基于 Diabetes 130-US Hospitals for Years 1999-2008 Data Set 数据集的糖尿病患者再入院预测与分析

姓名：贾亚伟

学号：2021Z8017782015

专业：电子信息

背景

糖尿病（Diabetes Mellitus）是最常见的慢性疾病之一，主要以高血糖为特征，其影响范围波及全球 4 亿多人，同时糖尿病具有明显的家族遗传特性，接近一半的糖尿病患者有家族遗传病史。糖尿病的病理原因主要包含两种情况，第一种是当胰腺无法产生充足的胰岛素（一种调节血糖或血葡萄糖的荷尔蒙）时引发的 1 型糖尿病（T1D），第二种是当所产生的胰岛素无法被人体有效地利用时引发的 2 型糖尿病（T2D）。T1D 通常被称为原发性糖尿病，这种糖尿病类型的发病机制通常是当胰腺中进行胰岛素分泌的 β 细胞受到损伤时，人体在短时间内没有充足的胰岛素供使用，从而导致血糖含量无法被及时降低至安全区间内，该过程也称作是胰岛β细胞的郎格罕氏（Langerhans）胰岛自身免疫性破坏。另一种更为常见的糖尿病类型的医学名称是非胰岛素依赖型糖尿病，简称 T2D。该类糖尿病通常由胰岛素抵抗或者是胰岛素分泌缺陷等因素引起，造成高血糖的直接原因是身体内的胰岛素没能得到有效利用。而引发 2 型糖尿病的主要原因通常包括生活方式、身体活动、饮食习惯和遗传等因素。

在中国过去三十多年的社会发展历史中，随着糖尿病患者人数不断上升，人们开始意识到这一普遍影响家庭生活和个人幸福的慢性疾病所带来的影响，但糖尿病患者的总数量仍然较二十多年前已经至少翻了一倍。

2017 年全世界范围内糖尿病患病人群的数量将达到近四亿两千五百万。而根据近年的增长率预测到 2045 年全球的糖尿病患者将达到六亿两千九百万人，这个数将超过全世界总人口数量的近十分之一，这一惊人的数字毫无疑问需要引起我们的高度重视。

据研究表明，2013 年至 2018 年期间，我国糖尿病的危险因素没有得到明显改善，部分危险因素甚至变得更为严重。例如，红肉摄入量过多的比例从 32.6%增加到 42.3%，身体活动不足率从 16.0%增加到 22.0%。

而根据我国判断肥胖的体重指数（BMI）标准，5 年间我国肥胖患病率从 14.1%上升到 16.5%；中心性肥胖从 31.6%增加到 35.4%。到 2018 年，一半左右成年人处于超重或肥胖状态。因此，考虑到我国庞大的糖尿病前期人群数量以及还在上升的肥胖态势等因素，如果不加大糖尿病防控力度，未来我国糖尿病患病率可能会进一步增加。

因此，糖尿病患者的病情预测就变得十分有意义，本实验将在UCI machine learning 提供的关于糖尿病的数据集的基础上探究糖尿病患者再入院的潜在风险特征因素。

数据概况

本实验使用的数据集为 UCI machine learning 提供的 Diabetes 130-US Hospitals for Years 1999-2008 Data Set 该数据集为对 1999 至 2008 年间经过不同医院实验室测试的十万名糖尿病患者进行描述的数据集，它经过临床专家的筛选后仅保留了 50 个最可能与糖尿病病情相关的属性，数据是代表弗吉尼亚联邦大学临床和转化研究中心提交的，该中心是 NIH CTSA 拨款 UL1 TR00058 的接受者和 CERNER 数据的接受者。该数据集代表了美国 130 家医院和综合交付网络的 10 年（1999-2008）糖尿病相关的临床护理数据。

该数据集包括代表患者和医院结果的 50 个特征。从数据库中提取满足以下标准的实验的信息。

(1) 是住院会诊（入院）。

(2) 这是一次糖尿病实验，即任何一种糖尿病都被输入系统作为诊断。

(3)停留时间不少于 1 天，最多 14 天。

(4) 在实验期间进行了实验室测试。

(5) 会诊期间给予药物治疗。

数据包含以下属性：患者编号、种族、性别、年龄、入院类型、住院时间、入院医生的医学专业、进行的实验室检测次数、HbA1c 检测结果、诊断、用药次数、糖尿病药物、门诊人数住院前一年的住院、急诊等等。

数据集特点：	多元	实例数量：	100000	领域：	生活类
属性特征：	整数	属性数量：	50	上传日期：	2014.5.3
相关任务：	分类/聚类	缺值：	是	浏览次数：	377464

表 1.数据集的相关信息

该数据集用在预测糖尿病患者是否会出院后再入院方面有好的效果，因此本实验也使用该数据集做相关的机器学习预测分类任务。

方法

本实验方法分为两个部分，第一部分为数据预处理阶段对于选取的数据集中数据的清洗即空缺值的预测填充；第二部分为根据清洗后的数据对糖尿病患者再入院的预测。

在数据清洗以及预处理阶段使用了 KNN（k 最邻近）算法以及随机森林算法对缺失值进行预测，根据算法的预测准确度择优进行数据的填充。

KNN（K-Nearest Neighbor）算法是 机器学习算法中最基础、最简单的算法之一。它既能用于分类，也能用于回归。KNN 通过测量不同特征值之间的距离来进行分类。KNN 算法的思想非常简单：对于任意 n 维输入向量，分别对应于特征空间中的一个点，输出为该特征向量所对应的类别标签或预测值。

KNN 算法是一种非常特别的机器学习算法，因为它没有一般意义上的学习过程。它的工作原理是利用训练数据对特征向量空间进行划分，并将划分结果作为最终算法模型。存在一个样本数据集合，也称作训练样本集，并且样本集中的每个数据都存在标签，即我们知道样本集中每一数据与所属分类的对应关系。输入没有标签的数据后，将这个没有标签的数据的每个特征与样本集中的数据对应的特征进行比较，然后提取样本中特征最相近的数据（最近邻）的分类标签。一般而言，我们只选择样本数据集中前 k 个最相似的数据，这就是 KNN 算法中 K 的由来，通常 k 是不大于 20 的整数。最后，选择 k 个最相似数据中出现次数最多的类别，作为新数据的分类。

随机森林算法是最常用也是最强大的监督学习算法之一，它兼顾了解决回归问题和分类问题的能力。随机森林是通过集成学习的思想，将多棵决策树进行集成的算法。对于分类问题，其输出的类别是由个别树输出的众数所决定的。在回归问题中，把每一棵决策树的输出进行平均得到最终的回归结果。

在对糖尿病患者再入院的预测中，使用了 LSTM（长短期记忆网络）、随机森林、逻辑回归和决策树算法来进行预测，最终并比对了各个算法的优劣。

长短期记忆网络（LSTM，Long Short-Term Memory）是一种时间循环神经网络，是为了解决一般的 RNN（循环神经网络）存在的长期依赖问题而专门设计出来的，所有的 RNN 都具有一种重复神经网络模块的链式形式。在标准 RNN 中，这个重复的结构模块只有一个非常简单的结构，例如一个 tanh 层。

logistic 回归又称 logistic 回归分析，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此因变量就为是否胃癌，值为“是”或“否”，自变量就可以包括很多了，如年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的，也可以是分类的。然后通过 logistic 回归分析，可以得到自变量的权重，从而可以大致了解到底哪些因素是胃癌的危险因素。同时根据该权值可以根据危险因素预测一个人患癌症的可能性。

决策树(Decision Tree）是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。

决策树是一种十分常用的分类方法。它是一种监督学习，所谓监督学习就是给定一堆样本，每个样本都有一组属性和一个类别，这些类别是事先确定的，那么通过学习得到一个分类器，这个分类器能够对新出现的对象给出正确的分类。

实现步骤

首先，收集到实验使用的数据集后的第一个任务就是将该数据集导入到项目中去，然后查看数据集的 shape 以及其数据集的描述。可以了解到共有 101766 个患者记录，每个记录具有 50 个属性。由于该数据及包含了同一患者的多次记录，而本实验的目的只为了研究糖尿病患者 30 天内再入院的可能，因此我们需要去除掉同一用户的重复信息，同时该数据集还包含了已经死亡以及临终关怀的患者，为更有利于研究我们也要把这一部分患者去除掉。

数据清洗的另一大任务就是去除过量含有空值的属性和将部分空值进行合理的赋值。

encounter_id	0.000000
patient_nbr	0.000000
race	2.741057
gender	0.000000
age	0.000000
weight	96.015606
admission_type_id	0.000000
discharge_disposition_id	0.000000
admission_source_id	0.000000
time_in_hospital	0.000000
payer_code	43.466766
medical_specialty	48.074257
num_lab_procedures	0.000000
num_procedures	0.000000
num_medications	0.000000
number_outpatient	0.000000
number_emergency	0.000000
number_inpatient	0.000000
diag_1	0.014291
diag_2	0.418733
diag_3	1.749246
number_diagnoses	0.000000
max_glu_serum	0.000000
A1Cresult	0.000000
metformin	0.000000
repaglinide	0.000000
nateglinide	0.000000

图 1. 各属性空值的分布情况

需要根据空值的分布图来进行部分属性的取舍，由于体重的空值已经占了 96%，属于高度数据缺失，因此将该列去除。并且支付人代码和医学专业这两个属性的空值量也比较大 43%和 48%，但可以通过使用分布预测与归因技术来填补其空缺值，因此将其保留。

在变量分析环节，首先观察了种族的分布，发现种族分布主要是以白种人为主其次是非洲裔美国人和亚裔，因此将该属性中占比 2.7%的缺失值替换为白种人；

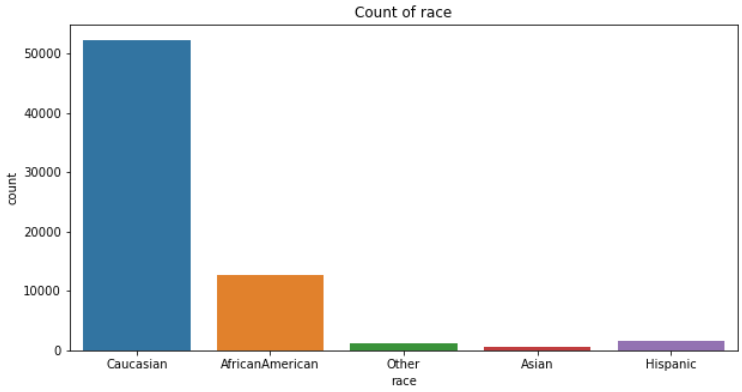


图 2. 参与实验患者的种族分布情况

在性别分布这一属性中，其中有 3 个性别值是未知的，选择把这些行给删除掉，从性别分布中可以看出女性人数多于男性人数，但差异很小；

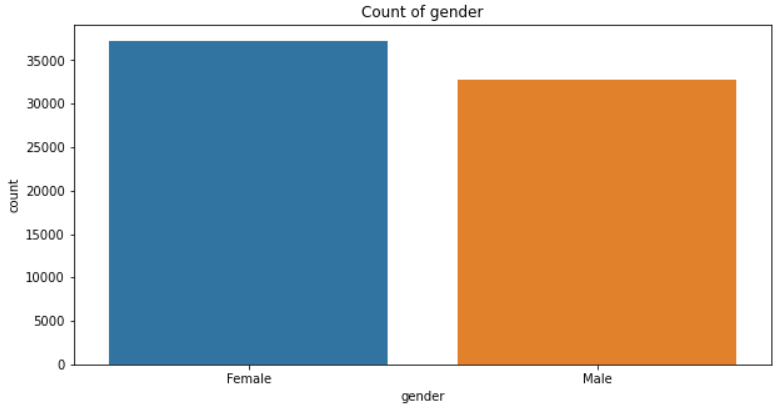


图 3. 参与实验患者的性别分布情况

在年龄分布中发现，年龄小于 40 岁的患者与年龄大于 40 岁的患者相比数量较少，患者主要分布在 70—80 年龄段内，为了便于后续研究将患者年龄分为 0—30、30—60、60—100 三类；

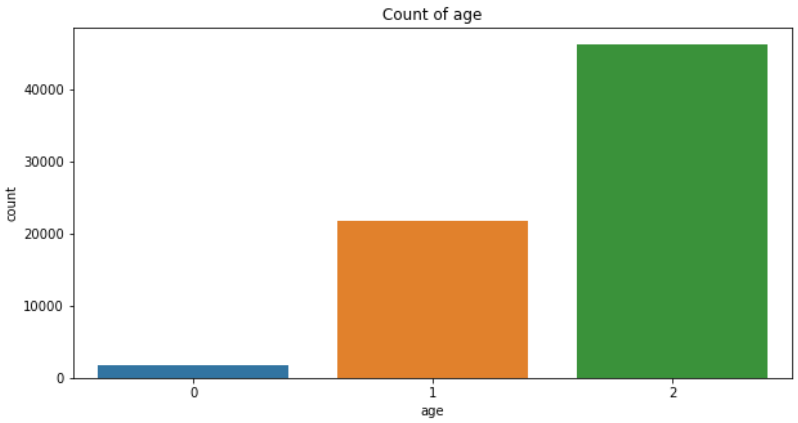
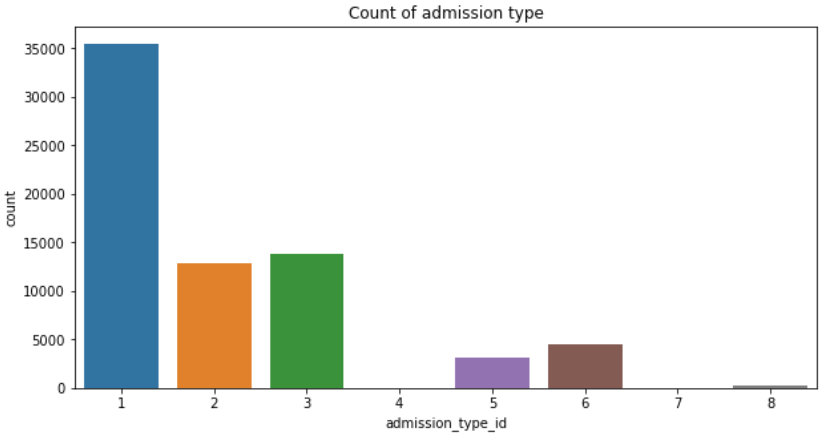


图 4. 参与实验患者的年龄分布情况

在入院类型分布中，可以看出大部分患者都是因急诊而入院；



admission_type_id	description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped

图 5. 参与实验患者的入院类型分布情况

在出院处置类型分布中，可以看出大部分患者都回了家，又因为原本的分类过于糅杂，因此在进行重新归类将原本的 29 类归为了 8 类；

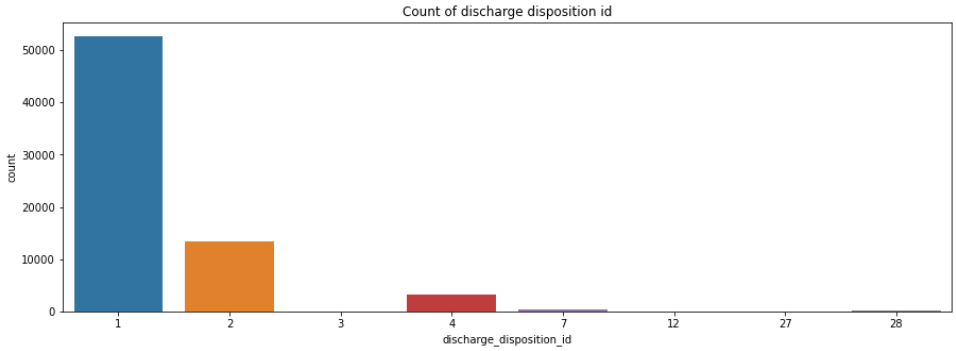


图 6. 参与实验患者的出院处置类型分布情况

再入院来源分布中可以看出大多是来自于急诊，其次就是转诊，后又将原本的 26 类入院类型归纳为 8 类；

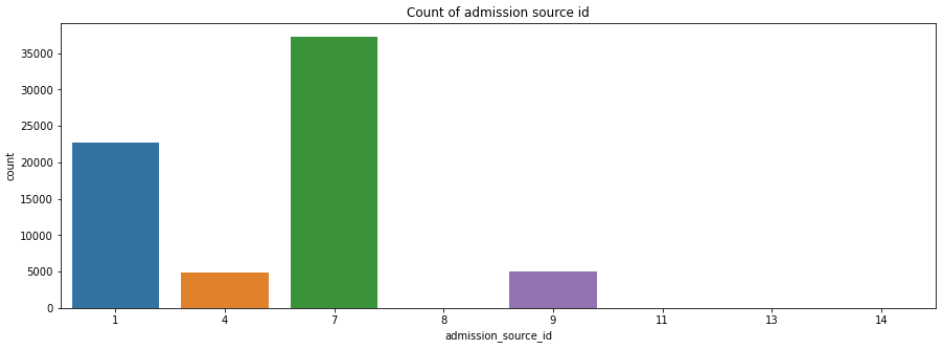


图 7. 参与实验患者的再入院来源分布情况

在住院时间分布中可以看出，患者的住院时间分为 1 天到 14 天不等，患者平均停留 4 天，大部分患者停留 3-4 天，患者很少停留超过 12 天；

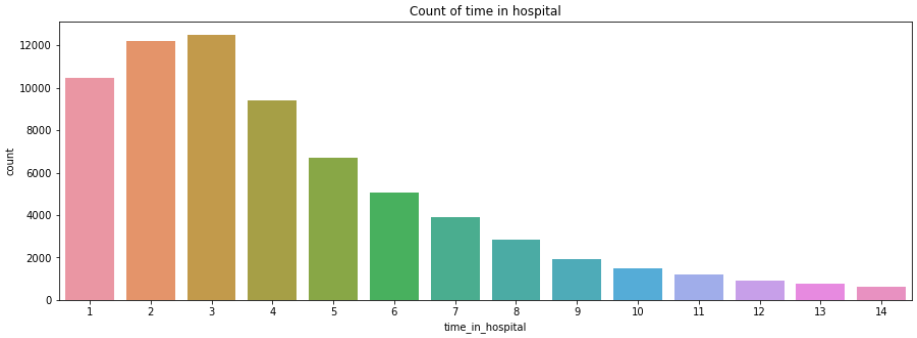


图 8. 参与实验患者的住院时间分布情况

在实验期间进行的实验的数量分布中可以看出，实验者的平均实验次数为 43 次；

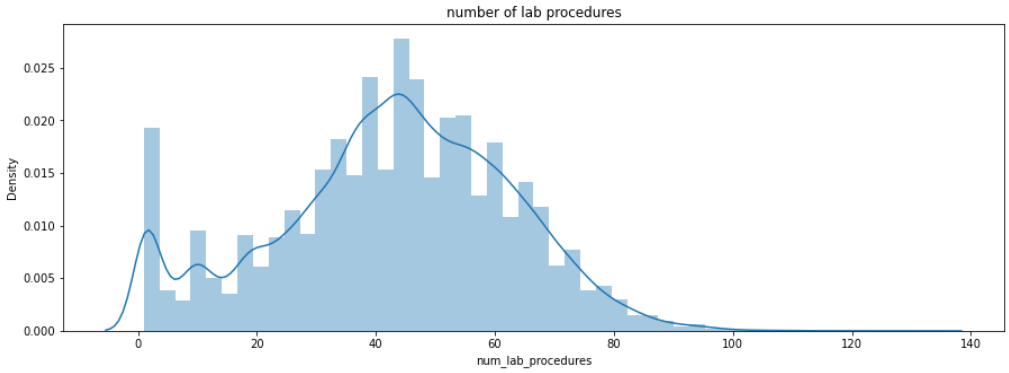


图 9. 参与实验患者的在实验期间进行的实验的数量分布情况

在实验期间执行的程序（实验室测试除外）的数量分布中可以看出，大部分患者没有再做相关的实验；

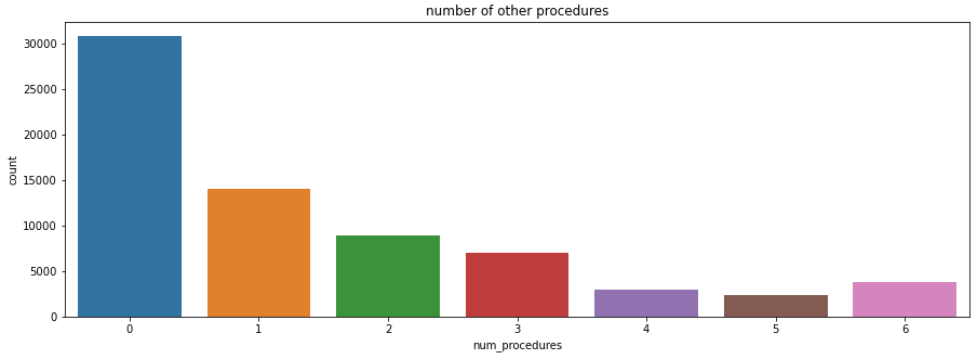


图 10. 参与实验患者的在实验期间执行的程序的数量分布情况

在患者的摄入药品种类分布中可以看出，大多数患者平均获得 16 种药物，只有 7 名患者服用了 70 多种药物；

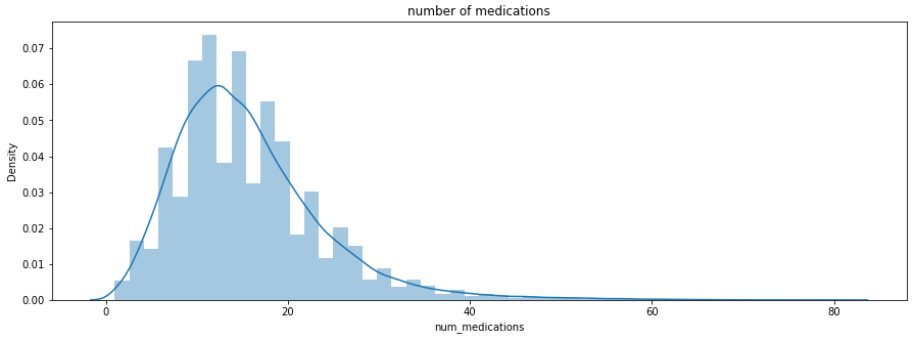


图 11. 参与实验患者的在实验期间摄入药品种类分布情况

在患者就诊前一年的就诊次数中可以看出，大多数患者没有任何门诊就诊记录，门诊次数超过 15 次的患者非常少；

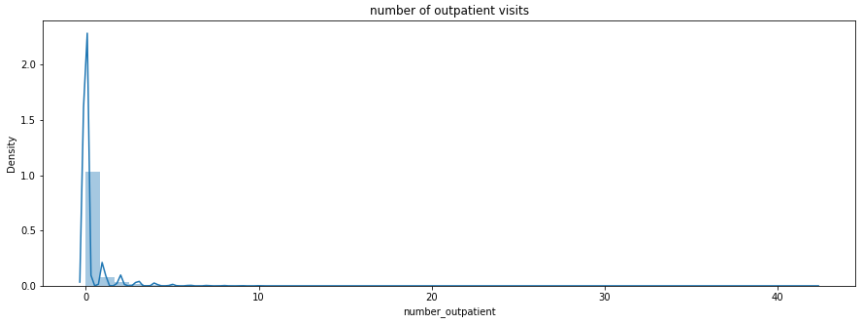


图 12. 参与实验患者的就诊前一年的就诊次数分布情况

在患者入院前一年的急诊次数分布中可以看出，大部分患者都没有急诊记录；

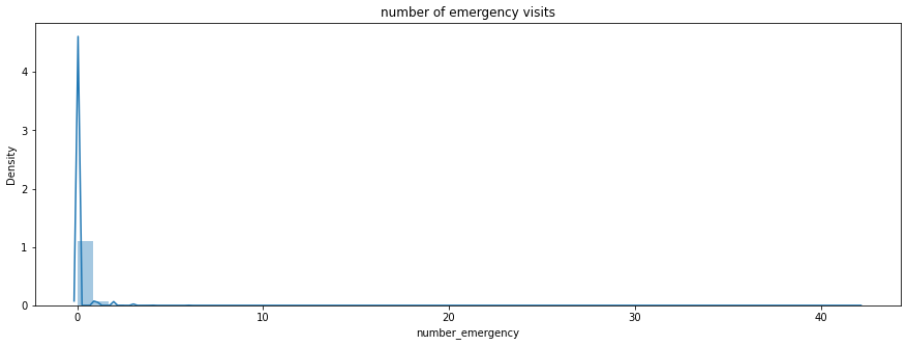


图 13. 参与实验患者入院前一年的急诊次数分布情况

在患者入院前一年的入院记录分布中可以看出，大部分患者就诊前都没有住院就诊记录；

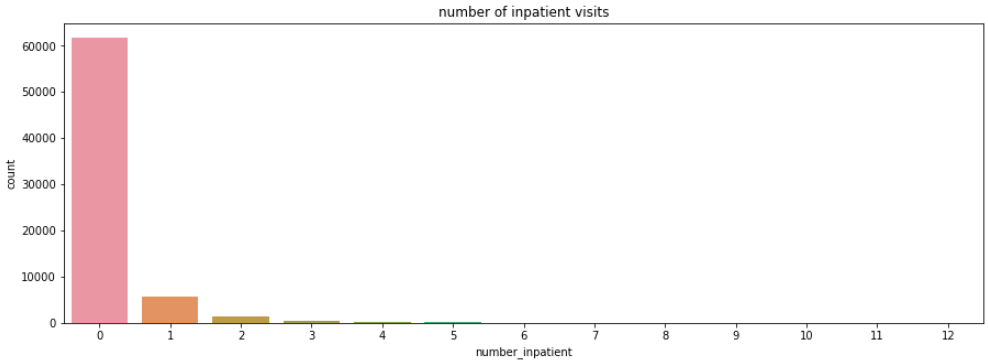


图 14. 参与实验患者入院前一年的入院记录分布情况

在三次诊断记录中，将诊断结果分为 icd9 code 的编码格式转化为数字编码，并分别统计不同次的诊断过程中诊断出不同疾病的分布情况，可以观察出大多数患者被诊断出患有呼吸系统疾病和其他疾病类型，nan 类别也随着诊断数的增加而增加；

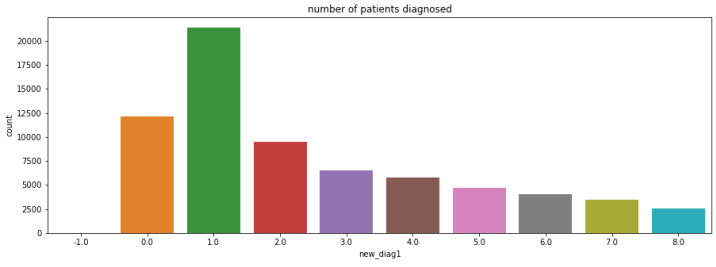


图 15. 参与实验患者第一次诊断的结果分布情况

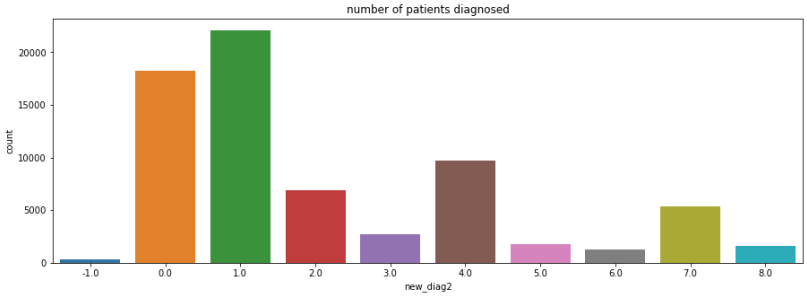


图 16. 参与实验患者第二次诊断的结果分布情况

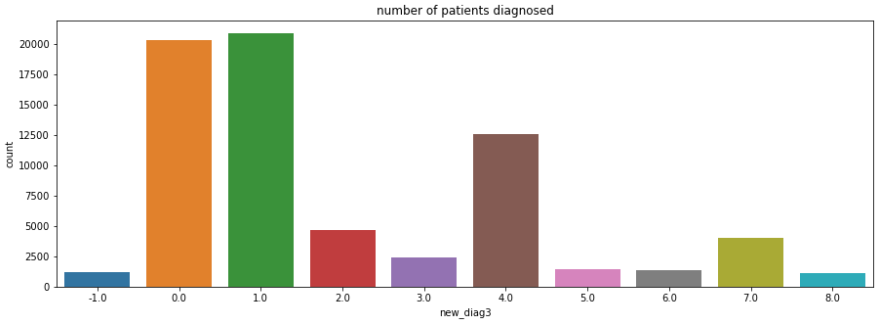


图 17. 参与实验患者第三次诊断的结果分布情况

在参与试验的患者的指输入系统的诊断数的统计中，可以看出大多数的患者都经历了 9 次诊断，超过 9 次诊断的患者数量比较少；

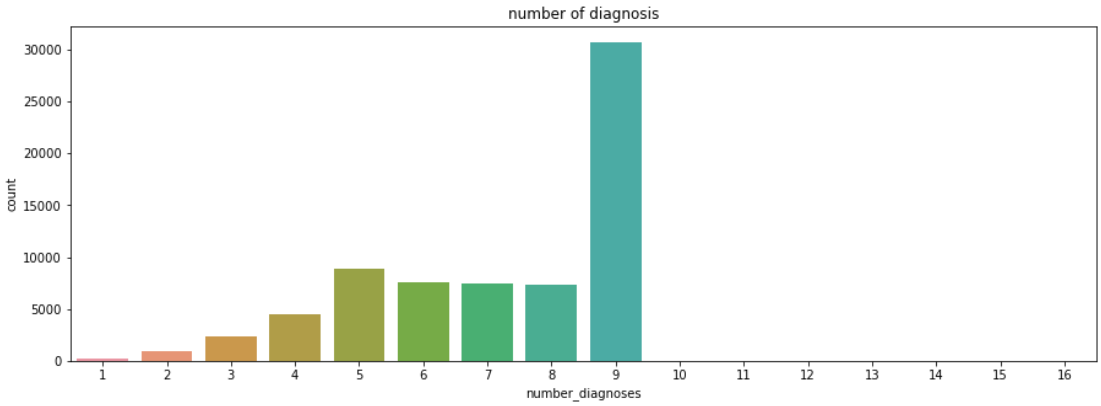


图 18. 参与实验患者诊断数分布情况

在参与试验的患者的葡萄糖血清测试其中值包含：“> 200”，“>300”，“正常”和“无”(如果未测量)，可以观察出大多数患者不接受这个测试在接受此测试的人中，大约一半的患者结果正常，另一半患者的结果在 >200 或 >300 的类别中。

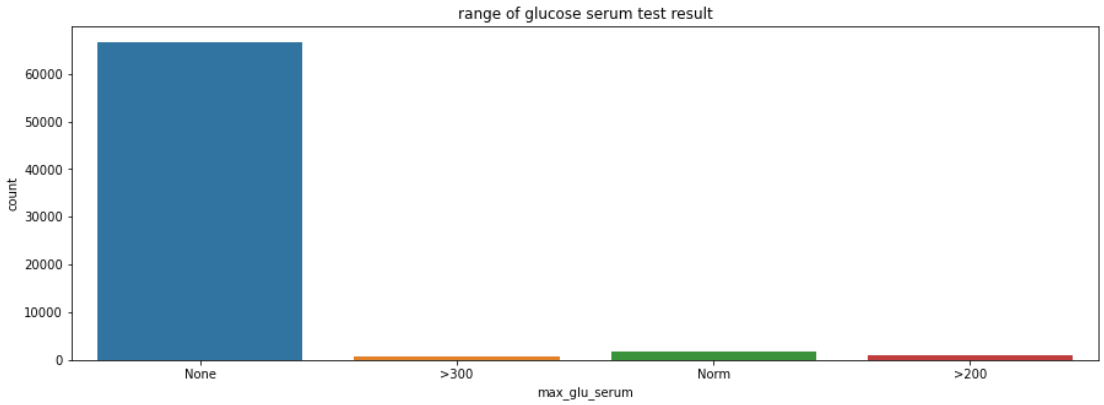


图 19. 参与实验患者葡萄糖血清测试结果分布情况

A1Cresult 指示结果范围或未进行测试。 值：如果结果大于 8%，则为“> 8”；如果结果大于 7%但小于 8%，则为“> 7”；如果结果小于 7%，则为“正常”；如果结果大于 7%，则为“无”没有测量，从统计结果中可以看出，大部分患者都不做此项检测，在接受此测试的人中，近一半的患者结果>8，另一半患者的结果要么>7，要么正常；

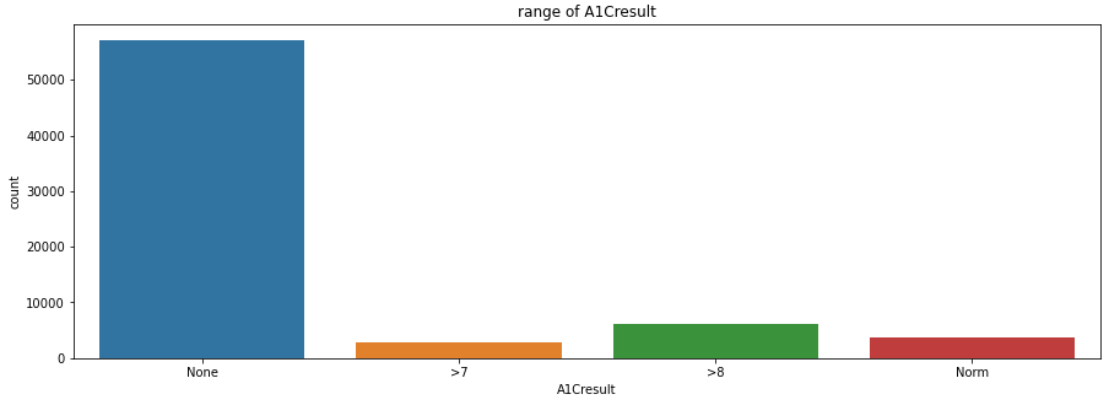


图 20. 参与实验患者 A1Cresult 测试结果分布情况

在实验期间所用药的剂量变化统计中，值：如果在实验期间增加剂量，则为“向上”，如果剂量减少，则为“向下”，如果剂量没有变化，则为“稳定”，如果未开药，则为“否”，从统计器 1 情况中我们可以观察到 examide, citoglipton 和 glimepiride-pioglitazone 的所有特征值都为 No。这 3 个特征无助于分类患者是否在 30 天内再次入院，因为所有值都相同。因此，让从数据集中删除这些特征，并且药物可以合并为一个特征，并且可以计算患者服用的药物数量；

change 指示糖尿病药物（剂量或通用名称）是否发生变化。价为：“改变”和“不变”，可以观察出大部分患者都没有换药；

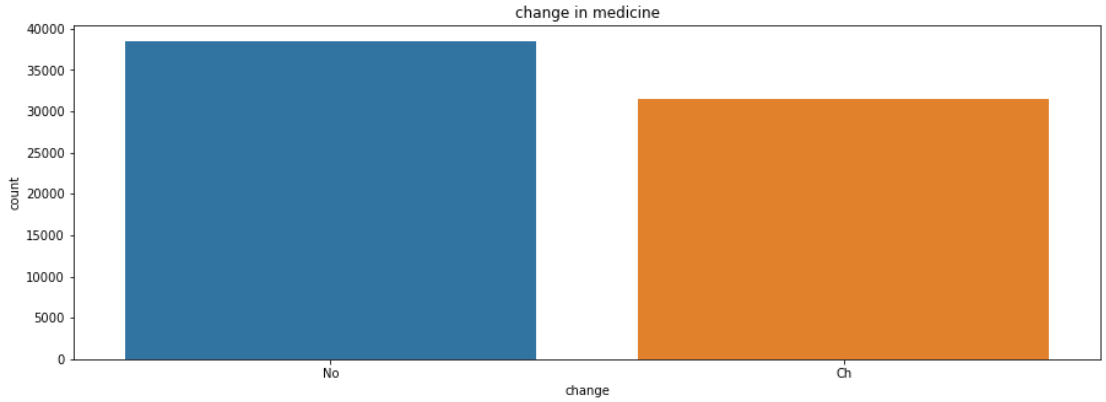


图 21. 参与实验患者是否换药统计分布情况

DiabetesMed 表示是否要糖尿病的处方药，从统计结果可以看出，大部分患者都使用了糖尿病处方药；

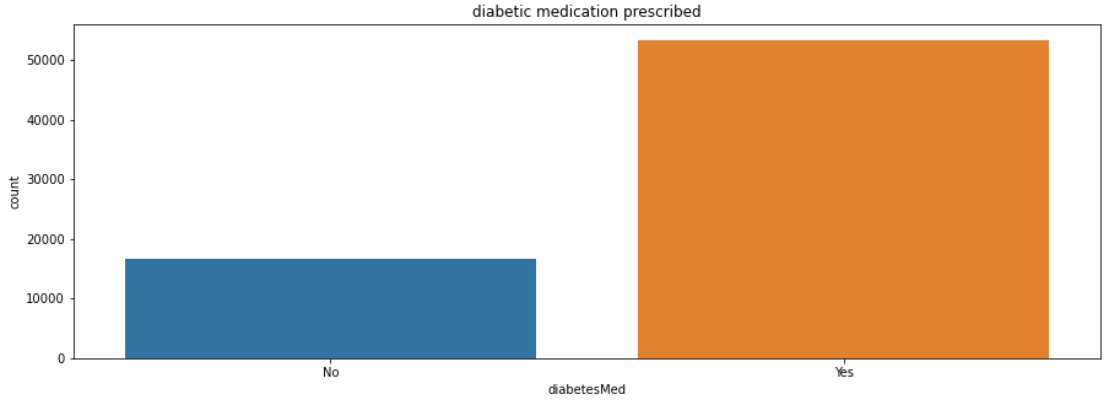


图 22. 参与实验患者是否糖尿病处方药分布情况

Readmitted 这是我们必须预测的变量，它指的是住院再入院的天数。值：“<30”表示患者在 30 天内再次入院，“>30”表示患者在 30 天内再次入院，“否”表示没有再入院记录，在此将“>30”和“否”的患者归为一类编码为 0，将“<30”编码为 1，从图中我们可以观察到，在 30 天内重新接纳的人数较少，大多数人要么没有重新接纳，要么在 30 天后重新接纳；

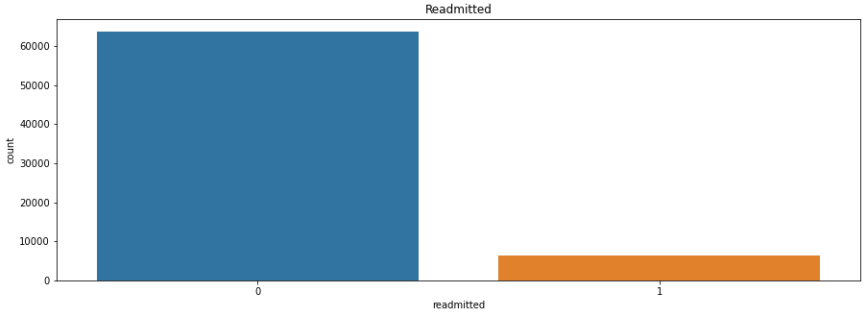


图 23. 参与实验患者 30 天内重返医院分布情况

在未进行空值填充前的 “Payer_code”数据分布如下图所示，可以观察出大部分患者付款由医疗保险（MC）完成，其他支付用的不多，只用了 MC、HM、BC。同时可以统计出共有 30414 个值缺失，需要后续进行填充；

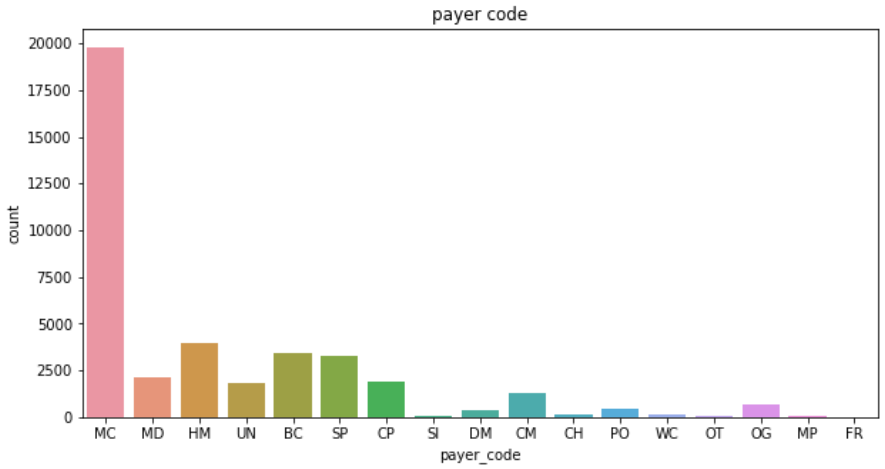


图 24. 参与实验患者支付形式分布情况（未填充空缺值）

在未进行空值填充前的“medical_specialty”属性可以观察出，大多数患者都是通过内科进行住院的，家庭/全科医师、心脏病学和急诊/创伤医师也占一些患者。其他类别的患者数量非常少。

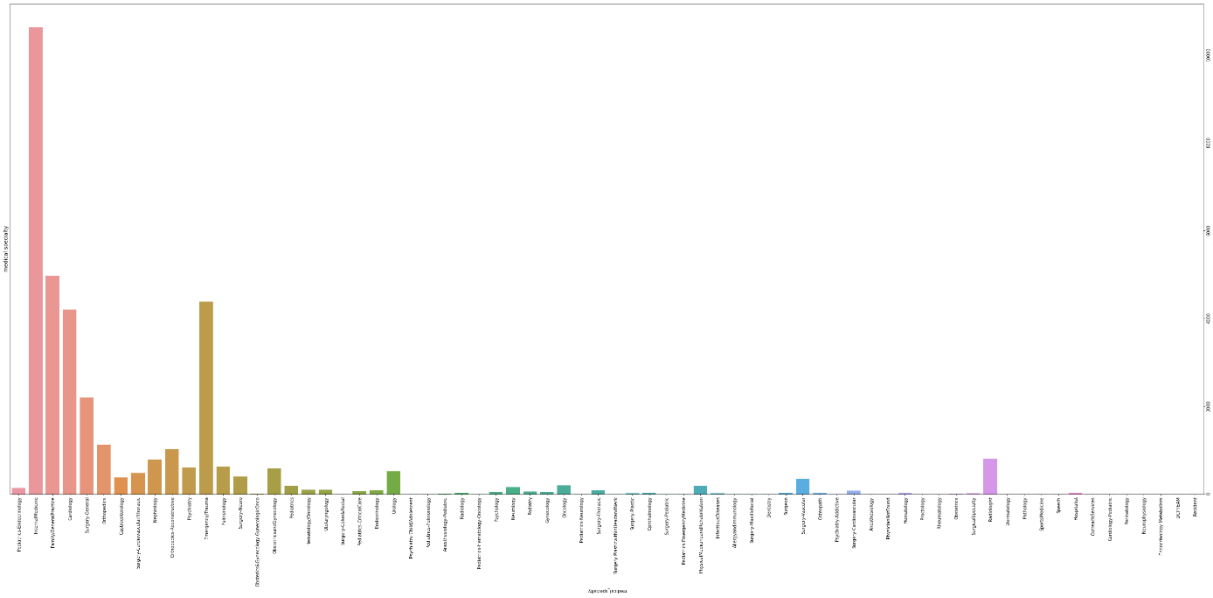


图 25. 参与实验患者参与的医疗专业分布情况（未填充空缺值）

在处理“medical_specialty”和“Payer_code”缺省值过程中，首先将数据分为 75%的训练集以及 25%的测试集。在“medical_specialty”属性使用随机森林进行数据测试的结果显示测试数据的预测得分为 53.63%，在使用 knn 模型进行预测过程中，使用了不同的 k 值进行测试，并得到最好效果的结果的预

测的得分为 43.22%，因此在“medical_specialty”属性中缺省值的填充使用随机森林算法；在“Payer_code”属性使用随机森林进行数据测试的结果显示测试数据的预测得分为 52.15%，在使用 knn 模型进行预测过程中，使用了不同的 k 值进行测试，并得到最好效果的结果的预测的得分为 48.32%，因此在“Payer_code”属性中缺省值的填充使用随机森林算法。

下图为“Payer_code”属性进行空值填充后的分布情况：

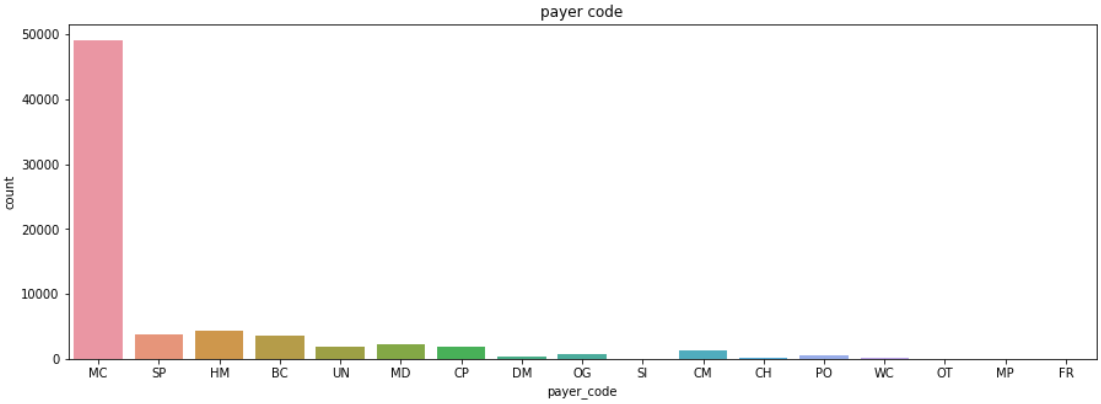


图 26. 参与实验患者支付形式分布情况（填充空缺值后）

下图为“medical_specialty”属性进行空值填充后的分布情况：

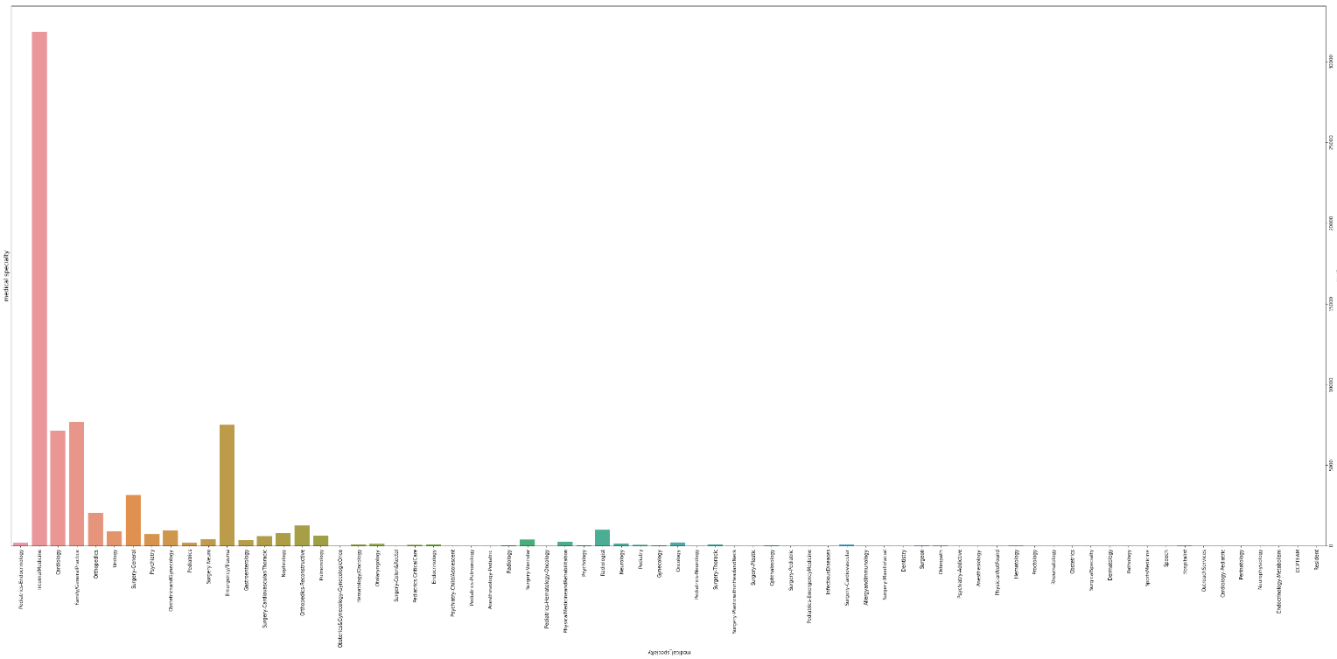


图 27. 参与实验患者参与的医疗专业分布情况（填充空缺值后）

在进行完变量分析后，可以得出以下结论：门诊、住院和急诊就诊可以合并为一个新的特征就诊；删除了药物的三个特征，因为它们不提供任何可能有助于预测患者再入院的信息；药物可以合并为一个特征，并且可以计算患者服

用的药物数量；诊断功能已从 icd9 代码更改为 10 个不同的类别；基于模型的插补应用于具有缺失值的特征；分类标签可以是一种热编码以将分类标签转换为数字数据；数据高度不平衡，只有 9% 的患者在 30 天内重新入院，因此必须进行过采样。

在进行完数据预处理后，就需要进行统计数据之间的相关以及依赖关系，从下图中，可以观察到 “num_medications”、“number_diagnoses”、“num_lab_procedures”等特征往往与住院时间呈正相关；诊断特征与其他特征的相关性非常低；重新接纳还显示与其他特征的低相关性表明与特征不存在线性关系。

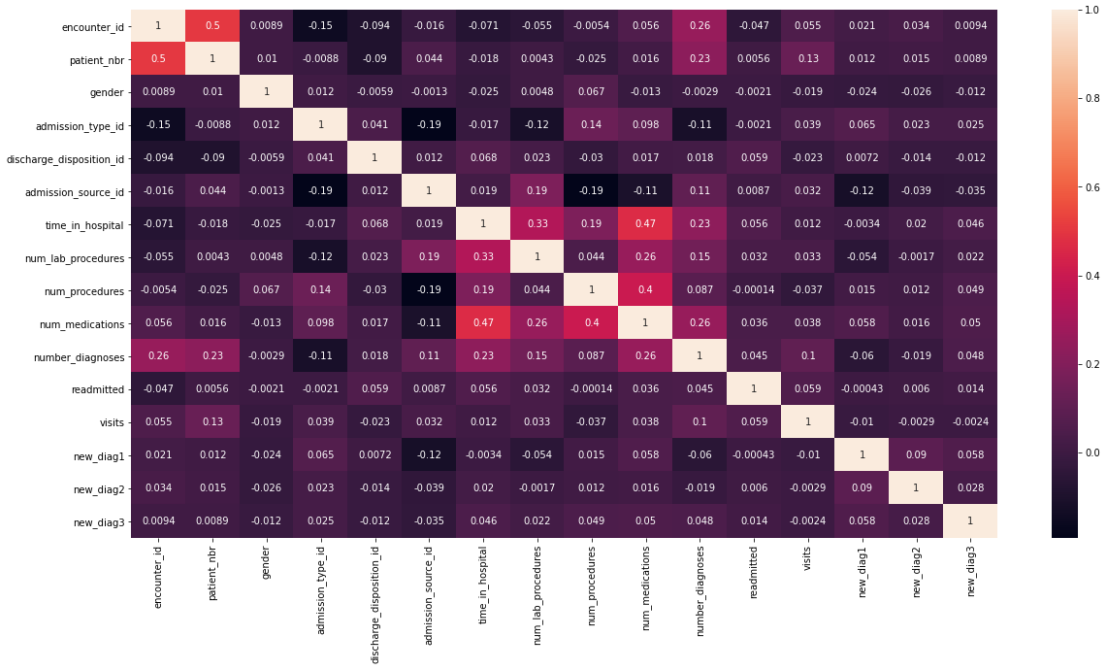


图 28. 属性与再入院之间的热图

在通过 VIF 进行统计检查多重共线性，从下图中可以看“number_diagnoses”的 vif 值是 15.7 因大于 10，故将此列删除，“age”的 vif 值为 10。将“number_diagnoses”列删除后，发现其余的 vif 都小于 10。

	variables	VIF
0	encounter_id	5.105695
1	patient_nbr	4.138981
2	gender	1.858407
3	age	10.028209
4	admission_type_id	3.435313
5	discharge_disposition_id	2.205312
6	admission_source_id	4.493303
7	time_in_hospital	4.484776
8	num_lab_procedures	7.149095
9	num_procedures	2.111642
10	num_medications	7.958893
11	number_diagnoses	15.777810
12	max_glu_serum	1.346388
13	A1Cresult	1.338864
14	metformin	1.892441
15	repaglinide	1.045946
16	nateglinide	1.026182
17	chlorpropamide	1.007063
18	glimepiride	1.233732
19	acetohexamide	1.000382
20	glipizide	1.625885
21	glyburide	1.616982
22	tolbutamide	1.001821
23	pioglitazone	1.263222
24	rosiglitazone	1.240341
25	acarbose	1.009318
26	miglitol	1.001497
27	troglitazone	1.000714
28	tolazamide	1.003143
29	insulin	6.394770
30	glyburide-metformin	1.031127
31	glipizide-metformin	1.001622
32	metformin-rosiglitazone	1.000485
33	metformin-pioglitazone	1.000702
34	change	6.289290
35	diabetesMed	8.096375
36	new_diag1	2.181135
37	new_diag2	1.866151
38	new_diag3	1.822291
39	number_outpatient	1.091817
40	number_emergency	1.077721
41	number_inpatient	1.135062

	variables	VIF
0	encounter_id	4.801163
1	patient_nbr	4.071332
2	gender	1.849637
3	age	8.540281
4	admission_type_id	3.421902
5	discharge_disposition_id	2.180916
6	admission_source_id	4.248665
7	time_in_hospital	4.417299
8	num_lab_procedures	7.012428
9	num_procedures	2.108065
10	num_medications	7.720680
11	max_glu_serum	1.341149
12	A1Cresult	1.338400
13	metformin	1.887780
14	repaglinide	1.045943
15	nateglinide	1.028169
16	chlorpropamide	1.006864
17	glimepiride	1.233289
18	acetohexamide	1.000380
19	glipizide	1.623752
20	glyburide	1.615293
21	tolbutamide	1.001818
22	pioglitazone	1.262402
23	rosiglitazone	1.240339
24	acarbose	1.009316
25	miglitol	1.001461
26	troglitazone	1.000688
27	tolazamide	1.003066
28	insulin	6.393975
29	glyburide-metformin	1.030748
30	glipizide-metformin	1.001620
31	metformin-rosiglitazone	1.000479
32	metformin-pioglitazone	1.000640
33	change	6.289254
34	diabetesMed	8.036848
35	new_diag1	2.180549
36	new_diag2	1.865265
37	new_diag3	1.814752
38	number_outpatient	1.091634
39	number_emergency	1.077589
40	number_inpatient	1.134727

图 29.删除“number_diagnoses”列前后各属性的 VIF 值

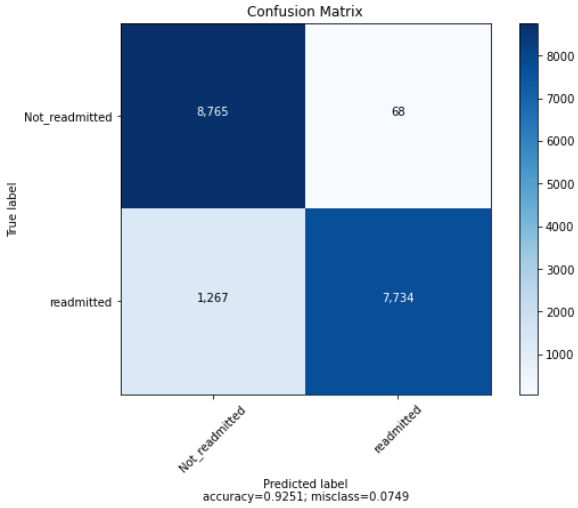
本实验采用十折交叉验证来进行模型的训练与预测，十折交叉验证英文名叫做 10-fold cross-validation，用来测试算法准确性，是常用的测试方法。将数据集分成十份，轮流将其中 9 份作为训练数据，1 份作为测试数据，进行试验。每次试验都会得出相应的正确率（或差错率）。10 次的结果的正确率（或差错率）的平均值作为对算法精度的估计，一般还需要进行多次 10 折交叉验证（例如 10 次 10 折交叉验证），再求其均值，作为对算法准确性的估计。

将 smote 应用于过采样，因为要预测的变量是不平衡的，并且在实施过程中仅在训练数据上使用 smote。

结果

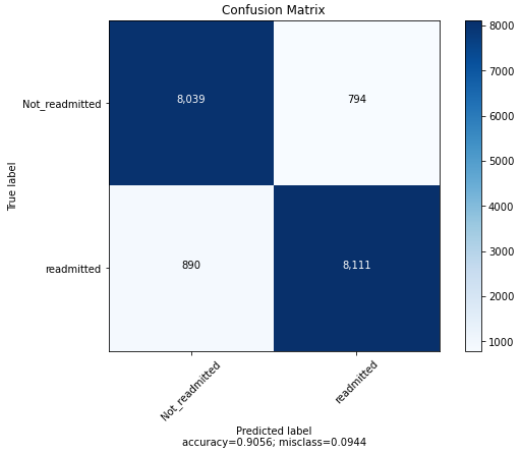
在最终的实验部分，首先使用逻辑回归来进行预测，使用十折交叉验证方法最终得到的是 10 次训练以及预测的准确度的平均值，最终使用逻辑回归再结合网格搜索的大最优参数预测的平均准确度为 0.925，及相关的模型评估结果以及混淆矩阵结果如下图所示：

```
Aver_Accuracy Score : 0.9251429853089604
Aver_Precision Score : 0.9912842860804921
Aver_Recall Score : 0.8592378624597267
Aver_F1 Score : 0.9205499018032494
Aver_AUC Score : 0.9257697293731895
```



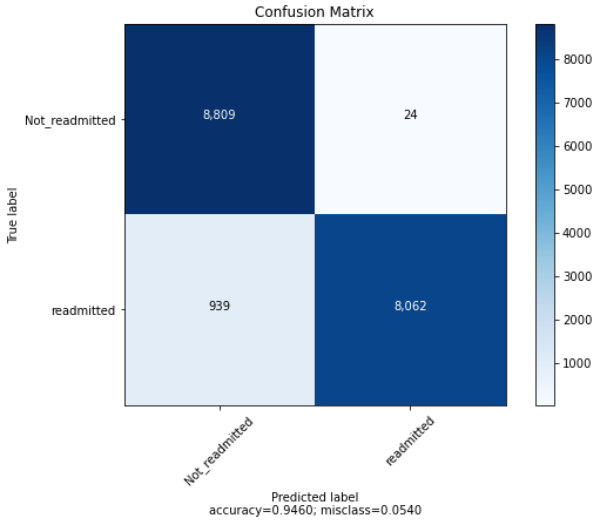
在使用决策树进行患者再入院的预测可以得到 0.9 左右的准确率，相关的模型评估结果以及混淆矩阵结果如下图所示：

```
Avg_Accuracy Score : 0.905573623415947
Avg_Precision Score : 0.9108366086468276
Avg_Recall Score : 0.9011220975447173
Avg_F1 Score : 0.9059533117390818
Avg_AUC Score : 0.9056159565047258
```



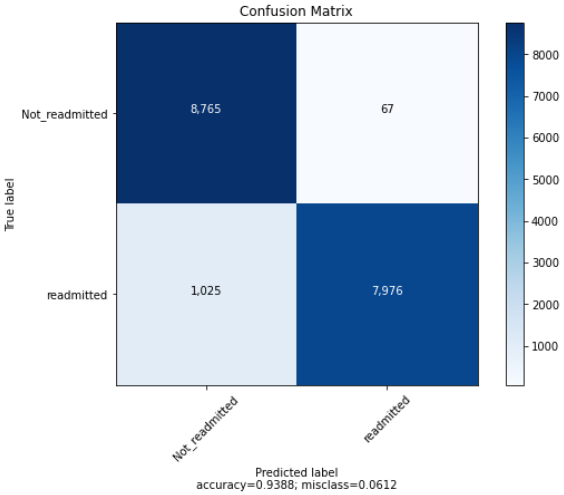
在使用随机森林进行患者再入院的预测可以得到 0.94 左右的准确率，相关的模型评估结果以及混淆矩阵结果如下图所示：

```
Avg_Accuracy Score : 0.9460020186161265
Avg_Precision Score : 0.9970319069997526
Avg_Recall Score : 0.8956782579713365
Avg_F1 Score : 0.9436413647802423
Avg_AUC Score : 0.9464805871539009
```

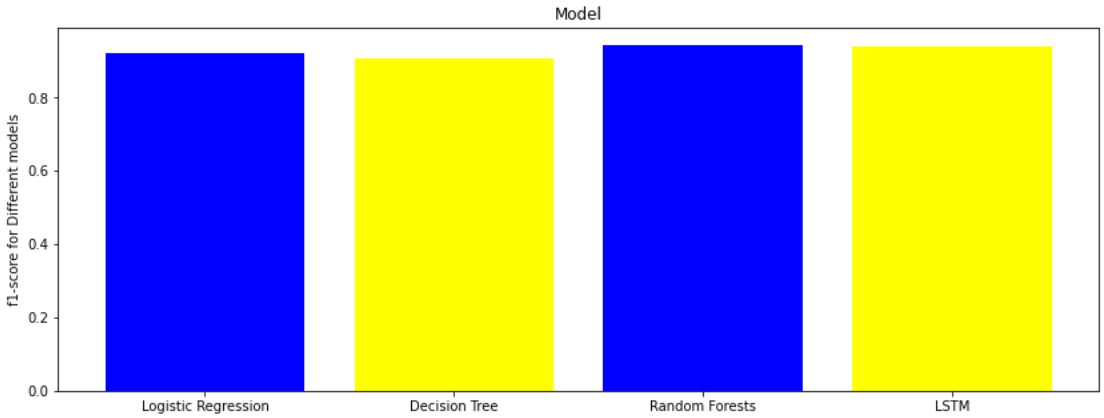


在使用 LSTM 进行患者再入院的预测可以得到 0.94 左右的准确率，相关的模型评估结果以及混淆矩阵结果如下图所示：

```
Avg_AUC:0.9607136200094113
Avg_f1:0.9387651
```



最后再进行统计制作出四种模型最终预测的 AUC 统计分布以及 f1 统计分布，



Model	AUC score	f1 score
Logistic Regression	0.925769729	0.9205499018
Decision Tree	0.905615956	0.9059533117
Random Forests	0.946480587	0.943641364
LSTM	0.9607136200	0.9387651

可以观察出，当只考虑 AUC 时，LSTM 模型的表现占主导地位；当只看 f1 值时，随机森林占主导地位。

讨论

通过此次实验实践了五种较为简单的数据挖掘算法，深刻理解了所实验算法的优劣。其中 K-Means 算法的优点为简单高效且当簇接近高斯分布时效果较好，缺点是 k 值再现实聚类时大多难以估计，且初始的中心点对聚类结果影响很大；Logistic 回归分析的优点为实现效率较高且可以解决多重共线性，主要应用于工业问题中，缺点是当特征空间很大时，逻辑回归的性能不是很好同时也不能很好地处理大量多类特征或变量，且依赖于全部的数据特征，当特征有缺失的时候表现效果不好；决策树的特点是可以不用对数据进行预处理但易出现过拟合现象；随机森林在并行处理超大数据集时能提供良好的性能表现同时也会自动避免过拟合现象的发生；LSTM 是 RNN 的一个优秀的变种模型，继承了大部分 RNN 模型的特性，同时解决了梯度反传过程由于逐步缩减而产生的 Vanishing Gradient 问题。

此次实验让我对数据挖掘的常用算法有了一定的掌握，对数据挖掘的思路有了初步的认识，对以后的科研生活有积极的推进作用。