

## I. IDENTIFIKAČNÍ ÚDAJE

<b>Název práce:</b>	<b>Estimating patient's life expectancy after a successful kidney transplant using machine learning methods</b>
<b>Jméno autora:</b>	<b>Kyrylo Stadniuk</b>
<b>Typ práce:</b>	bakalářská práce
<b>Fakulta:</b>	Fakulta jaderná a fyzikálně inženýrská (FJFI)
<b>Katedra:</b>	Katedra matematiky
<b>Oponent práce:</b>	Ing. Jiří Franc Ph.D.
<b>Pracoviště oponenta práce:</b>	Katedra matematiky, FJFI, ČVUT v Praze, Trojanova 13, 120 00 Praha 2

## II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

<b>Zadání</b>	<b>náročnější</b>
<i>Hodnocení náročnosti zadání závěrečné práce.</i>	
Zadání bakalářské práce je velmi komplexní a ambiciózní. Student měl za úkol nejen zpracovat a analyzovat velká a cenzurovaná reálná data pomocí metod strojového učení, ale také dataset rozšířit, provést úpravy skórovacího algoritmu pro transplantace ledvin a na závěr integrovat výsledky do softwarového nástroje. Samotné zadání předpokládá součinnost třetích stran a klade velkou náročnost i na projektové řízení samotného projektu, aby byl zvládnutelný v rozsahu bakalářské práce.	

<b>Splnění zadání</b>	<b>splněno</b>
<i>Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.</i>	
Zadání práce bylo víceméně splněno, ale mám k němu několik výhrad. V práci je příliš prostoru věnováno bodům 1 a 2 na úkor ostatních bodů, které definují samotný autorův přínos. V bodu 3 student měl za úkol předzpracovat a rozšířit datovou sadu, jelikož pak ale pracoval s jinými daty než bylo původně zamýšleno, není v práci zdůvodněno proč data již nerozšiřoval a proč je dále neupravoval vzhledem k jejich charakteru. Zadání je příliš restriktivní co se týče návrhu skórovacího algoritmu a integrace do TX Matching. Student tyto body částečně splnil, ale více se věnoval vlastní softwarové implementaci. Ocenil bych, kdyby bylo v práci vlastní řešení více diskutováno vzhledem k původnímu zadání, které bylo v některých bodech až příliš konkrétní.	

<b>Zvolený postup řešení</b>	<b>vhodný s výhradami</b>
<i>Posuďte, zda student zvolil správný postup nebo metody řešení.</i>	
Přestože v teoretické části jsou diskutovány tři různé přístupy (regrese, klasifikace, analýza přežití), je ve výsledné práci zvolen jen poslední z nich. Buď bych se od začátku zaměřil jen na něj, nebo pak ty ostatní implementoval a porovnal je mezi sebou. Když jim je v teorii věnováno tolik prostoru, je zdůvodnění, proč nebyly použity, velmi stručné. Největší a zásadní výhradu mám ale k samotnému přístupu k datům. Ta jsou sbírána za dlouhé časové období, přesto je s nimi v práci pracováno jako s nezávislými a stejně rozdělenými pozorováními. Jelikož je úkolem forecasting, je vždy potřeba ukázat a porovnávat predikce na budoucích datech vzhledem k trénovací množině. Zde tento základní princip ve zpracování dat pomocí ML nevidím. Dále mi v práci chybí podrobnější statistiky pro trénovací, testovací a validační sadu pro kontrolu možného přetřénování. Myslím, že se práce zaměřila více na implementaci MVP, než na řádnou analýzu, která ale měla být vzhledem k zadání a oboru studia prioritní.	

<b>Odborná úroveň</b>	<b>průměrná</b>
<i>Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.</i>	
Metody strojového učení v analýze přežití nejsou standardním obsahem bakalářského studia, tudíž student musel většinu znalostí získat samostatně. Také způsob práce s velkými daty na výpočetním clusteru přesahuje klasický studijní rámec. Je	

otázkou nakolik byl student k daným řešením veden školitelem, který by měl mít v dané oblasti praktické znalosti a nakolik si musel cestu prošlapávat sám metodou pokus omyl.  
Zvolené a prezentované metody jsou standardní co se týče aplikace na cenzorovaná data v analýze přežití, čemuž také odpovídá možnost použití knihovny scikit learn a absence potřeby implementovat vlastní řešení.

**Formální a jazyková úroveň**

**průměrná**

*Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.*

Práce je psána dobrou a srozumitelnou angličtinou, ale úvodní a teoretické kapitoly jsou z mého pohledu často zbytečně rozsáhlé, zejména ve věcech, které pak nejsou v aplikační části použity. Bohužel to vede i k tomu, že není zachována konzistence značení co se týče označení features, nebo odhadovaných parametrů napříč jednotlivými sekcemi. U obrázků je často uvedena bez jednotková veličina time, nebo je špatný popis jako u 3.1.

**Výběr zdrojů, korektnost citací**

**výborné**

*Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.*

Zdroje jsou relevantní a jsou citovány správně, ale úvodní rešeršní pasáže se hodnotí v dnešní době generativní AI složitěji. Proto bych uvítal, kdyby se student více zaměřil na svůj přínos než na rešerši, která tvoří polovinu práce. V práci mi ale chybí více porovnání s podobnými analýzami ML metod pro data přežití, případně data selhání. Práci by jistě prospělo uvedení podobných přístupů u cenzurovaných dat, i když z jiných oblastí, než transplantace ledvin.

**Další komentáře a hodnocení**

*Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.*

Jak již zaznělo, práce je velmi rozsáhlá a je vidět, že student musel odvést velké penzum práce z různých oblastí. Z mého pohledu by šlo některé pasáže, především v rešeršní části, vypustit a lépe čtenáři vysvětlit důvody pro zvolený přístup. Velmi oceňuji uvedení kódu na Githubu, kde je i vidět velké množství autorova kódu.

**III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE**

*Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.*

Celkově se práce nehodnotí snadno, jelikož díky ambicióznímu zadání a složitosti dát nebylo žádné přímé a snadné řešení. Je otázkou, jaké body byly pro studenta a jeho školitele klíčové a které naopak upozadili.

K obhajobě navrhuji následující otázky:

1-V práci je zmíněna metoda zobecněných lineárních modelů, které se používají v analýze přežití a Coxův PH model. Oba modely umožňují počítat poměr šancí a vliv jednotlivých proměnných. Jak oba modely spolu souvisí a jaký je mezi nimi rozdíl?

2-Jak byly zpracovány interakce mezi jednotlivými proměnnými a lze z modelů některé silné interakce vyšších řádů vyčíst?

3-Jak se testovala nezávislost a stejné rozdělení dat (iid) a pokud data nejsou iid, jak bylo naloženo s train/val/test rozdělením datasetu vzhledem k časové proměnné sběru dat?

4-Můžete v rámci obhajoby ukázat zkoumané metriky a křivky (obr. 4.1 -4.4) nikoli napříč jednotlivými modely na testovací sadě, ale pro jednotlivé modely na trénovací, validační a testovací sadě?

5-Byla provedena kalibrace modelu co se týče predikovaných pravděpodobností přežití pro dané období? Jsou výsledné modely dobře kalibrovány? Pokud možno při obhajobě ukažte.

Vzhledem k uvedeným skutečnostem hodnotím práci klasifikačním stupněm C, ale pokud student zodpoví uvedené otázky a prokáže orientaci v dané oblasti během obhajoby, jsem ochoten kvůli rozsahu práce a složitosti zadání, klasifikaci během obhajoby posunout.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **C - dobře**.

Datum: 20.8.2024

Podpis:

