



**ZZN - Projekt 2022/2023**

# **Data o obchodování na burze**

## **Řešení**

Autoři: Bc. Ondřej Studnička (xstudn00), Bc. Jan Lorenc (xloren15)  
Datum: 14. listopadu 2022

# **Obsah**

<b>1</b>	<b>Zadání</b>	<b>2</b>
<b>2</b>	<b>Analýza dat</b>	<b>2</b>
<b>3</b>	<b>Predikce vývoje ceny indexů na burze</b>	<b>3</b>
<b>4</b>	<b>Detekce výkyvů cen indexů</b>	<b>4</b>
<b>5</b>	<b>Analýza podobnosti indexů</b>	<b>5</b>
	<b>Závěr</b>	<b>6</b>
	<b>Příloha A</b>	<b>7</b>
	<b>Příloha B</b>	<b>9</b>
	<b>Příloha C</b>	<b>11</b>

# 1 Zadání

Cílem projektu je analyzovat zadaná data o obchodování na burze, vymyslet 3 dolovací úlohy a po schválení je implementovat v prostředí RapidMiner. Úlohy k řešení jsou následující:

## 1. *Predikce vývoje ceny indexů na burze*

Obchodník zajímá, jak se bude cena indexu vyvíjet. Podle této informace může provádět obchody.

**Formulace:** Pro každý index provedte regresní analýzu funkční křivky historických dat s cílem predikce hodnoty funkce pro neznámé vstupní hodnoty (budoucnost).

**Přínos:** Obchodník tak může nakupovat podíly indexů, které mají růst a prodávat ty, které mají klesat.

## 2. *Detekce výkyv cen indexů*

Pro obchodníka je velice důležité poznat, kdy se index nachází v tzv. dipu nebo tzv. peaku. Je to pro něj indikátor přicházející potenciální korekce. Výkyvy v cenách indexů jsou však běžné a je těžké zjistit, jestli se jedná o dip/peak nebo jen standardní výkyv.

**Formulace:** Pro každý index provedte detekci anomalií ve funkční křivce historických dat s cílem odhalení zajímavých vývojů v cenách.

**Přínos:** Nacházíme-li se aktuálně podle historických dat v anomálii, lze učinit rozhodnutí nákupu podílu ve slevě (v dipu, pokud je cena nižší, než je běžné) nebo prodeje či shortování (v peaku, kdy je cena vyšší než obvykle). Zároveň mám všeobecný přehled o volatilitě indexu.

## 3. *Analýza podobnosti indexů*

Z pohledu diverzifikace není dobré kupovat podíly indexů, které rostou/klesají zároveň. Korelace mezi indexy bývá často dána globální ekonomickou situací či stejných akcií zahrnutých v různých indexech.

**Formulace:** Proveďte shlukovou či korelační analýzu nad vývojem cen indexů s cílem zjistit podobnost.

**Přínos:** Namísto investování do 2 indexů, které však roustou/klesají stejně, obchodník může diverzifikovat a místo druhého indexu koupit jiný, zcela nezávislý.

# 2 Analýza dat

Data popisují denní vývoj cen indexů na burzách z celého světa. Údaje jsou převzaty ze severu Yahoo Finance, který je již několik desítek let poskytuje. Ceny jsou uvedeny v méně státu, ve kterém se s indexem obchoduje.

Dataset obsahuje celkem 3 soubory. V indexInfo.csv lze nalézt seznam indexů v datasetu s dodatečnými informacemi, jakými jsou region, burza a měna. Soubory indexData.csv a indexProcessed.csv obsahují stejná data pouze s tím rozdílem, že z indexProcessed.csv byly odstraněny null hodnoty a má navíc sloupec s uzavírající cenou v dolarech bez ohledu na původní měnu.

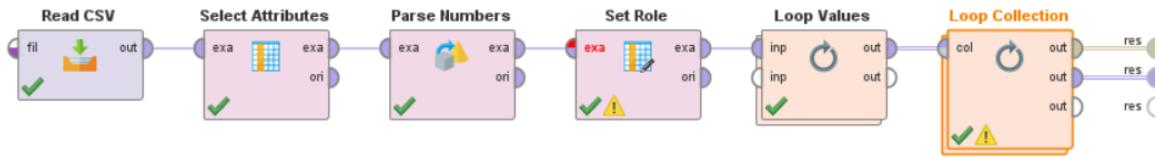
Pracovat se tedy bude primárně se souborem indexProcessed.csv. Data obsahují následující sloupce:

- Index – zkratka indexu, pro který je daný údaj
- Date – datum uvedeného obchodního dne
- Open – cena podílu indexu při otevření burzy
- High – nejvyšší cena podílu indexu za daný den
- Low – nejnižší cena podílu indexu za daný den
- Close – cena podílu indexu při uzavření burzy
- Adj Close – upravená cena podílu indexu po uzavření burzy na základě nějaké skutečnosti
- Volume – množství obchodovaných podílů indexu v daný den
- CloseUSD – cena podílu indexu při uzavření burzy v dolarech

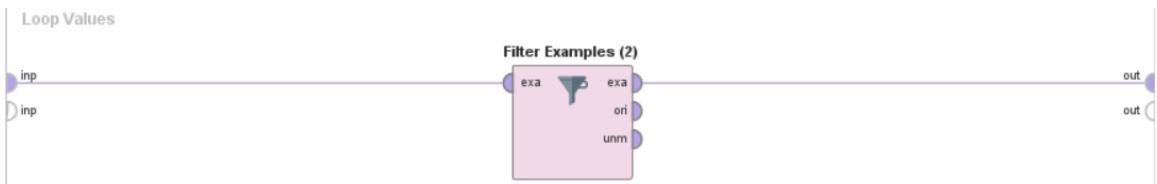
### 3 Predikce vývoje ceny indexů na burze

Pro tuto úlohu není potřeba mnoho atributů z datové sady. Hlavním důvodem je korelace atributů Open, High, Low, Close, Adj Close, CloseUSD a taky to, že prakticky Close z jednoho dne je Open následujícího dne (s případnou mírnou změnou z aftermarket hodin). Pro časový horizont několika desetiletí je toto zcela bezpředmětné. Vybrány tedy byly atributy index, neboť chceme predikovat vývoj ceny pro každý index zvlášť, Date, který je v podstatě jediný parametr pro predikci a CloseUSD, jenž udává cílové hodnoty. Za cílovou hodnotu mohl být vybrán libovolný z atributů Open, High, Low, Close, Adj Close, CloseUSD, nicméně tento jsme zvolili pro jednotu měny napříč indexy.

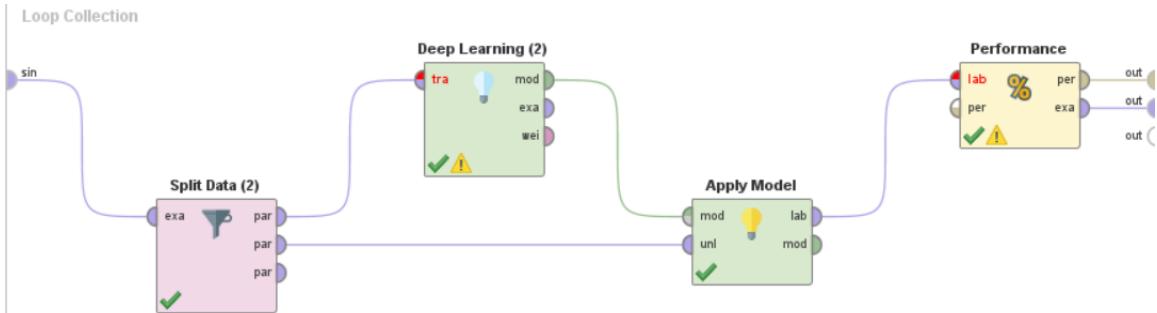
Jelikož predikce je regresní úloha, jako první nabízela se řešení pomocí lineární či polynomální regrese. Tyto metody jsou však příliš triviální a velmi nepřesné, nehledě na to, že u polynomální regrese lze jen velmi těžko najít vhodný rád, který by generalizoval různá data. Zvolena tedy byla třívrstvá neuronová síť s ReLU aktivacemi a kvadratickou chybovou funkcí. Na následujících obrázcích 1, 2, 3 je popisán výsledný model.



Obrázek 1: První blok pouze načítá soubor *indexProcessed.csv*. V následujícím bloku jsou vybrány atributy index, Date a CloseUSD. Poté dochází k přeměně datumu, který má typ string ve formátu yyyy-MM-dd, na číselný atribut. Další blok nastaví roli atributu CloseUSD na "label", což značí cílovou hodnotu, de facto Y, a za X je pak použité číselné datum. Blok Loop Values popisuje příští obrázek 2, ale v podstatě jen rozštěpí dataset na několik menších pro každý index. Poslední blok provádí trénování modelu pro každý index a na výstup dává testovací dataset s predikcemi a přesnost modelu.



Obrázek 2: Subproces bloku Loop Values pouze filtruje původní dataset dle indexu a vytvoří menší datasety pro každý index zvlášť.



Obrázek 3: Uvnitř iterace pro dataset daného indexu dochází k následujícímu. V prvním bloku se náhodně rozdělí data na trénovací a testovací v poměru 80:20. Trénovací dataset jde na vstup bloku s neuronovou sítí, jehož výstupní model společně s testovacími daty putuje do bloku aplikující model na daná data. Výstupní dataset s predikcemi je ještě hodnocen výkonostním blokem, měřícím chybu RMSE.

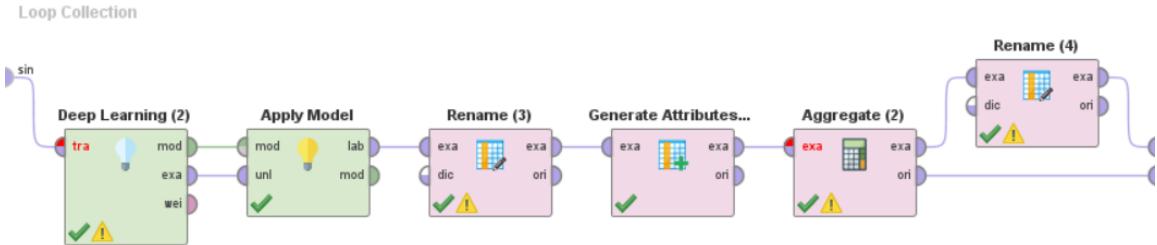
Následující tabulka 1 zobrazuje RMSE chybu pro každý index a v příloze A lze nalézt grafy s korektními a predikovanými hodnotami testovacích dat.

Index	RMSE
000001.SS	81.287
399001.SZ	418.763
GDAXI	1181.282
GSPTSE	942.762
HSI	357.418
IXIC	877.052
J203.JO	213.491
N100	139.684
N225	17.950
NSEI	9.694
NYA	758.582
SSMI	1122.687
TWII	41.704

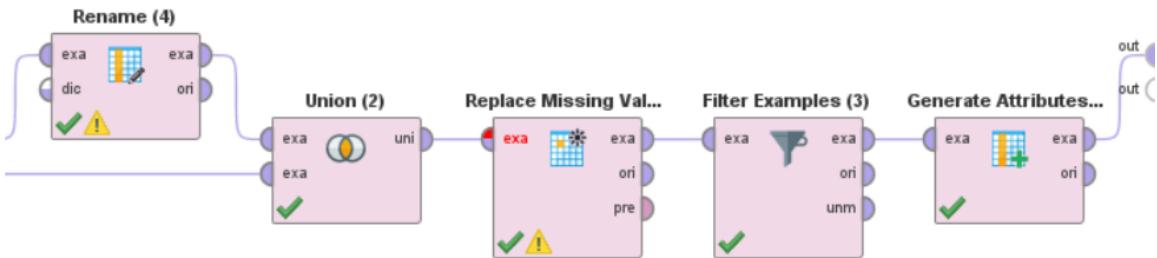
Tabulka 1: Výsledky RMSE pro jednotlivé indexy

## 4 Detekce výkyv cen indexů

V této úloze jsme zvolili stejná data jako u předchozí úlohy predikce 3. Z toho důvodu i příprava dat a práce s atributy vypadá stejně a tedy hlavní proces také odpovídá obrázku 1. Odlišnost spočívá až ve zpracování datasetů pro jednotlivé indexy uvnitř bloku Loop Collection, který popisují diagramy 4, 5.



Obrázek 4: Diagram popisuje první polovinu procesu detekce odlehlých hodnoty pro daný index.



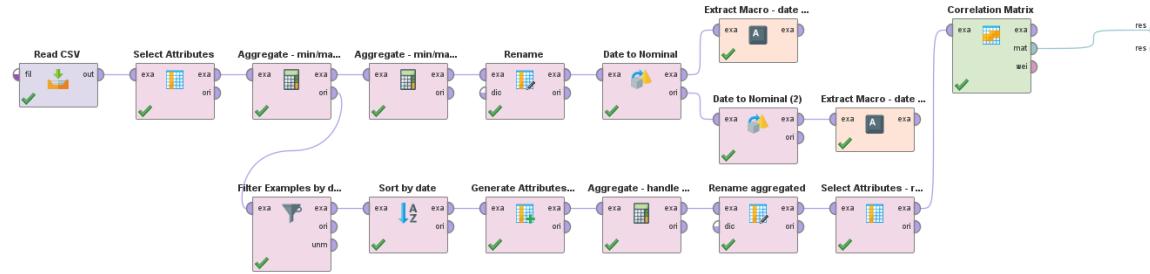
Obrázek 5: Diagram popisuje druhou polovinu procesu detekce odlehlých hodnoty pro daný index.

Nejprve jsou všechna data approximována 6-vrstvou regresní neuronovou sítí s ReLU aktivacemi a kvadratickou chybou. Ta slouží k podobnému účelu jako síť z úlohy predikce 3, nicméně zde je cílem přesnější approximace, a proto byla zvolena síť hlubší. Approximovaná hodnota je dále použita jako

ground-truth pro detekci outlier. V bloku Rename je atribut ”prediction(CloseUSD)” vytvořené modelem na predictionCloseUSD, neboť závorky se nelší výrazům v dalších výpočtech. Následující blok Generate Attributes vytvoří atribut Variance s hodnotami odchylek od reálných a predikovaných CloseUSD hodnot. Jelikož v Generate Attributes není k dispozici funkce pro výpočet mediánu, následující blok Aggregate počítá medián odchylek a blok Rename tento výsledek přejmenovává na MAD (Median absolute deviation). Správně by odchylky pro MAD měly být počítány od mediánu časové řady, nicméně jelikož se hodnoty řady vyvíjí, byla za tento střed zvolena approximace neuronovou sítí. Následně se pomocí bloků Union, Replace Missing Values a Filter Examples připojí nový sloupec MAD do původní tabulky. Generate Attributes poté spočítá Min, Max, Outliers a Correct atributy. Jelikož MAD reprezentuje odchylku, jeho trojnásobek udává okraj normálního rozložení. Za korektní data v této úloze považujeme data v oblasti normálního rozložení kolem approximované hodnoty. Min a Max jsou tedy odečtené/přičtené trojnásobky MAD od/k predictionCloseUSD, Correct jsou hodnoty predictionCloseUSD mezi Min a Max, Outliers jsou pak zbylé hodnoty. Výsledek detekce outlierů pro každý index lze nalézt v příloze B.

## 5 Analýza podobnosti indexů

V rámci poslední úlohy jsme zkoumali závislosti mezi jednotlivými indexy pomocí korelace. Pro zpracování využíváme atributy Index, Date a CloseUSD. Atribut Date je třeba využít pro nalezení vhodného intervalu, v rámci něhož chceme korelací zkoumat. S CloseUSD jsme schopni sledovat denní přírustky nebo úbytky. Celý proces zkoumání korelace je zachycen diagramem 6.



Obrázek 6: Diagram popisuje proces hledání korelace mezi jednotlivými indexy.

V první fázi procesu načteme CSV soubor indexProcessed.csv. Pomocí operátoru Select Attributes vybereme ze vstupní datové sady výše zmíněné atributy. Následně je potřeba nalézt časový interval, ve kterém máme data pro všechny indexy. Pomocí funkce Aggregate nalezneme nejménší a největší možné datum pro každý z indexů. Následně z těchto nalezených dat (opět pomocí funkce Aggregate) vybereme největší minimální datum a nejménší maximální datum. Tyto dva výsledky nám dávají časový rámec, v rámci něhož budeme zkoumat závislost. Pro pohodlnou práci je pro každé datum vytvořeno makro (date\_from, date\_to), které se později využívá dále v rámci procesu.

V druhé větvi procesu vyfiltrujeme datovou sadu pomocí hodnot zapsaných v makrech a vzorky seřadíme podle data. Dále pomocí funkce Generate Attributes vytvoříme pro každý unikátní název indexu nový atribut (vzniknou tedy atributy HSI, NYA, ...). Do jejich hodnot je zapsáno CloseUSD (v případě, že název atributu odpovídá hodnotě v atributu Index) nebo otazník (v opačném případě) značící chybějící hodnotu.

Protože v současném stavu stále pro každé datum existuje více řádků, kde je vždy vyplňena pouze jedna hodnota CloseUSD pro daný index a zbytek je doplněn otazníky, tak je potřeba provést poslední agregaci. V tomto případě jako aggregační atribut využijeme jednotlivé indexy a seskupovat budeme podle Date. Po aggregaci dochází pro přehlednost k přejmenování atributů.

Z dat již nyní můžeme odebrat Date, pro další operace již není potřeba. Posledním krokem je zavolání operátoru Correlation Matrix nad upravenými daty. Výslednou matici (viz příloha B) vyvedeme na výstup procesu.

## Závěr

Byly provedeny tři dolovací úlohy. V rámci první jsme natrénovali neuronovou síť schopnou predikovat hodnoty cen indexů na burze. Síť neapproximuje perfektně, nicméně to je žádoucí vzhledem ke generalizaci. Ve druhé úloze jsme opět natrénovali prediktivní hlubokou síť, která tentokrát approximuje lépe a pomocí metody MAD jsme detekovali odlehlé hodnoty v časových řadách. Nakonec byla provedena korelační analýza a zjistili jsme podobnost mezi indexy.

## Příloha A - Grafy skutečných a predikovaných cen indexů

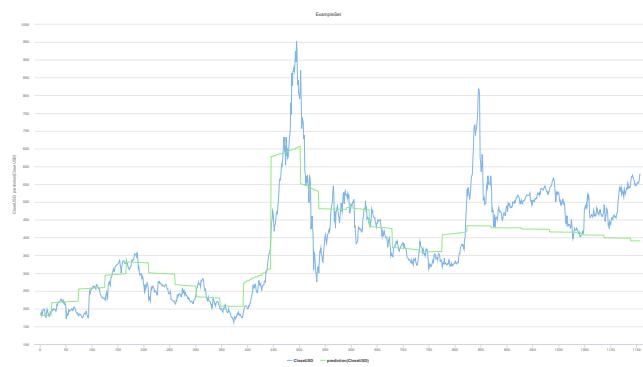
Legenda:

Osa X: vývoj v čase

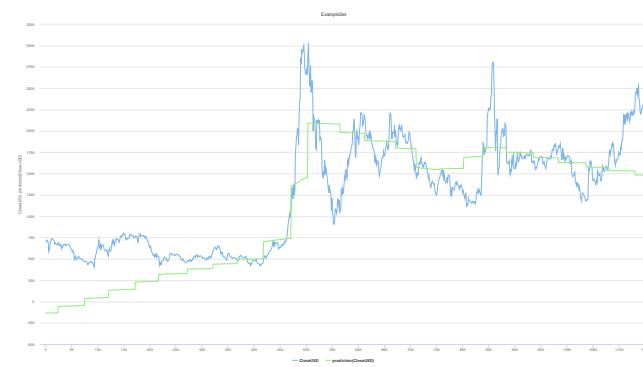
Osa Y: cena v USD

Modrý graf: reálná cena

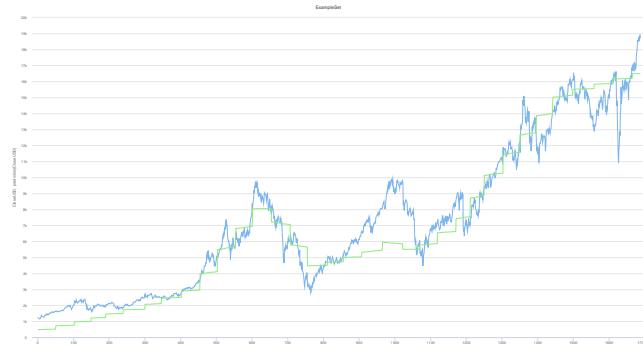
Zelený graf: predikovaná cena



000001



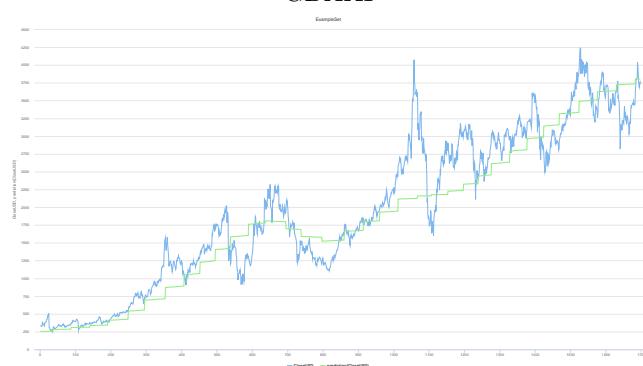
399001



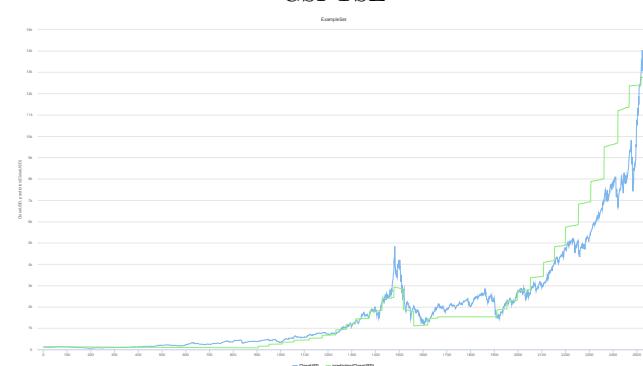
GDAXI



GSPTSE



HSI



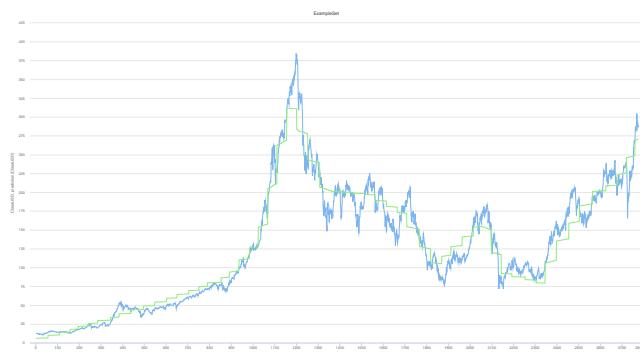
IXIC



J203



N100



N225



NSEI



NYA



SSMI



TWII

## Příloha B - Grafy odlehčených hodnot

Legenda:

Osa X: vývoj v čase

Osa Y: cena v USD

Modrý graf: aproximovaná cena

Zelený graf: očekávané hodnoty

Červený graf: odlehčené hodnoty / anomálie



000001



399001



GDAXI



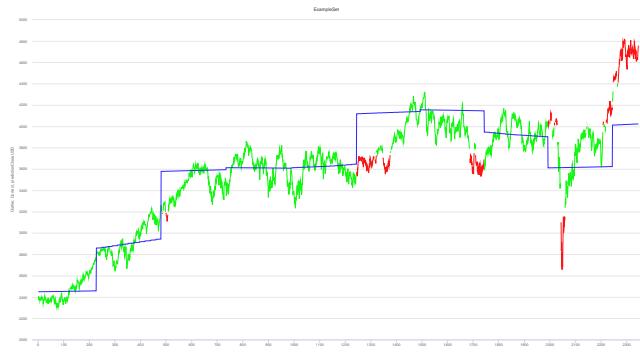
GSPTSE



HSI



IXIC



J203



N100



N225



NSEI



NYA



SSMI



TWII

## Příloha C - Korelační matice zachycující závislosti mezi indexy

Attribut...	HSI	NYA	IXIC	000001....	N225	N100	399001....	GSPTSE	NSEI	GDAXI	SSMI	TWII	J203JO
HSI	1	0.822	0.660	0.517	0.782	0.823	0.337	0.794	0.789	0.799	0.701	0.705	0.779
NYA	0.822	1	0.909	0.542	0.957	0.952	0.439	0.972	0.966	0.948	0.915	0.924	0.933
IXIC	0.660	0.909	1	0.504	0.899	0.821	0.557	0.867	0.936	0.846	0.863	0.963	0.797
000001....	0.517	0.542	0.504	1	0.681	0.669	0.866	0.539	0.574	0.697	0.570	0.505	0.650
N225	0.782	0.957	0.899	0.681	1	0.956	0.554	0.905	0.946	0.960	0.920	0.883	0.934
N100	0.823	0.952	0.821	0.669	0.956	1	0.475	0.928	0.928	0.980	0.901	0.823	0.948
399001....	0.337	0.439	0.557	0.866	0.554	0.475	1	0.435	0.479	0.532	0.487	0.569	0.466
GSPTSE	0.794	0.972	0.867	0.539	0.905	0.928	0.435	1	0.935	0.929	0.883	0.901	0.915
NSEI	0.789	0.966	0.936	0.574	0.946	0.928	0.479	0.935	1	0.922	0.881	0.921	0.903
GDAXI	0.799	0.948	0.846	0.697	0.960	0.980	0.532	0.929	0.922	1	0.897	0.852	0.948
SSMI	0.701	0.915	0.863	0.570	0.920	0.901	0.487	0.883	0.881	0.897	1	0.847	0.870
TWII	0.705	0.924	0.963	0.505	0.883	0.823	0.569	0.901	0.921	0.852	0.847	1	0.815
J203JO	0.779	0.933	0.797	0.650	0.934	0.948	0.466	0.915	0.903	0.948	0.870	0.815	1

Obrázek 7: Korelace mezi indexy