Technická univerzita v Liberci

Ekonomická fakulta

Semestrální projekt

Predikce akademické úspěšnosti

Jméno studenta: Pavel Švec

Ročník: 1.

Akademický rok: 2023/2024

Datum vypracování: 16.5.2024

Obsah

1 Úvod

V rámci řešení úlohy pro nasazení modelů je použit dataset pocházející z výzkumné studie v Portugalsku *Using Data Mining to Predict High School Student Performance.* Cílem průzkumu je sestavení vhodných modelů a jejich porovnání, následné nasazení a průzkum hlavních prediktorů pro nadměrnou konzumaci alkoholu studentů během výukového týdne.

Poor	0-3,4
Weak	3,5-9,4
Sufficient	9,5-13,4
Good	13,5-15,4
Very good	15,5-17,4
Excellent	17,4-20

Obrázek 1 Akademické známkování v Portugalsku

Pro potřeby tvorby modelu budeme tvořit proměnnou představující úspěšné studium, což překračuje hranici alespoň 10 bodů. Tato bodová hranice představuje 50 % a odpovídá tak stupnici E dle ECTS. Zdrojová data jsou dostupná zde:

https://www.kaggle.com/datasets/gabrielluizone/high-school-alcoholism-and-academic-performance?rvi=1

Díky již proběhlé přípravě dat není nutné data nijak dále upravovat. Je tak rovnou možné začít postupně data rozdělit dle typů, stanovit cílovou proměnnou a vyzkoumat vhodné modely k nasazení v závěrečné části práce.

2 Analýza dat a tvorba modelu

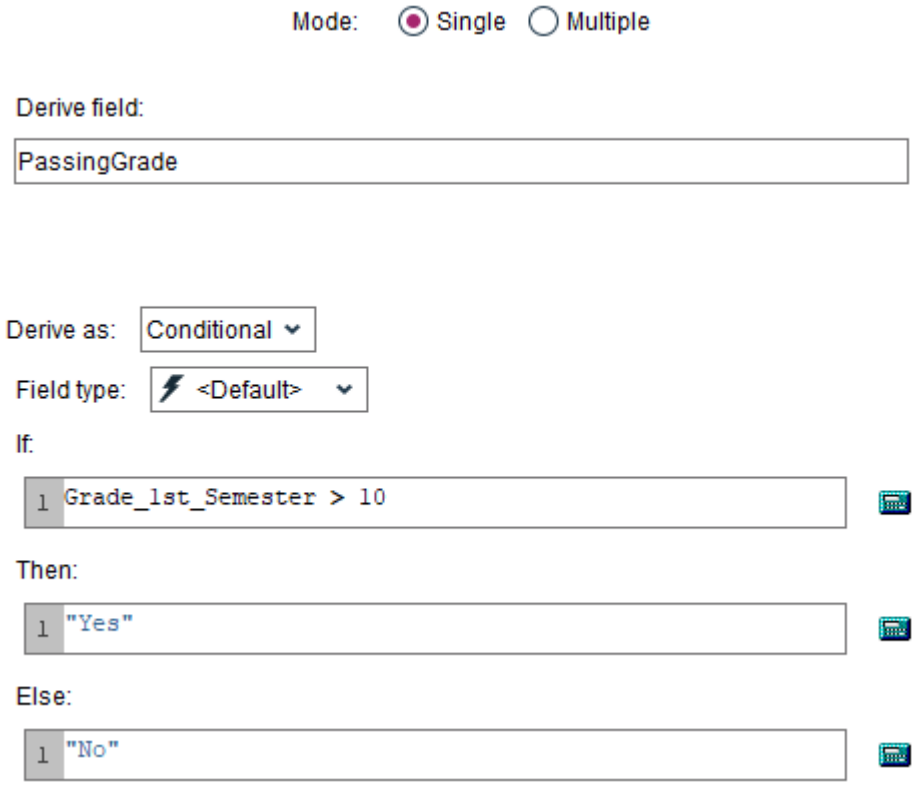
K analýze dat je použit program IBM SPSS Modeler. Nejprve je v rámci programu nastaven datový soubor přeložený do angličtiny pomocí uzlu *Var. File*. Oddělovače proměnných jsou čárky a každý nový řádek symbolizuje jiného žáka a jeho údaje.

Field	Measurement	Values	Missing	Check	Role
School	Flag	"Mousinho da...		None	Input
Gender	Flag	Male/Female		None	Input
Age	Continuous	[15,22]		None	Input
Housing_Type	Flag	Urban/Rural		None	Input
Family_Size	Flag	"Up to 3"/"Abo...		None	Input
Parental_Status	Flag	Separated/"Li...		None	Input
Mother_Education	Nominal	"High School"...		None	Input
Father_Education	Nominal	"High School"...		None	Input
Mother_Work	Nominal	Health,Home...		None	Input
Father_Work	Nominal	Health,Home...		None	Input
Reason_School...	Nominal	"Course Pref...		None	Input
Legal_Respons...	Nominal	Father,Mother...		None	Input
Commute_Time	Nominal	"15 to 30 min"...		None	Input
Weekly_Study_T...	Nominal	"2 to 5h","5 to ...		None	Input
Extra_Education...	Flag	Yes/No		None	Input
Parental_Educat...	Flag	Yes/No		None	Input
Private_Tutoring	Flag	Yes/No		None	Input
Extracurricular_...	Flag	Yes/No		None	Input
Attended_Daycare	Flag	Yes/No		None	Input
Desire_Graduat...	Flag	Yes/No		None	Input
Has_Internet	Flag	Yes/No		None	Input
Is_Dating	Flag	Yes/No		None	Input
Good_Family_R...	Nominal	Excellent,Fair,...		None	Input
Free_Time_After...	Nominal	High,Low,Mo...		None	Input
Time_with_Frie...	Nominal	High,Low,Mo...		None	Input
Alcohol_Weekd...	Nominal	High,Low,Mo...		None	Input
Alcohol_Weeke...	Nominal	High,Low,Mo...		None	Input
Health_Status	Nominal	Fair,Good,Po...		None	Input
School_Absence	Continuous	[0,32]		None	Input
Grade_1st_Sem...	Continuous	[0,19]		None	Input
Grade_2nd_Se...	Continuous	[0,19]		None	None

Obrázek 2 Uzel type – nastavené typy proměnných v souboru

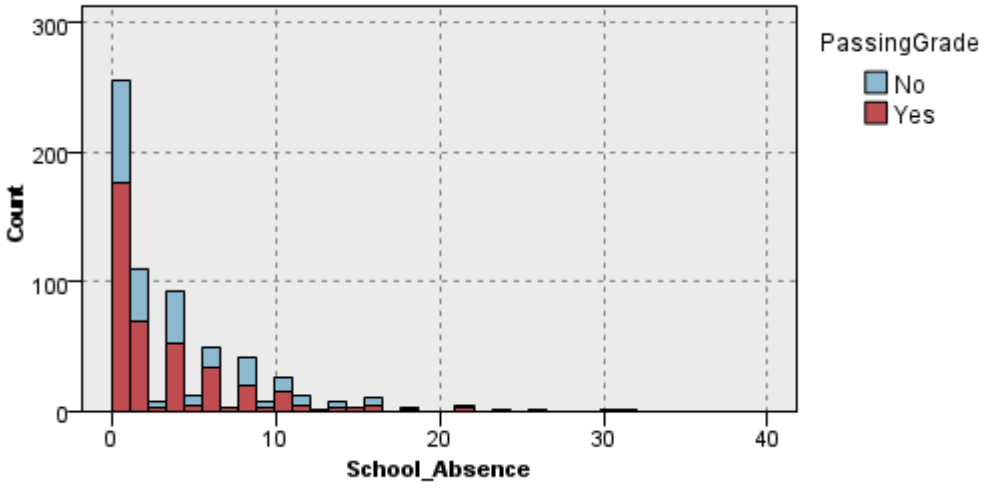
V rámci výzkumu jsem se rozhodl zaměřit se více na známky z prvního semestru a na jejich základě vytvořit modely. Většina studentů vyhledává úspěšnost právě v prvních semestrech studia, protože v těch dalších již mají lepší představu co očekávat, a i škola dokáže lépe predikovat počet úspěšných. Cílem je ale stanovit pouze zdali student dosáhne výsledku vyššího 50 % či nikoliv. Proto je nutné vytvořit novou proměnnou, která bude sloužit jako cílová, a to pomocí uzlu *Derive*. Nová proměnná bude nabývat následujících hodnot:

- *Yes* – ano, známka přesahuje 50 %,
- *No* – ne, známka není dostatečná.



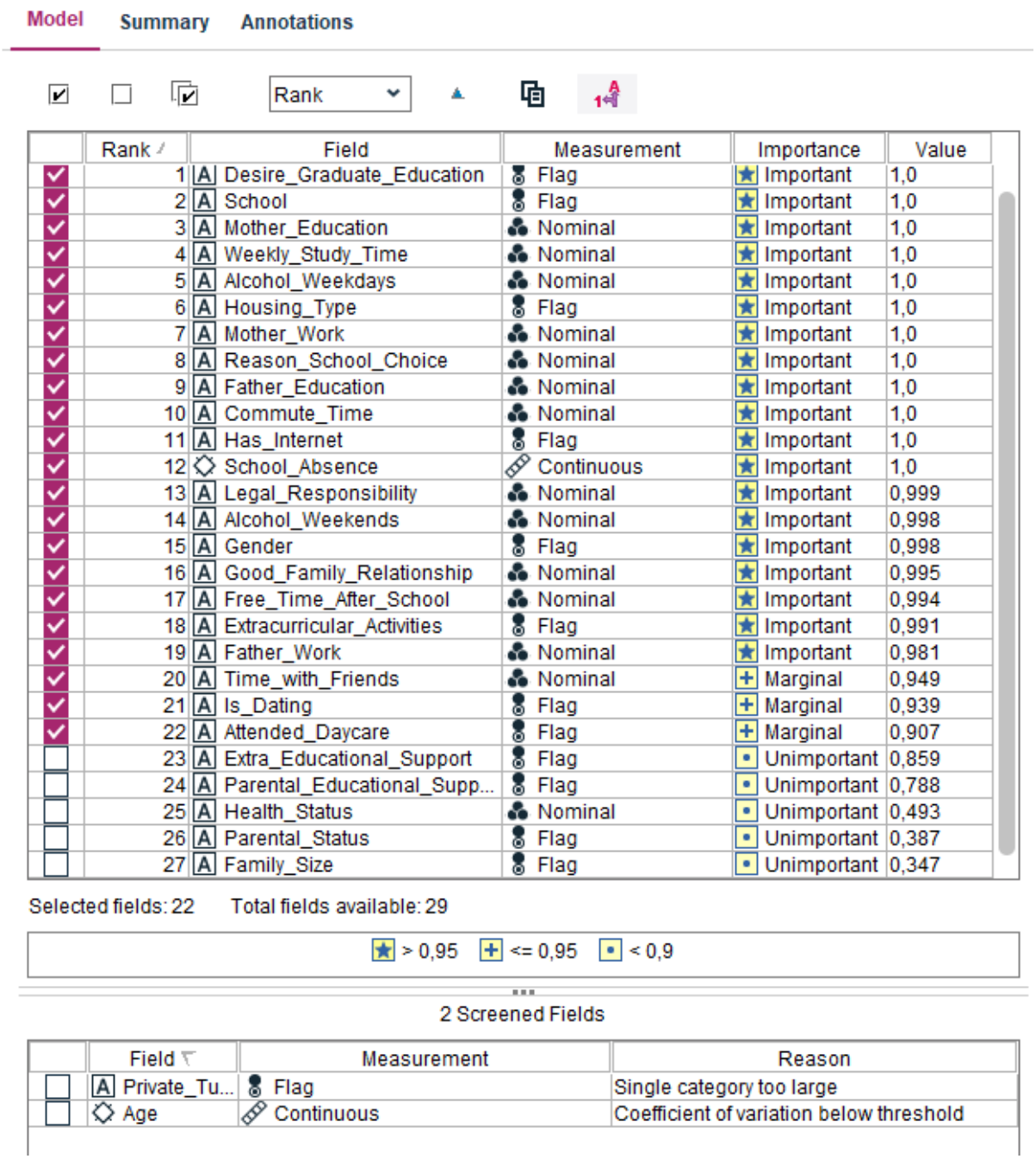
Obrázek 3 Tvorba nové cílové proměnné „PassingGrade“

Celkový soubor tak nyní obsahuje 30 proměnných po vyřazení *Grade_1st_Semester* a *Grade_2nd_Semester*. Na základě uzlu Data audit, je možné zjistit, že soubor je poměrně kvalitní. Jediné proměnné, kde se vyskytují odlehlá data jsou *Age* a *School_Absence*.



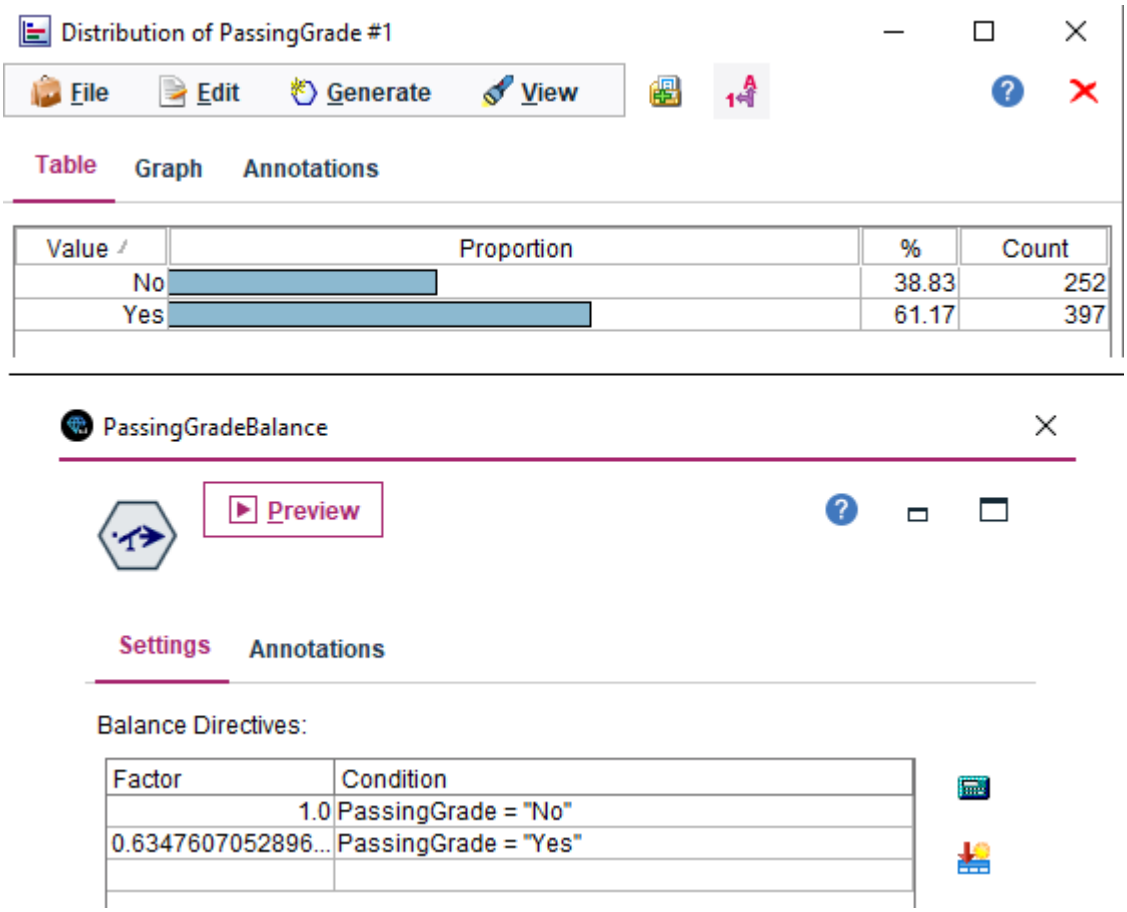
Obrázek 4 Histogram školní absence

Tyto odlehlé hodnoty jsem se rozhodl v souboru ponechat vzhledem k jejich minimální četnosti. Náhled do data auditu je dostupný ve Streamu přiloženém k této práci.



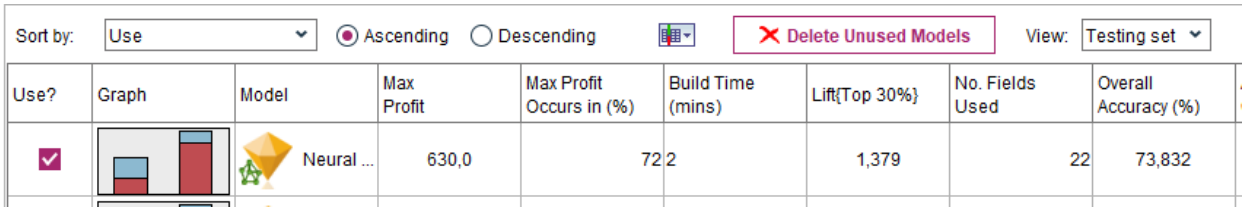
Obrázek 5 Uzel Feature selection – Důležitost proměnných

Na základě modelu pro určení důležitosti proměnných pro sestavení modelu bylo vyřazených několik proměnných. Byla také vyřazena proměnná *Age,* čímž se eliminovali s tím související odlehlé hodnoty. Dále je učiněno rozhodnutí, které rozděluje úlohu na 2 řešení, a to s pomocí vybalancování cílové proměnné a bez balancování cílové proměnné. Na základě toho tak vzniknou dva vhodné modely, které bude možné porovnat.

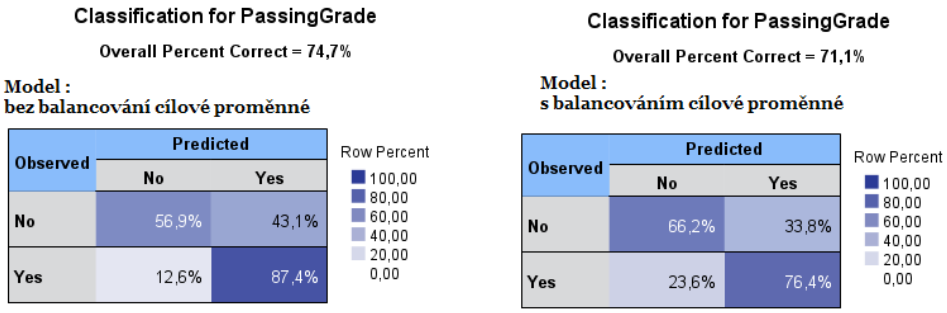


Obrázek 6 Původní struktura cílové proměnné a její vybalancování

Vybalancování cílové proměnné proběhlo pomocí metody zredukovaní počtu *Yes* dle faktoru ca. 0,63. Stream byl tak rozdělen do dvou částí. Dále je pro obě části použit uzel *Partition,* kde jsou stanovena data na testování a trénování modelů. Jejich poměr je 1:1 a tvoří 50 % a 50 % ze zbývajících dat. Pomocí uzlu *Auto Classifier* bylo vybráno několik vhodných modelů. U obou řešení vychází jako nejlepší model neuronová síť na základě výkonu model, nejvyšší hodnota *Lift* a poměrně vysoká přesnost. Mezi další poměrně úspěšné modely se řadí logistická regrese, LSVM, CHAID nebo C5 1.

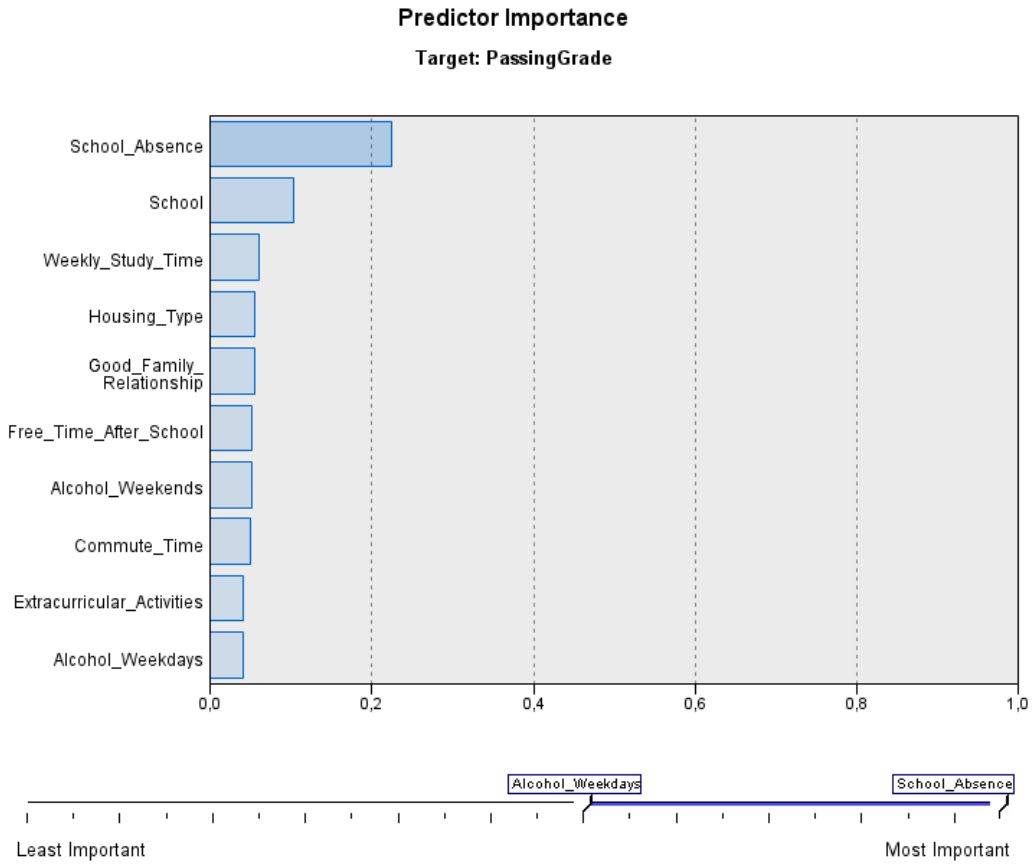


Obrázek 7 Auto classifier – náhled na nejlepší vybraný model

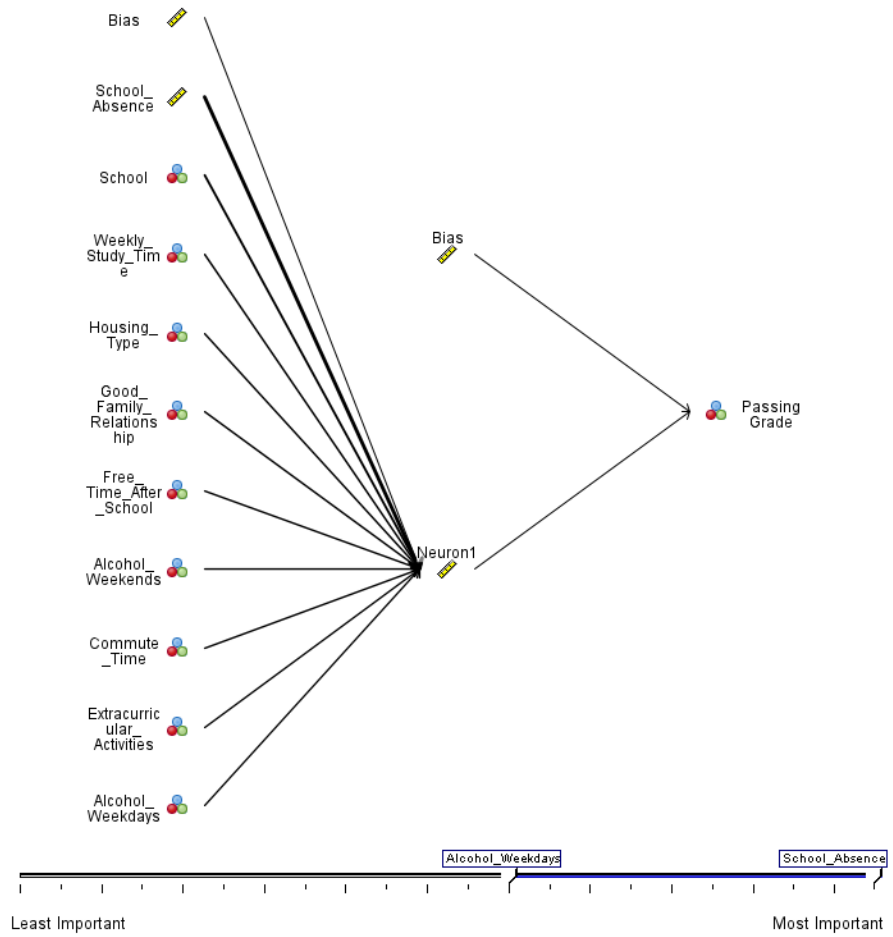


Obrázek 8 Porovnání matic chyb I. a II. řádu modelů s balancováním (vpravo) a bez balancovaní (vlevo) neuronových sítí

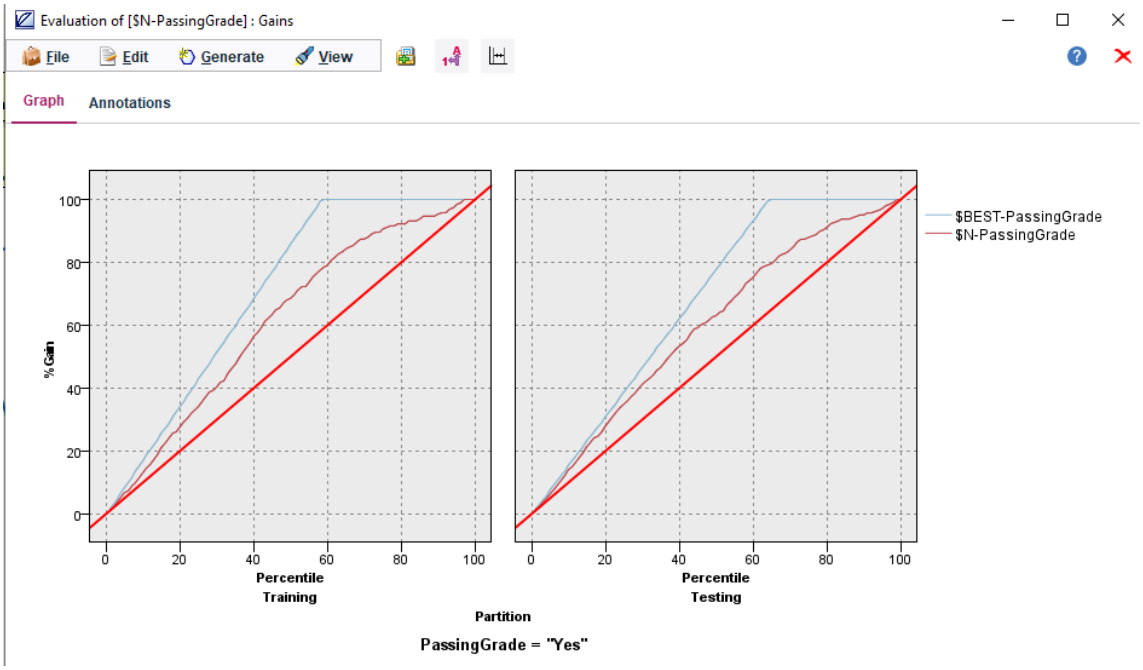
V rámci tvorby modelů tak došlo ke tvorbě dvou řešení. Model, který nebyl vybalancovaný má méně falešných positiv, ale model, který byl vybalancovaný má méně falešných negativ. Prospěšné studenty dokáže lépe identifikovat model bez balancování. Celkově vychází jako přesnější model neuronové sítě bez vybalancování na základě přesnosti (74,7 %). Je tak použit jako model pro nasazení a je dále interpretován.



Obrázek 9 Důležité prediktory v rámci modelu



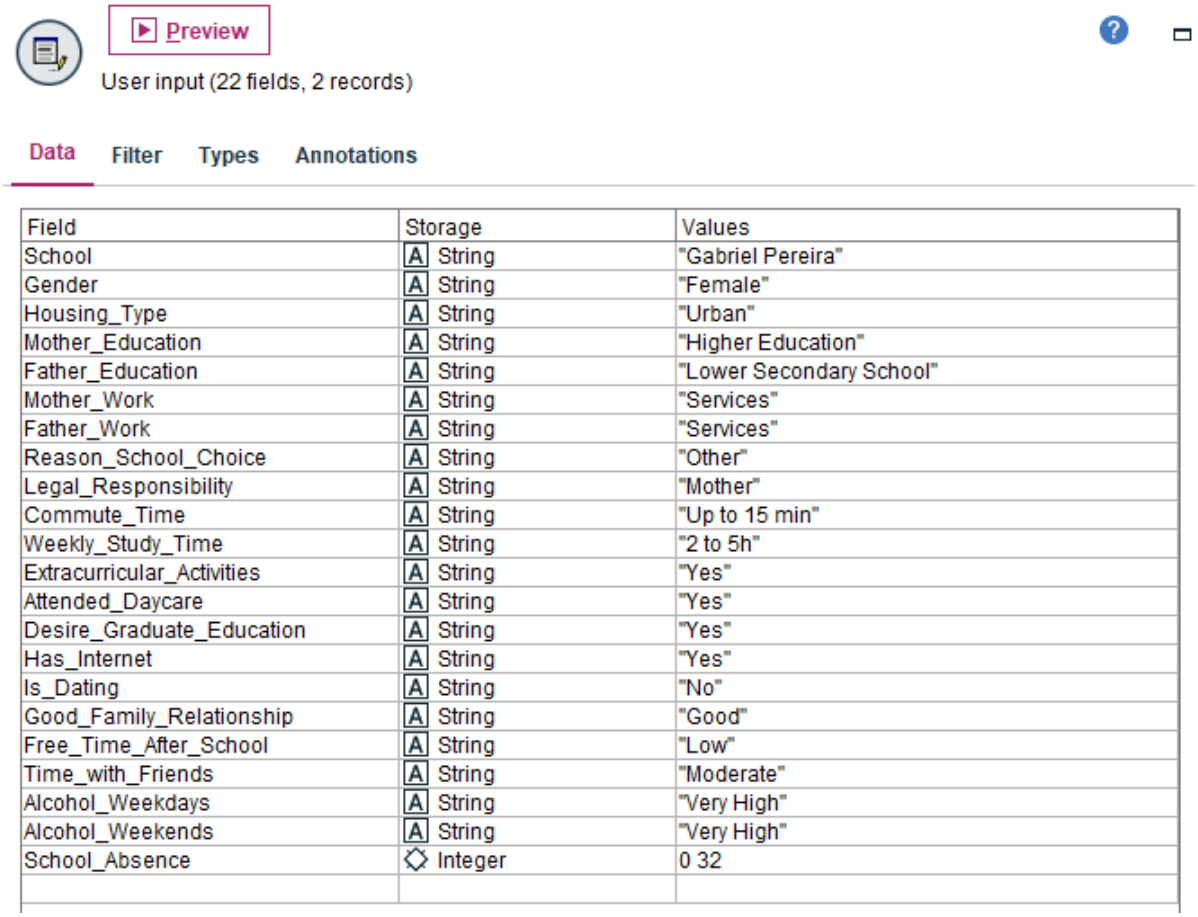
Obrázek 10 Struktura neuronové sítě



Obrázek 11 Evaluace modelu neuronové sítě

3 Nasazení modelu

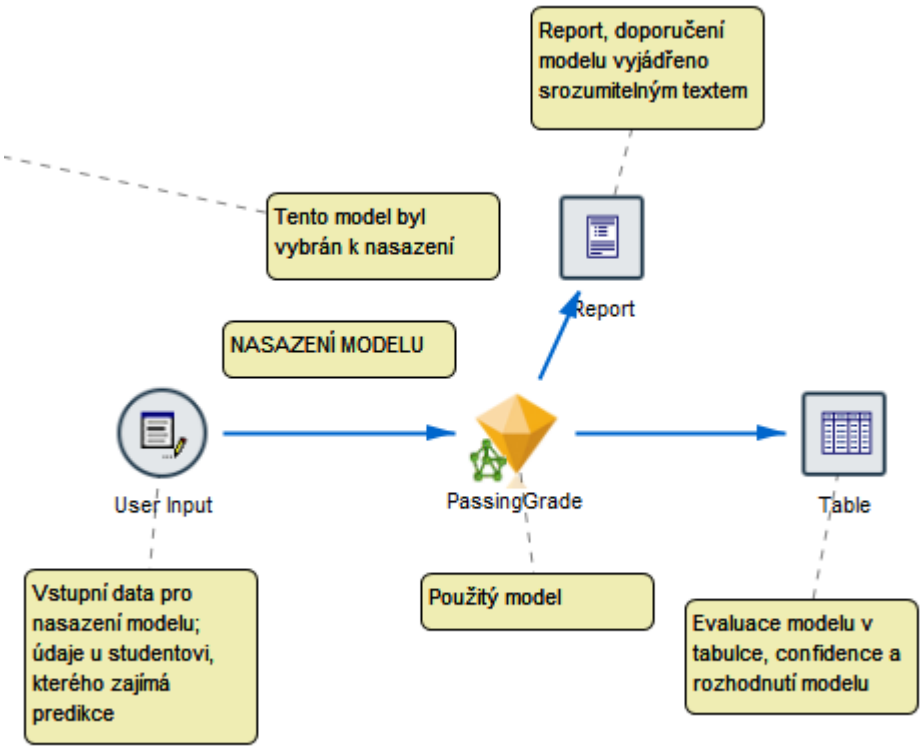
K nasazení modelu bude použit uzel *User input* a model bez vybalancování cílové proměnné. Je nutné ponechat na paměti, že je model nutné přizpůsobit pro potřeby organizace v rámci změn a model dále trénovat a upravovat na nových datech.



Field	Storage	Values
School	String	"Gabriel Pereira"
Gender	String	"Female"
Housing_Type	String	"Urban"
Mother_Education	String	"Higher Education"
Father_Education	String	"Lower Secondary School"
Mother_Work	String	"Services"
Father_Work	String	"Services"
Reason_School_Choice	String	"Other"
Legal_Responsibility	String	"Mother"
Commute_Time	String	"Up to 15 min"
Weekly_Study_Time	String	"2 to 5h"
Extracurricular_Activities	String	"Yes"
Attended_Daycare	String	"Yes"
Desire_Graduate_Education	String	"Yes"
Has_Internet	String	"Yes"
Is_Dating	String	"No"
Good_Family_Relationship	String	"Good"
Free_Time_After_School	String	"Low"
Time_with_Friends	String	"Moderate"
Alcohol_Weekdays	String	"Very High"
Alcohol_Weekends	String	"Very High"
School_Absence	Integer	0 32

Obrázek 12 Ukázka vstupních dat pro evaluace

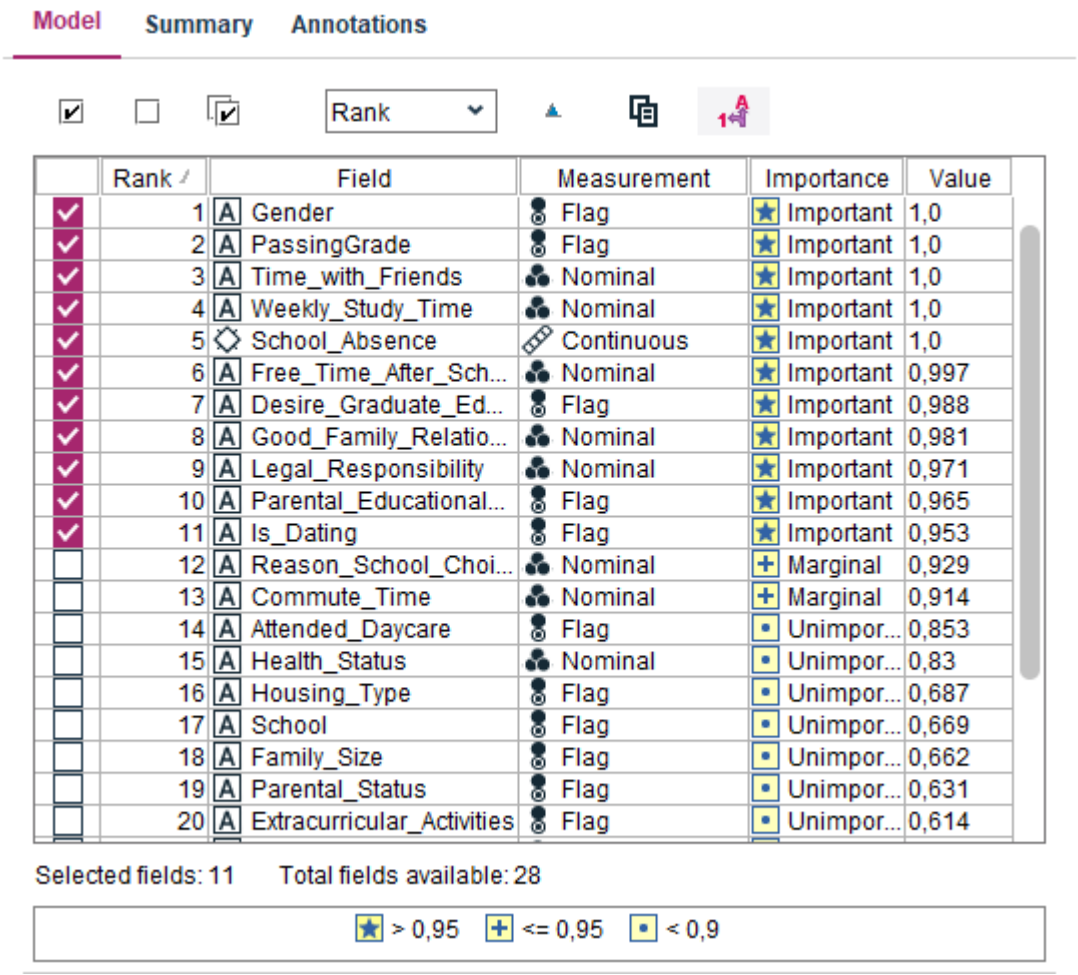
V následujícím obrázku je vidět kompletní zapojení evaluačního streamu, kdy z uzlu *User input* je na základě vstupních 22 datových údajů vytvořena predikce modelu, zdali je žák je úspěšným studentem, či nikoliv (úspěšnost >50 %). Pomocí uzlu *Table* je možné nahlédnout na přidané proměnné *$N-PassingGrade* (predikce modelu *Yes* nebo *No*) a *$NC-PassingGrade* (výjádření jistoty předpokladu modelem). Po nastavení vysoké školní absence, vysoké konzumace alkoholu, nízkého studijního času, dlouhého dojezdu do školy, méně kvalitní původní základní školy a špatných rodinných vztahů je velmi snadné docílit predikce *No* (úspěšnost <50 %).



Obrázek 12 Nasazení modelu

4 Dodatečná úloha – predikce alkoholismu

Na závěr předpokládejme teoretickou situaci – škola se snaží zamezit konzumaci alkoholu a snaží se odchytit žáky s nadbytečnou konzumací. Je nutné vytvořit model, který by umožnil vynaložit méně sil pro identifikaci co největšího množství žáku bez oslovení celé školy osobně. Cílovou proměnnou se tak stane *Alcohol_Weekdays*, protože konzumace alkoholu během týdne má vyšší dopad než konzumace o víkendech. Z modelu je vyřazena proměnná *Alcohol_Weekends*.



Obrázek 13 Vybrané proměnné pro tvorbu modelu; důležitost – uzel Feature selection

Po vybrání vhodných proměnných jsou opět rozdělena data na testovací a trénovací: poměr je 1:1. Na základě vyzkoušení několika modelů se jako vhodné jeví neurální síť, C&R Tree, Tree-AS, Quest. Vzhledem k hojnému využití modelu neurální sítě v této semestrální práci byl vybrán opět tento model.

Classification for Alcohol_Weekdays

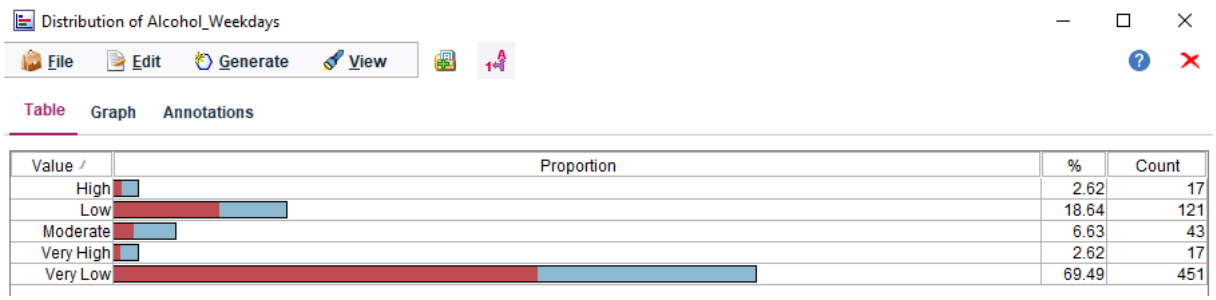
Overall Percent Correct = 68,9%

Observed	Predicted				
	High	Low	Moderate	Very High	Very Low
High	0,0%	0,0%	0,0%	0,0%	100,0%
Low	0,0%	3,2%	0,0%	3,2%	93,7%
Moderate	4,8%	0,0%	0,0%	9,5%	85,7%
Very High	0,0%	25,0%	0,0%	16,7%	58,3%
Very Low	0,0%	0,4%	0,0%	0,4%	99,1%

Row Percent
100,00
80,00
60,00
40,00
20,00
0,00

Obrázek 13 Výsledek klasifikace pro cílovou proměnnou Alcohol_Weekdays

Již z předchozí analýzy dat bylo zřejmé, že naprostá většina studentů není častými konzumenty alkoholu. Avšak i tak byl model schopný identifikovat alespoň nějaké konzumenty s velkou konzumací alkoholu (*Very high*). Model není tak přesný a je lepší model nenasazovat a najít více studentů s problémy s alkoholem a na základě toho tvořit nový model.



Obrázek 14 Problém s identifikací alkoholiků – nedostatečné množství dat

Závěr

Podařilo se sestrojit model, který je možné použít pro predikci úspěšnosti studenta ve studiu na základě dílčích předpokladů. Modely, které se osvědčily byly neuronové sítě. Pro identifikaci alkoholismu během týdne se však v souboru nenachází dostatek dat, a tak by bylo nutné pro konstrukci modelu pro takovou identifikaci získat větší objemy dat, a to konkrétně na cílované alkoholiky s hodnotou *High* (vysoká konzumace) a *Very High* (velmi vysoká konzumace).

Zdroje

P. Cortez e A. Silva. *Usando a Mineração de Dados para Prever o Desempenho do Aluno do Ensino Médio*. Em A. Brito e J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008, Porto, Portugalsko, 2008, EUROSIS, ISBN 978-9077381-39-7.

Luiz, Gabriel. High School Alcoholism and Academic Performance. Kaggle. Dostupné z: https://www.kaggle.com/datasets/gabrielluizone/high-school-alcoholism-and-academic-performance?rvi=1

Příloha

Příloha 1 Náhled na celý stream