męczące w szczególności, gdy osoba pracująca nad projektem zna język zapytań SQL i potrafi zauważyć, że każdą z modyfikacji można było zapisać w postaci paru linijek kodu, co zaoszczędziłoby trochę czasu.
Ponadto, trudnością okazało się zdefiniowanie podejścia analitycznego w zakresie uczenia maszynowego do projektu. Modele regresyjne, które mogą być stworzone dla zbioru (np. model przewidujący liczbę nieobecności w miesiącu) wydawały się mało uniwersalne i interesujące, dlatego zdecydowano się rozważyć problem klasyfikacji dla tych danych. Zauważono, że w takim przypadku każdy analityk, który będzie używać modelu, może sam zmienić granice dla typów nieobecności. Dodatkowo, informacja o typie nieobecności pracownika jest łatwiejsza do zinterepretowania dla innych osób, które nie mają wiedzy w zakresie analizy danych, niż gdyby model zwracał średnią liczbę godzin nieobecnych w miesiącu dla danego pracownika.
Ponadto, znalezienie dobrego modelu klasyfikujacego okazało sie wymagające ze względu na problem przeuczenia, który dotyczył wszystkich rozważanych modeli. Próba rozwiązania tego problemu wymagała czasochłonnego testowania różnych parametrów.
Praca w programie SAS Miner była przyjemnym doświadczeniem. Ze względu na liczne funkcje w postaci kafelków w prosty sposób może testować modele nawet osoba, która nie ma dużej wiedzy na ich temat.

Bibliografia

[1] SAS documentation - dokumentacja w programie SAS Miner

[2] `https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work` - źródło zbioru danych