

并非所有特征都值得关注：使用语言模型的图引导依赖学习进行表格数据生成

Zheyu Zhang Shuo Yang Bardh Prenkaj Gjergji Kasneci

Technical University of Munich

{ name.surname } @tum.de

Abstract

大型语言模型 (LLMs) 已经在通过建模文本化的特征-值对生成表格数据方面展示出了强大潜力。然而，表格数据本质上表现出稀疏的特征级依赖性，其中许多特征交互在结构上并不显著。这导致了一个根本的不匹配，因为 LLMs 的自注意力机制不可避免地将关注分散到所有对上，从而削弱了对关键关系的注意力，特别是在具有复杂依赖关系或语义模糊特征的数据集中。为了应对这一限制，我们提出了 GraDe (图引导的依赖学习)，这是一种新颖的方法，通过将稀疏的依赖图明确地整合到 LLMs 的注意力机制中。GraDe 采用一个轻量级的动态图学习模块，由外部提取的功能依赖引导，优先考虑关键特征交互，同时抑制无关的交互。我们在各种真实世界的数据集上进行的实验表明，GraDe 在复杂数据集上比现有基于 LLM 的方法的表现提升高达 12 个百分点，同时在合成数据质量方面取得了与最先进方法相当的结果。我们的方法入侵性极小但有效，为具有结构感知的表格数据建模提供了一种实用的解决方案。

1 介绍

表格数据构成了无数现实世界应用的基础，这些应用涵盖了医疗分析(?)、财务建模(?)、人口统计研究(?)以及科学实验(?)。生成高质量的合成表格数据对于数据增强、隐私保护和模型测试变得越来越重要(?)。然而，由于表格式数据集固有的复杂结构特性，这一任务呈现出独特的挑战。

与图像或文本不同，表格数据在特征之间表现出不同的关系模式。有些特征具有很强的相关性，而其他特征则完全相互独立(?)。更为关键的是，某些特征可能通过逻辑或业务规则唯一地决定其他特征的值。数据库理论将这些模式形式化为函数依赖(FD)，即通过已知一个特征(或一组特征)的值可以确定另一个特征的值(?)。例如，邮政编码可以唯一地确定城市和州，而与年龄或收入等个人属性无关。如果不保持这些依赖性，合成数据将遭受逻辑

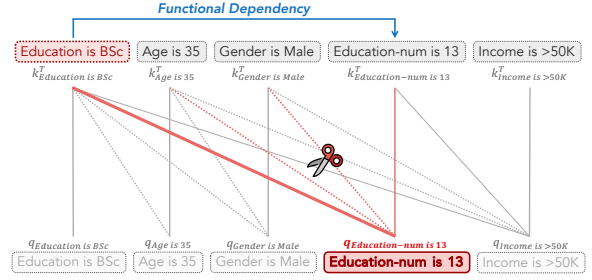


Figure 1: 大规模语言模型的密集注意力模式与表格数据的稀疏依赖结构之间的结构性不匹配。

不一致的问题，例如客户 ID 与多个矛盾的个人详细信息相关联，从而使生成的数据在实际应用中不切实际。

近年来，表格数据生成取得了快速进展(????)。在这些方法中，大型语言模型 (LLMs) 由于其在大量语料库上进行预训练的强大泛化能力以及对复杂现实世界关系的理解，提供了一个引人注目的解决方案(?)。首次通过将数据转换为文本序列(例如，「年龄是 39，收入是 $\leq 50K$ 」)并在训练过程中随机打乱特征顺序，将 LLMs 应用于这一任务，使得模型能够在任何特征子集的条件下生成数据。后来表明，这种随机打乱会扰乱模型捕获特征之间依赖关系的能力，因此提出了固定拓扑顺序的方法。这揭示了一个根本性的设计挑战：如何在基于 LLM 的方法中同时保持灵活的条件设定和特征关系的准确建模。

这一挑战的根本原因在于大型语言模型 (LLM) 与表格数据之间的结构性不匹配。如图 1 所示，LLM 使用注意力机制，其中每个标记潜在地与每个其他标记有关(?)。相比之下，表格数据具有稀疏的、非顺序的依赖关系，其中大多数特征是条件独立的。这种根本性的脱节创造了一个两难境地：随机特征排序提供了灵活性，但削弱了依赖建模，而固定排序更好地捕捉关系，但牺牲了灵活性。当特征-值对线性化为文本时，LLM 面临着正确解释每一对而防止无关特征干扰的固有困难任务(?)。为了充分利用 LLM 处理表格数据，我们需要一种既能保持灵活排序又能显式建模表格数据中

固有的稀疏依赖结构的方法。

为了弥合这种结构差距，我们提出了 GraDe (图引导依赖学习)。与以特征排序进行隐式学习的先前方法不同，GraDe 通过将可学习的稀疏依赖图引入注意力机制 (§??)，显式增强了 LLM。我们的方法在动态建模标记级关系时利用外部提取的功能依赖作为指导 (§3.2)，从而在结构上重要的连接上进行集中关注。模型通过多目标损失来平衡语言流畅性、结构稀疏性和功能依赖对齐 (§3.3) 进行训练。为了提高参数效率，我们还引入了 GraDe-Light，这是一种仅更新我们注意模块的变体，同时保持有竞争力的性能。总体而言，我们的贡献如下：

- 问题形式化：我们形式化了 LLM 的密集注意模式与表格数据的稀疏依赖结构之间的结构性不匹配，为理解当前方法的局限性提供了一个框架 (§??)。
- 新颖架构：我们开发了一种图引导注意力机制，该机制动态建模词元之间的稀疏依赖关系，并将功能依赖知识作为软监督整合进去，在保持底层 LLM 架构的同时增强结构意识 (§3)。
- 实证验证：我们通过实验证明，GraDe 和 GraDe-Light 都在现有的基于 LLM 的方法上表现出显著的改进，同时取得了与最新方法相当的结果。我们的方法在具有复杂依赖结构的数据集上表现尤其突出，并在低资源环境中保持优势 (§5)。

2 相关工作

用于表格建模的深度生成模型 早期对表格数据生成的研究采用了深度生成模型，如 GANs、VAEs 和扩散模型 (?)。CTGAN (?) 和 CTAB-GAN (?) 通过条件采样和专门的标准化策略处理混合类型数据 (?)，而 TVAE (?) 采用了一种变分框架以实现更稳定的训练。近期基于扩散的方法如 TabDDPM (?) 和 TabSyn (?) 通过对多模分布的迭代去噪提高了样本质量。值得注意的是，GOGGLE (?) 引入了特征依赖的显式图模型。尽管有这些进展，大多数生成方法仍然难以灵活捕捉表格数据中固有的稀疏依赖结构。

用于表格建模的语言模型 另一研究方向通过将行序列化为文本序列并应用大型语言模型生成来建模表格数据。GReaT (?) 在这一方向上率先将特征-值对序列化为文本，并采用随机特征排列以增强灵活性。然而，这种随机化削弱了模型学习结构依赖关系的能力 (?)。随后的研究工作如 TabLLM (?) 和 TAPTAP (?) 提出

了精细的序列化格式和大规模表格预训练，以更好地保留关系模式。最近的研究如 P-TA (?) 通过策略优化引入对抗性反馈，进一步探索了基于 LLM 的表格生成，强调了超越文本序列化的更强结构引导的必要性。

在各领域中，模型架构与数据结构对齐的重要性越来越受到重视。各种方法旨在实现这种对齐：稀疏注意模型将标记的交互限制在相关的邻域内；图增强方法结合了明确的关系结构；而专门的架构如 GraSAME 则直接将结构信息注入注意机制中。这些发展反映了越来越多的证据表明标准的密集注意机制无法充分捕捉真实数据固有的稀疏性。我们的 GraDe 方法通过引入一个图引导的注意机制，将这种见解拓展到表格建模中，利用功能性依赖关系聚焦于有意义的特征关系，同时保留语言模型的生成灵活性。

表格数据的生成建模旨在学习关于表格样本的概率分布。给定一个由独立同分布样本组成的训练数据集，其目标是通过学习生成分布的参数来逼近该分布。传统方法通常直接在特征空间中建模这种分布，而基于语言模型的方法通常将表格数据转化为文本表示，以便将联合分布建模为：其中，表示特征-值对的线性化文本序列，例如“特征是，...，特征是”。

在基于 Transformer 的语言模型中，自注意力机制可以被视为输入标记上的一个完全连接图。对于一个标记序列，标记之间的注意力权重计算为：其中是与标记相关的查询和键向量。这个机制可以被解释为定义了一个完全连接的有向图，其中代表标记，表示具有权重的边。这个结构假设任何标记都可以直接影响其他任何标记，这适用于自然语言，其中长距离依赖关系是频繁且具有上下文相关性的。

相反，表格数据通常表现出稀疏的关系结构，其中每个特征仅依赖于其他特征的一个小子集。这样的关系可以表示为一个稀疏图，其中顶点表示特征，边表示成对的依赖关系。形式上，每个特征值的生成过程可以描述为：特征依赖的一小集合是一个函数映射，表示独立噪声。如所观察到的，这样的稀疏关系结构是现实世界数据生成过程的特征。

在将语言模型应用于表格数据生成时，密集注意结构 G_{LM} 和稀疏关系结构 G_{Tab} 之间出现了关键的不匹配。这种不匹配导致了两个主要问题：(1) 注意机制在所有标记对之间分配了不可忽略的权重，随着特征数量的增加，其捕捉少数关键依赖关系的能力被冲淡；(2) 语言模型没有关于表格特征依赖性的任何归纳偏差，必须纯粹通过数据学习此类结构模式，随着表格复杂性的增加，这变得越来越困难。

受到这些限制的启发，我们建议将功能依赖

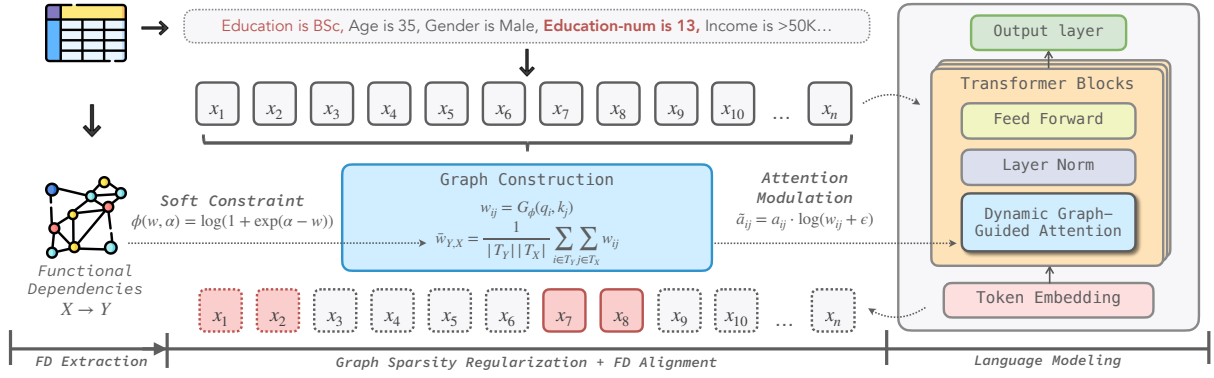


Figure 2: GraDe 的流程。按照 ?，我们首先将表格数据线性化为文本，并随机排列特征顺序，然后将序列输入到一个语言模型中。在训练过程中，会学习到一个动态的令牌级依赖图，由功能性依赖的软约束引导，并用于调节注意力权重，以实现结构感知的表格建模。

结构明确地融入到语言模型中，具体细节将在以下部分中详述。

3 图引导的依赖学习

我们提出了一种新方法 GraDe，该方法增强了语言模型以更好地捕捉表格数据中固有的稀疏依赖结构。基于表格数据在特征关系中包含自然稀疏性的观察，GraDe 将这种结构信息整合到语言模型的注意力机制中。

3.1 概述

最近的研究表明，表格数据中的稀疏依赖关系可以通过关系结构更准确地捕捉到 (?)。虽然语言模型的自注意力机制在令牌之间创建关系表示，但我们假设它们的密集连接模式（即每个令牌关注所有其他令牌）可能会稀释对表格数据中关键关系的关注。

GraDe 通过学习一个动态稀疏图来应对这一挑战，该图引导注意力机制优先考虑结构上重要的依赖关系。我们的方法包含三个关键创新：(1) 图引导的注意力动态地建模稀疏的标记级关系；(2) 将外部派生的功能依赖整合为监督信息；(3) 稀疏性正则化以将注意力集中在有意义的连接上。与之前的方法不同，GraDe 仅对注意力模块进行最小的修改，同时保留了预训练语言模型的语义能力。图 2 展示了我们的架构。

对于一序列的标记，GraDe 为每个注意力头学习一个加权有向图，该图由邻接矩阵 $\mathbf{W} \in [0, 1]^{n \times n}$ 表示，其中每个条目 w_{ij} 量化了从标记 j 到标记 i 的有向关系的重要性。该图由一个轻量级的两层神经网络 G_ϕ 参数化：

$$w_{ij} = G_\phi(q_i, k_j) \quad (1)$$

$$= \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot [q_i; k_j])), \quad (2)$$

，其中 $q_i, k_j \in \mathbb{R}^{d_k}$ 是查询和键向量， $W_1 \in \mathbb{R}^{d_k \times 2d_k}$ 和 $W_2 \in \mathbb{R}^{1 \times d_k}$ 是可学习参数， σ 表示 sigmoid 激活函数。该模块在头部维度 d_k 上操作，从而实现参数效率。

注意力调制 标准自注意力通过查询-键的点积计算比对分数 (?)，将所有的标记对视为同等相关的注意力候选。为了鼓励模型关注重要关系，GraDe 通过一个对数门控机制整合学习到的依赖权重来调整这些分数：

$$\tilde{a}_{ij} = a_{ij} \cdot \log(w_{ij} + \epsilon), \quad (3)$$

其中 a_{ij} 是未归一化的注意力分数， w_{ij} 是学习到的边权重， ϵ 是用于数值稳定的小常数。这个转换充当门控机制：当 $w_{ij} \approx 1$ （表示强依赖）时，它保留原始分数；当 $w_{ij} \approx 0$ 时，它强烈抑制分数。调整后的注意力分数 $\tilde{a}_{ij}$ 通过 softmax 归一化，确保获得有效的注意力分布。

在推理过程中，我们通过仅计算涉及新生成标记的边来优化自回归设置的图计算，从而实现长序列的高效解码。

3.2 融合函数依赖性

动态图模块在序列中捕捉令牌级别的依赖性，而表格数据主要在特征级别表现出重要的关系 (?)。这些关系通常以功能依赖的形式出现，其中一组特征唯一地决定另一组特征。

为了弥合在 token 级和特征级依赖之间的差距，我们将关于 FDs 的先验知识纳入到训练目标中。对于每个已知的 FD $X \rightarrow Y$ ，我们计算从表示 X 中特征的 tokens 到表示 Y 中特征的 tokens 的平均连接强度：

$$\bar{w}_{Y,X} = \frac{1}{|T_Y||T_X|} \sum_{i \in T_Y} \sum_{j \in T_X} w_{ij}, \quad (4)$$

其中 T_X 和 T_Y 分别是与 X 和 Y 特征相对应的 token 索引集。

基于这些平均连接强度，我们定义了一个 FD 对齐损失：

$$\mathcal{L}_{\text{FD}} = \sum_{X \rightarrow Y \in \mathcal{F}} \phi(\bar{w}_{Y,X}, \alpha), \quad (5)$$

其中 $\mathcal{F}$ 是已知 FD 的集合， ϕ 是一个施加最小强度阈值 α 的约束函数。

我们使用光滑的 softplus 公式实现这一约束：

$$\phi(w, \alpha) = \log(1 + \exp(\alpha - w)), \quad (6)$$

这种公式能够灵活地模拟功能依赖关系，同时适应现实世界数据中固有的变异性。通过平滑地惩罚不够强的连接强度，模型被鼓励专注于关键的结构关系，而不会强加过于僵硬的约束。

3.3 多目标微调

GraDe 的训练使用一种复合损失，该复合损失共同优化语言建模质量、结构稀疏性和与已知依赖关系的对齐：

为了保持语言模型的序列生成能力，我们采用标准的自回归训练与交叉熵损失：其中 x 是从表格数据中提取的线性化输入序列。

为了反映表格数据关系中固有的稀疏性，我们在学习的邻接矩阵上施加 L1 惩罚：

$$\mathcal{L}_{\text{sparse}} = \frac{1}{L} \sum_{l=1}^L \|\mathbf{W}^{(l)}\|_1, \quad (7)$$

其中 L 是变压器层的数量， $\mathbf{W}^{(l)}$ 是第 l 层的图矩阵。这个正则化鼓励模型仅为具有结构意义的依赖项分配非零权重，并避免密集的注意力分配。

FD 对齐 FD 对齐损失（如公式 5 所示）指导模型通过增强依赖特征标记之间的强连接来遵循已知的特征级约束。

总训练损失 整体训练损失结合了语言建模目标、稀疏正则化和 FD 对齐：

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda_{\text{sparse}} \cdot \mathcal{L}_{\text{sparse}} + \lambda_{\text{FD}} \cdot \mathcal{L}_{\text{FD}}, \quad (8)$$

其中 λ_{sparse} 和 λ_{FD} 是超参数，分别控制稀疏性和依赖对齐的影响。

这种多目标方法使 GraDe 能够学习到既对语言模型有效又与底层表格数据依赖关系结构一致的表示。

我们还引入了 GraDe-Light，这是一种轻量级变体，仅更新动态图引导的注意模块，同时冻结所有其他参数。这种方法显著减少了可训练参数，同时保持了模型的大部分依赖建模能力。我们评估了这两种变体，以权衡计算效率和生成质量之间的取舍。

4 实验设置

模型 我们使用具有 1.24 亿个参数的 GPT-2 (?) 作为我们所有主要实验的基础模型。我们在该模型上评估了完整的 GraDe 和 GraDe-Light 变体。附录 ?? 中报告了使用 3.55 亿参数的 GPT-2 Medium 进行的额外实验。我们使用核采样 (?) 进行文本生成。

在我们的主要实验中，我们使用了五个来自不同领域的真实世界数据集，这些数据集具有复杂的依赖关系：鸟类学（生态）、疾病和糖尿病（医学）、收入（社会）以及房产（房地产）。数据集的大小从 700 到 40,000 个样本不等，特征数量为 5 到 30 个，包括连续、分类和整数类型。对于扩展实验 (§7.1)，我们使用了贷款数据集，这是一个大规模金融数据集，包含约 250,000 个样本。详细描述和数据集统计信息在附录 ?? 中提供。

FDs 提取 函数依赖检测在数据库领域是一个成熟且研究充分的问题 (?)。我们使用 HyFD (?)，这是一种用于发现函数依赖的混合算法，自动从我们的数据集提取函数依赖。我们使用 TANE (?) 验证结果。检测到的函数依赖用作表格数据建模的先验知识。关于检测算法的更多细节以及每个数据集提取的函数依赖详见附录 ??。

我们将 GraDe 与五种代表性的表格数据生成方法进行比较：基于 VAE 的 TVAE (?)，采用模式特定归一化；基于 GAN 的方法 CTGAN (?) 和 CTABGAN+ (?)，其中 CTGAN 使用模式特定归一化来处理复杂分布，而 CTABGAN+ 引入下游损失和用于混合变量的专用编码器；基于扩散的方法 TabSyn (?)，在潜在空间中操作以捕捉列间关系；以及基于 LLM 的方法 GReaT (?)。

根据之前关于表格数据生成的研究 (???)，我们从三个关键维度评估合成数据质量：(1) 实用性：我们评估合成数据能否通过在合成数据上训练模型并在真实测试数据上评估它们来有效替代真实数据 (?)；(2) 保真性：我们衡量属性间统计依赖关系的保留情况，这通常比匹配个体分布更为重要；以及 (3) 隐私性：我们计算合成样本与其在原始数据集中最近的记录之间的距离，验证生成的数据是否类似于但不复制训练样本。对于所有评估，我们报告五次运行不同随机种子的平均结果以确保可重复性。

Dataset		Original	TVAE	CTGAN	CTABGAN+	TabSyn	GReaT	GraDe	GraDe-Light
Bird (↑)	DT	99.97 ±0.02	70.59 ±0.00	94.28 ±0.17	66.46 ±0.21	99.75 ±0.06	99.63 ±0.07	99.70 ±0.00	99.78 ±0.02
	RF	100.00 ±0.00	79.87 ±0.46	99.09 ±0.08	75.54 ±0.11	99.97 ±0.12	100.00 ±0.00	100.00 ±0.00	100.00 ±0.00
	LR	100.00 ±0.00	87.04 ±0.00	100.00 ±0.00	73.99 ±0.00	100.00 ±0.00	98.29 ±0.00	100.00 ±0.00	98.85 ±0.00
Sick (↑)	DT	98.70 ±0.10	95.87 ±0.26	86.28 ±1.20	92.40 ±0.42	74.65 ±1.05	91.42 ±0.61	96.05 ±0.42	92.19 ±0.37
	RF	98.46 ±0.18	95.34 ±0.24	94.97 ±0.00	94.99 ±0.26	96.21 ±0.34	96.40 ±0.15	98.04 ±0.15	97.01 ±0.22
	LR	89.54 ±0.00	94.57 ±0.00	58.01 ±0.00	82.65 ±0.00	66.23 ±0.00	83.31 ±0.00	90.60 ±0.00	91.72 ±0.00
Income (↑)	DT	81.14 ±0.03	80.28 ±0.12	79.87 ±0.17	76.61 ±0.31	80.73 ±0.13	79.09 ±0.14	80.94 ±0.19	79.63 ±0.12
	RF	85.15 ±0.15	82.62 ±0.13	82.25 ±0.13	83.38 ±0.18	80.22 ±0.11	83.68 ±0.12	84.05 ±0.06	83.65 ±0.19
	LR	80.45 ±0.00	78.99 ±0.00	79.19 ±0.00	77.07 ±0.00	81.14 ±0.00	80.19 ±0.00	80.94 ±0.00	80.02 ±0.00
Diabetes (↑)	DT	74.68 ±1.30	72.21 ±0.49	59.48 ±0.88	62.99 ±1.23	75.06 ±2.08	65.97 ±0.66	78.05 ±0.66	67.08 ±1.40
	RF	74.94 ±0.88	75.97 ±0.71	57.01 ±1.50	64.81 ±1.50	75.32 ±0.92	67.40 ±0.76	77.27 ±1.42	71.56 ±1.76
	LR	69.48 ±0.00	73.38 ±0.00	57.14 ±0.00	72.08 ±0.00	70.78 ±0.00	67.53 ±0.00	68.83 ±0.00	66.88 ±0.00
Housing (↓)	DT	0.24 ±0.01	0.35 ±0.00	0.50 ±0.00	0.39 ±0.00	0.30 ±0.01	0.34 ±0.00	0.27 ±0.00	0.31 ±0.01
	RF	0.18 ±0.00	0.30 ±0.01	0.37 ±0.00	0.28 ±0.01	0.22 ±0.00	0.24 ±0.01	0.21 ±0.01	0.23 ±0.00
	LR	0.29 ±0.00	0.33 ±0.00	0.35 ±0.00	0.29 ±0.00	0.29 ±0.00	0.30 ±0.00	0.31 ±0.00	0.31 ±0.00

Table 1: 机器学习效率实验。最佳结果标记在红色中。“Original”指的是在原始真实数据上进行训练。分类准确率 (↑) 和回归平均绝对百分误差 (↓) 已报告。

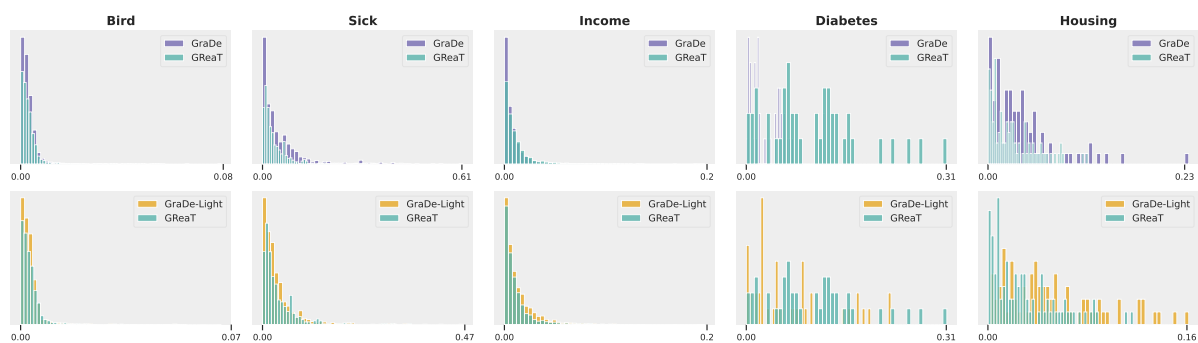


Figure 3: 五个数据集之间的相关性误差直方图。每个条形代表真实数据和合成数据相关系数的绝对差值，值越接近零表示保真度越高。GraDe (紫色) 和 GraDe-Light (黄色) 的误差分布相比 GReaT (绿色) 更集中于零附近。

5 结果

5.1 合成数据质量

我们通过在生成的数据集上训练判别模型并在真实测试集中测试其性能来评估合成数据的实际效用。此设置反映了合成数据是否能够在下游学习任务中有效替代真实数据。为了确保在模型类别上的通用性，我们使用了三种常用分类器：决策树 (DT)、随机森林 (RF) (?), 以及线性/逻辑回归 (LR)。表 1 显示了与原始训练数据集上的模型相比的各数据集的 MLE 分数。GraDe 展示了强大的性能，尤其是在复杂医学数据集上相较于 GReaT (我们最接近的基于 LLM 的竞争对手) 有显著的改进，在糖尿病数据集上实现了多达 12 % 的提升，在病情 (Sick) 数据集上则有 5 % 的提升。这些结果突出显示了在具有复杂结构关系的领域中显式依赖建模的优势。我们的 GraDe-Light 变体在使用比 GReaT 少 1 亿个可训练参数的情况下实现了类似的性能，突显出了我们图引导方法的高效性。

列相关性误差 特征关系的保持对于生成真实的表格数据至关重要。遵循 ?，我们通过测量真实数据与合成数据中皮尔逊相关系数的绝对差异来评估忠实度。对于类别特征，我们首先应用独热编码，然后计算相关性。图 3 显示了比较 GraDe、GraDe-Light 和 GReaT 在不同数据集上的相关性误差直方图。我们将此次比较集中在基于 LLM 的方法上，因为以往的工作表明，自动回归模型通常难以捕捉联合分布 (?), 使这对基于 LLM 的表格数据生成而言成为一个特别具有挑战性的方面。我们的结果表明，GraDe 的图引导方法通过显式建模功能依赖性并将注意力引向结构上相关的连接，提供了明显的优势。

距最近记录的距离 (DCR) 隐私保护是合成表格数据的关键应用，要求生成的样本要类似但不重复训练示例。根据 ?，我们使用 L1 范数来测量每个合成样本与真实训练示例之间的最小距离。对于数值特征，差异通过观测到的特征范围进行归一化；对于分类特征，如果值匹配，则距离定义为 0，否则为 1。最终的 DCR 分数计算为所有合成样本的最小距离的平均

Method	Bird	Sick	Income	Diabetes	Housing	Average
TVAE	.434 $\pm$.032	.262 $\pm$.023	.156 $\pm$.025	.360 $\pm$.010	.165 $\pm$.006	.275 $\pm$.019
CTGAN	.180 $\pm$.022	.702 $\pm$.040	.560 $\pm$.044	.783 $\pm$.020	.259 $\pm$.010	.497 $\pm$.027
CTABGAN+	.279 $\pm$.027	.386 $\pm$.032	.533 $\pm$.036	.516 $\pm$.014	.197 $\pm$.008	.382 $\pm$.023
TabSyn	.008 $\pm$.004	.202 $\pm$.019	.397 $\pm$.033	.385 $\pm$.014	.152 $\pm$.006	.229 $\pm$.015
GReaT	.010 $\pm$.005	.098 $\pm$.011	.165 $\pm$.059	.368 $\pm$.012	.117 $\pm$.004	.152 $\pm$.018
GraDe	.002 $\pm$.001	.081 $\pm$.009	.161 $\pm$.019	.323 $\pm$.012	.103 $\pm$.003	.134 $\pm$.009
GraDe-Light	.007 $\pm$.004	.145 $\pm$.065	.147 $\pm$.018	.355 $\pm$.014	.112 $\pm$.003	.153 $\pm$.021

Table 2: 与最近记录的距离 (DCR) 得分 (越低越好)。最好的结果用 红色 标记。

值。表 2 报告了所有五个数据集的 DCR 分数。虽然大多数模型在较简单的数据集上表现尚可，但我们的方法在具有复杂依赖关系的医学数据集（如 Sick, Diabetes）上显示了显著的改进。这一增强在敏感领域中特别有价值，因为较低的 DCR 分数表明与训练记录的相似性较小，并且与降低的成员推断风险相关联 (?)。这些结果展示了显式依赖建模如何使语言模型更好地平衡准确性和隐私性。

5.2 低数据环境下的建模

数据导向的生成方法，如 GAN 或基于扩散的方法，通常在数据稀缺的场景中表现不佳，需要大量的训练样本来捕捉复杂的分布 (?)。相反，基于 LLM 的方法可以利用丰富的预训练知识来有效地对表格数据进行建模，即使样本有限 (??)。我们通过使用 250 到 4000 个示例的训练集在 Bird 和 Income 数据集上评估这一能力，将 GraDe 与 GReaT 在相同的数据子集上进行比较。图 4 显示，GraDe 在所有数据方案中均持续优于 GReaT，在极度稀缺（250 个示例）情况下的提升幅度高达 15 %。虽然随着数据的增加差距缩小，但 GraDe 始终保持其优势。尽管参数较少，GraDe-Light 也实现了相当的改进。我们在 Income 数据集上观察到更明显的改进，其中提取的功能依赖关系更多且结构更复杂。这种经验模式表明，当底层依赖关系更丰富或难以隐式学习时，GraDe 从结构指导中受益更多。

我们通过分别去除两个组件的消融实验评估我们方法中每个组件的贡献：(1) GraDe w/o Sparsity，去除了图稀疏性正则化，和 (2) GraDe w/o FDs，去除了 FD 对齐损失。表格 3 对比了这些变体与完整的 GraDe 模型和 GReaT 在所有数据集上使用由生成数据训练的决策树的性能。

这两个组件提升了 GraDe 的有效性，其中 FD 对齐具有更强的影响。去除 FD 对齐会显著降低医疗数据集（如 Sick 和 Diabetes）的性能，因为这些数据集中数值属性和模糊的特征名称使得隐性学习变得困难。对于具有更清晰依赖

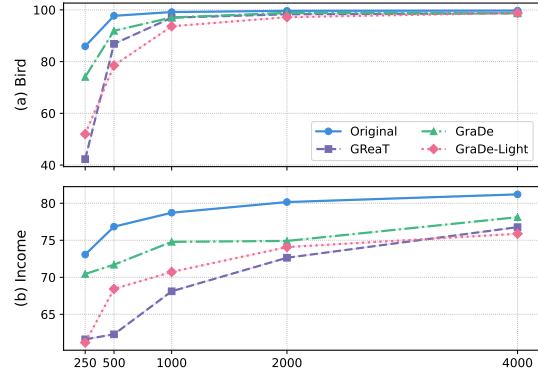


Figure 4: 在低数据情况下，不同训练样本规模的分类准确率。"Original" 表示在原始真实数据上训练的模型；其他则是在合成数据上训练的模型。

	Bird Acc $\uparrow$	Sick Acc $\uparrow$	Income Acc $\uparrow$	Diabetes Acc $\uparrow$	Housing MAPE $\downarrow$
GraDe	99.70	96.05	80.94	78.05	0.27
w/o Sparsity	+0.23	-0.40	-0.51	-0.63	+0.01
w/o FDs	-0.03	-6.51	-0.51	-5.47	-0.03
GReaT	-0.07	-4.63	-1.85	-12.08	-0.07

Table 3: 在五个数据集上进行消融研究，比较 GraDe、其消融变体和 GReaT 基线。分类任务使用准确率 (Acc)，回归任务使用平均绝对百分比误差 (MAPE)。

关系的数据集（如 Bird、Income、Housing），影响较小，并且去除稀疏正则化有时会提高性能，可能是因为在强局部模式的数据集中，注意力分布受益于增加的灵活性。这些结果表明，显式依赖指导对复杂数据结构最有价值，而这两个组件协同作用以优化整体性能，为我们的假设提供了实证支持，即密集注意可能会削弱对关键关系的关注。

6 结论

在这项工作中，我们解决了 LLMs 密集注意模式与表格数据稀疏依赖之间的结构不匹配问题。我们的解决方案 GraDe，通过可学习的依赖图引导 token 交互，将功能关系整合到注意力机制中。这种方法帮助模型在不改变基础架构的情况下，关注有意义的连接。跨多种数据集的实验显示出在实用性、保真度和隐私方面的一致性提升。GraDe 在复杂的医疗数据集上相比之前的 LLM 方法有高达 12 % 的优越表现，即使只有 250 个训练样本也仍然有效。轻量级的 GraDe-Light 变体在减少近一亿个可训练参数的同时获得了类似的结果。这些发现证实了结构性归纳偏差显著增强了 LLM 在表格建模中的表现。通过弥合强大的语义表示与特定领域结构意识之间的差距，GraDe 为生成高质量合成表格数据提供了实用且有效的解决方案。

7 局限性

尽管 GraDe 通过函数依赖有效增强了表格数据建模，但仍存在一些限制。首先，极高维度的表格可能会超过大型语言模型的上下文限制并增加计算需求。尽管 GraDe-Light 减少了参数要求，宽表格的顺序生成仍然效率低下。由于 GraDe 适用于任何仅解码的大型语言模型，未来具有扩展上下文能力的模型可能自然会解决这些扩展限制。

其次，如附录 ?? 所讨论的，我们的方法依赖于外部提取的函数依赖，这可能会继承数据集特定的偏差或忽略细微的关系。然而，自动 FD 提取能够实现可扩展且与领域无关的结构监督，而不需要昂贵的专家注释。虽然 GraDe 通过软约束和隐式建模来缓解潜在的偏差，但关键应用可能仍然会从专家验证或混合方法中受益，以进一步增强鲁棒性。

最后，我们观察到在训练过程中相较于基线模型收敛速度更慢。这是因为我们的图引导注意模块是从零开始训练的，而语言模型组件受益于广泛的预训练。这些图结构需要额外的训练迭代才能有效地与预训练的表达对齐。未来的工作可以探索迁移学习策略或改进初始化技

术，以加速图组件的收敛。

作者仅使用 ChatGPT 来润色最终稿件中的文本。

这部分通过补充实验扩展了我们的评估，提供了 GraDe 在不同场景下效果的更深入见解。具体来说，我们在一个大规模不平衡数据集上评估了该方法 (§7.1)，测量生成数据中内在约束的保持情况 (§7.2)，并研究了其对于更大骨干模型的可扩展性 (§??)。此外，我们使用基于鉴别器的度量评估数据保真度 (§??)，检查对功能依赖权重敏感性的敏感性 (§??)，并评估稀疏约束对高维数据的影响 (§??)。

7.1 扩展到大规模不平衡数据

实际的大规模数据集，尤其是在贷款违约预测等金融领域，通常表现出严重的不平衡问题，少数类通常是至关重要的却代表性不足。为了评估 GraDe 在这些挑战上的表现，我们在一个大规模贷款数据集上进行了额外的机器学习效率实验，该数据集约包含 250,000 个样本，但仅有 12 个% 的正例是基于客户行为预测贷款违约的。这项实验直接评估了在应用于大规模不平衡数据集时由不同方法生成的合成数据的实用性。

表 4 给出了 AUC 和 F1 分数作为评估指标，与准确率相比，这些指标更能代表在不平衡数据上的性能。结果揭示了不同方法之间显著的性能模式。TVAE 完全未能生成少数类样本，可能是由于后验坍塌(?)。虽然 CTGAN 和 CTABGAN+ 设法捕捉到了一些少数类模式，但他们的性能仍然有限，AUC 分数仅略高于随机概率。TabSyn 表现出强大的性能，特别是在 RF 和 LR 模型中，利用其潜在空间建模方法。GraDe 在 DT 模型中取得最佳性能，在其他分类器中仍具有竞争力，优于其他基于 LLM 的方法。这些结果突显了 GraDe 在捕捉复杂依赖关系上的有效性，即使在具有挑战性的类不平衡场景中，准确建模少数类对于下游应用是至关重要的。

7.2 内在约束保真度

根据 ?，我们评估合成数据在多大程度上保留了现实世界数据集中自然存在的内在约束。表 5 报告了三个具有代表性的约束的违反率：地理坐标与州界匹配（鸟类），教育水平与教育代码匹配（收入），以及房屋坐标落在加州边界内（住房）。

GraDe 以 1.12 % 的最低平均违规率显著优于传统生成器。这种卓越的性能源于我们对功能依赖关系的明确建模，它捕捉了特征之间的结构关系。虽然 GReaT 的表现也不错（1.88 % 违规），但我们的图引导方法提供了更一致的

约束维护。这些结果证实 GraDe 在保持统计特性和事实一致性方面，维护了高保真度合成数据的关键方面。

原则上，GraDe 可以应用于任何仅解码语言模型。为了研究模型规模对我们方法的影响，我们使用具有 3.55 亿参数的 GPT-2 Medium (?) 作为骨干模型进行额外的实验。表格 6 展示了在这个更大的模型下，GraDe 和 GraDe-Light 的 MLE 结果。这些结果补充了我们的主要发现，并证明了我们的方法在不同模型规模上的可扩展性。

为了进一步评估合成数据的真实性，超越在 §5 中报告的相关误差直方图，我们采用了一种基于鉴别器的评价方法。我们训练了一个支持向量机 (?)，使用 5 折交叉验证来区分真实和合成样本。对于高质量的合成数据，鉴别器的准确率应接近 50% (随机概率)，这表明分类器无法可靠地区分真实和合成分布。

表格 7 显示了我们生成的样本有效地混淆了在真实数据上训练的分类器。虽然 TabSyn 在接近 50 % 的准确率上实现了最接近理想的表现，但我们的方法优于其他基线。TabSyn 的优势可能源于其基于得分的潜在空间指导，使生成接近于模仿原始分布的高概率区域。然而，由于合成样本与真实样本之间显著的特征重叠，这种接近的近似可能引发隐私问题。

在我们的消融研究中，我们观察到删除功能依赖对齐损失对类似于 Sick (?) 的医学数据集有最显著的影响。这可能是因为此类数据集包含许多大量的数值特征和语义模糊的特征名称 (例如在 Sick 中的 “T3”，“TT4”)，使得语言模型在没有结构指导的情况下难以隐式捕捉依赖关系。

为了更好地理解这一效果，我们研究了不同的 FD 对齐损失权重 λ_{FD} (如方程 8 所示) 对挑战性数据集性能的影响。我们在 Sick 上训练 GraDe，采用不同的 λ_{FD} 值，并使用决策树分类器评估合成数据质量。图 ?? 展示了结果，其中虚线表示在原始数据上训练的模型。在 Sick 中确定的 λ_{FD} 最佳值为 0.1。

为了在更具挑战性的场景中进一步验证我们稀疏性约束的有效性，我们在 UCI 心肌梗死并发症 (MIC) 数据集 (?) 上进行了消融研究，该数据集包含 124 个用于预测患者并发症的特征。与我们最初针对 5 至 30 个特征的实验相比，这种高维设定提供了对稀疏性约束影响的更全面评估。

我们在四种不同配置下生成了合成数据，并使用决策树 (DT) 分类器评估性能，结果取五个随机种子的平均值。表格 8 显示了每个配置的分类准确率。

结果证实，在这个高维设置中，稀疏性约束

展示了显著更强的效果，相较于没有稀疏性的变体，性能提升了 3.23 个百分点。稀疏性和 FD 约束均证明是互补的，其中 FD 约束自身通过选择性地强调关键依赖关系，起到了额外的稀疏性的作用。这验证了我们的假设，即更大、更复杂的数据集能够更好地展示我们约束引导方法的优势。

所有模型在 PyTorch (?) 中实现，并使用来自 Huggingface (?) 的预训练语言模型作为我们的基础。我们修改了它们的源代码以集成我们的图引导注意力模块。对于基线模型，我们使用其各自作者推荐的超参数。所有实验均在具有 80GB 内存的 NVIDIA A100 GPU 上进行。

我们在所有数据集上对 GraDe 和 GraDe-Light 保持一致的超参数配置。我们使用 AdamW 优化器，学习率为 5×10^{-5} 。根据 GPU 内存的限制，我们使用固定的批量大小 64。稀疏性和功能依赖正则化权重 (λ_{sparse} 和 λ_{FD}) 分别设置为 0.001 和 0.1。对于采样步骤，我们将温度设置为 0.7，核采样参数设置为 0.95，适用于所有实验和数据集。并且按照 ?，我们使用正则表达式将生成的文本转换回表格式 (?)。

对于机器学习效率 (§5, §7.1, ?? 和 ??)，低数据场景 (§5.2)、消融研究 (??) 和判别器实验 (??)，我们额外使用了来自 Scikit-Learn 包的决策树 (DT)、随机森林 (RF)、线性/逻辑回归 (LR) 和支持向量机 (SVM) 模型 (?)。表 9 总结了我们的实验中所有模型使用的超参数设置。对于不平衡贷款数据集的缩放实验 (§7.1)，我们还为所有分类器启用了类别加权，以合理地解决类别不平衡问题。

为了提高参数效率，我们引入了 GraDe-Light，这是一种轻量级的变体，仅更新动态图引导的注意力模块，同时冻结所有其他参数。如表 10 所示，这种方法将 GPT-2 的可训练参数数量减少了 77.2%，对于 GPT-2 Medium 则减少了 71.6%。尽管减少了如此多的参数，GraDe-Light 在大多数数据集上仍保持了竞争力的性能，为受限的计算环境提供了一种实用的解决方案，同时保留了我们方法的核心优势。

我们在六个真实世界的数据集上评估我们的方法，其中五个用于我们的主要实验，两个用于我们的附加实验。以下是每个数据集的介绍：对于所有数据集，我们采用 80%/20% 的训练-测试拆分进行模型训练和评估。表 11 提供了所有数据集的综合统计信息。

为了将函数依赖作为结构先验纳入 GraDe 中，我们采用了数据分析文献中成熟的算法。具体来说，我们采用 HyFD (?) 作为我们的主要提取方法，并以 TANE (?) 作为验证参考。

	DT		RF		LR	
	AUC $\uparrow$	F1 $\uparrow$	AUC $\uparrow$	F1 $\uparrow$	AUC $\uparrow$	F1 $\uparrow$
Original	.739 $\pm$.000	.347 $\pm$.000	.814 $\pm$.002	.434 $\pm$.003	.635 $\pm$.000	.268 $\pm$.000
TVAE	-	-	-	-	-	-
CTGAN	.505 $\pm$.001	.211 $\pm$.001	.522 $\pm$.001	.210 $\pm$.002	.521 $\pm$.000	.210 $\pm$.000
CTABGAN+	.521 $\pm$.001	.217 $\pm$.000	.538 $\pm$.002	.222 $\pm$.001	.543 $\pm$.000	.224 $\pm$.000
TabSyn	<u>.571 $\pm$.001</u>	.246 $\pm$.001	<u>.645 $\pm$.001</u>	<u>.269 $\pm$.002</u>	<u>.594 $\pm$.000</u>	<u>.244 $\pm$.000</u>
GReaT	.569 $\pm$.000	.254 $\pm$.000	.622 $\pm$.002	.252 $\pm$.002	.553 $\pm$.000	.224 $\pm$.000
GraDe	<u>.587 $\pm$.001</u>	<u>.271 $\pm$.000</u>	<u>.642 $\pm$.003</u>	<u>.267 $\pm$.004</u>	<u>.581 $\pm$.000</u>	<u>.238 $\pm$.000</u>
GraDe-Light	.576 $\pm$.001	<u>.260 $\pm$.000</u>	.611 $\pm$.004	.245 $\pm$.002	.550 $\pm$.000	.218 $\pm$.000

Table 4: 关于大规模不平衡贷款数据集（约 25 万个样本，其中 12% 为正类）的机器学习实验的附加结果。最佳结果用 **红色** 标记，第二佳结果用 underline 标记。我们在此报告 AUC 和 F1 分数（ $\uparrow$ ）。

Method	Bird	Income	Housing	Average
	lat, long $\rightarrow$ State	education $\rightarrow$ education_num	latitude, longitude $\rightarrow$ CA	
TVAE	15.75 $\pm$ 0.63	18.76 $\pm$ 0.47	14.14 $\pm$ 0.53	16.22 $\pm$ 0.54
CTGAN	21.09 $\pm$ 0.70	20.64 $\pm$ 0.49	32.91 $\pm$ 0.72	24.88 $\pm$ 0.64
CTABGAN+	16.78 $\pm$ 0.65	14.35 $\pm$ 0.43	17.15 $\pm$ 0.57	16.09 $\pm$ 0.55
TabSyn	4.23 $\pm$ 0.35	0.00 $\pm$0.01	8.30 $\pm$ 0.42	4.18 $\pm$ 0.26
GReaT	1.98 $\pm$ 0.24	0.15 $\pm$ 0.05	3.51 $\pm$ 0.28	1.88 $\pm$ 0.19
GraDe	0.64 $\pm$0.14	0.08 $\pm$ 0.04	2.64 $\pm$0.24	1.12 $\pm$0.14
GraDe-Light	1.52 $\pm$ 0.21	0.13 $\pm$ 0.04	3.54 $\pm$ 0.28	1.73 $\pm$ 0.18

Table 5: 在合成数据中，内在约束的违反率（以%表示，得分越接近 0 % 越好），并附有 95 % 的置信区间。最佳结果用 **红色** 标记。

TANE (?) : 是一种经典的、基于格的 FD 发现算法，通过广度优先、逐层遍历系统地枚举所有属性组合。它计算元组的分区并详尽地验证所有候选依赖关系，保证了完整性和可靠性。虽然理论上严谨，但 TANE 的计算复杂度随着属性数量的增加呈指数增长，因此在特征众多或大规模数据集上使用并不实际。

HyFD (?) : 一种更新的混合方法，专为有效处理大规模、真实世界数据集而设计。该算法分为两个关键阶段：

在第一阶段（候选生成阶段），HyFD 采用基于采样的方法检查数据集的子集，以快速生成潜在 FD 候选的全面列表。具体来说，它选择随机的数据元组对，构建差异集（捕捉元组对之间的属性差异），并利用这些集合高效地修剪不可能或不太可能的 FD 候选。这种基于采样的策略相比于穷举列举大大降低了计算成本。

第二阶段（验证阶段）系统地验证先前识别的 FD 候选，通过在完整数据集上构建基于分区的等价类。HyFD 通过有效地精炼和裁剪这些分区，严谨地确认或舍弃每个候选 FD。该混合策略有效地平衡了计算效率和完整性，支持精确和近似的函数依赖，并确保在不同的数据条件下（包括噪声或稀疏性）具有高度的可靠性。

结构指导的实用选择 鉴于其在完整性和效率方面的强烈权衡，我们在整个实验中使用 HyFD 来提取结构先验。其可扩展性和对噪声的鲁棒性使其成为实际应用中的一个实用选择，在这些应用中，穷举枚举是不切实际的。

自动故障检测提取的挑战 一个潜在的担忧是在使用自动提取的功能依赖时，这些结构可能不完整或具有数据集特异性，这引发了一个问题：它们是否能在没有人工验证的情况下可靠地指导建模。然而，我们的方法并不将提取的功能依赖视为绝对真理，而是将其作为来自数据本身的有用结构提示。像 HyFD 这样的算法能够识别统计支持的模式，提升模型性能而无需专家输入。GraDe 的设计承认这些依赖可能是不完美的；在训练过程中，模型自然会学习哪些依赖是可靠的，哪些是不可靠的，并相应地调整它们的影响力。这使得我们的方法在需要耗费大量时间进行人工验证的现实应用中显得实用。

通过互补建模实现稳健性 GraDe 通过两种互补的策略来实现可靠性。首先，它采用了一种软约束形式，使得模型能够根据其观察到的可靠性灵活地整合依赖关系。这意味着 GraDe 可以自然地减少噪声或错误依赖的影响，同时从有效的依赖中获益。其次，我们的方法结合了两种结构信息的来源：从提取的 FDs 中获取的

Dataset		Original	TVAE	CTGAN	CTABGAN+	TabSyn	GReaT	GraDe	GraDe-Light
Bird (↑)	DT	99.97 ±0.02	70.59 ±0.00	94.28 ±0.17	66.46 ±0.21	99.75 ±0.06	99.63 ±0.07	99.76 ±0.02	99.78 ±0.00
	RF	100.00 ±0.00	79.87 ±0.46	99.09 ±0.08	75.54 ±0.11	99.97 ±0.12	100.00 ±0.00	100.00 ±0.00	99.96 ±0.02
	LR	100.00 ±0.00	87.04 ±0.00	100.00 ±0.00	73.99 ±0.00	100.00 ±0.00	98.29 ±0.00	100.00 ±0.00	98.35 ±0.00
Sick (↑)	DT	98.70 ±0.10	95.87 ±0.26	86.28 ±1.20	92.40 ±0.42	74.65 ±1.05	91.42 ±0.61	95.47 ±0.77	92.90 ±0.38
	RF	98.46 ±0.18	95.34 ±0.24	94.97 ±0.00	94.99 ±0.26	96.21 ±0.34	96.40 ±0.15	97.77 ±0.05	97.51 ±0.27
	LR	89.54 ±0.00	94.57 ±0.00	58.01 ±0.00	82.65 ±0.00	66.23 ±0.00	83.31 ±0.00	91.52 ±0.00	89.14 ±0.00
Income (↑)	DT	81.14 ±0.03	80.28 ±0.12	79.87 ±0.17	76.61 ±0.31	80.73 ±0.13	79.09 ±0.14	80.85 ±0.09	80.61 ±0.19
	RF	85.15 ±0.15	82.62 ±0.13	82.25 ±0.13	83.38 ±0.18	80.22 ±0.11	83.68 ±0.12	84.26 ±0.10	84.66 ±0.07
	LR	80.45 ±0.00	78.99 ±0.00	79.19 ±0.00	77.07 ±0.00	81.14 ±0.00	80.19 ±0.00	80.90 ±0.00	80.13 ±0.00
Diabetes (↑)	DT	74.68 ±1.30	72.21 ±0.49	59.48 ±0.88	62.99 ±1.23	75.06 ±2.08	65.97 ±0.66	76.62 ±0.88	72.21 ±1.12
	RF	74.94 ±0.88	75.97 ±0.71	57.01 ±1.50	64.81 ±1.50	75.32 ±0.92	67.40 ±0.76	73.90 ±0.95	75.32 ±0.71
	LR	69.48 ±0.00	73.38 ±0.00	57.14 ±0.00	72.08 ±0.00	70.78 ±0.00	67.53 ±0.00	69.48 ±0.00	68.18 ±0.00
Housing (↓)	DT	0.24 ±0.01	0.35 ±0.00	0.50 ±0.00	0.39 ±0.00	0.30 ±0.01	0.34 ±0.00	0.30 ±0.00	0.29 ±0.00
	RF	0.18 ±0.00	0.30 ±0.01	0.37 ±0.00	0.28 ±0.01	0.22 ±0.00	0.24 ±0.01	0.21 ±0.01	0.22 ±0.00
	LR	0.29 ±0.00	0.33 ±0.00	0.35 ±0.00	0.29 ±0.00	0.29 ±0.00	0.30 ±0.00	0.30 ±0.00	0.31 ±0.00

Table 6: 机器学习效率实验的其他结果。在此，GraDe 和 GraDe-Light 使用 GPT-2 Medium 作为基础模型。最佳结果标记在 红色。报告分类准确率 (↑) 和回归平均绝对百分比误差 (↓)。

Method	Bird	Sick	Income	Diabetes	Housing	Average
TVAE	74.95 ±0.25	76.22 ±2.26	67.27 ±1.21	67.59 ±6.58	60.37 ±0.97	69.28 ±2.26
CTGAN	59.90 ±0.98	81.54 ±2.66	62.58 ±0.81	85.50 ±3.22	62.89 ±0.77	70.48 ±1.69
CTABGAN+	63.77 ±1.35	64.35 ±2.13	58.49 ±0.42	60.75 ±2.47	60.41 ±0.44	61.55 ±1.36
TabSyn	49.32 ±0.37	54.82 ±2.74	55.41 ±0.85	45.36 ±2.79	50.48 ±0.71	51.08 ±1.49
GReaT	53.92 ±0.27	65.48 ±3.85	65.15 ±0.67	68.40 ±3.87	58.14 ±1.35	62.22 ±2.00
GraDe	52.33 ±0.67	62.33 ±1.62	59.49 ±0.85	59.28 ±2.37	56.54 ±0.80	57.99 ±1.26
GraDe-Light	53.43 ±0.74	63.44 ±2.66	64.64 ±0.90	66.45 ±2.61	58.40 ±0.47	61.27 ±1.48

Table 7: 判别器度量（准确性在%，越接近 50 % 的分数越好），具有 95 % 的置信区间。最佳结果标记为 红色，次佳结果为 underlined。

Method	DT Accuracy (%) ↑
GraDe (full)	89.41
GraDe w/o Sparsity	86.18
GraDe w/o FDs	87.35
w/o both	84.41

Table 8: 关于稀疏性约束在高维数据（124 个特征）中的有效性的消融研究。结果显示决策树的准确性是在 5 个随机种子上平均得出的。

DT	RF		LR	SVM
max_depth	max_depth	n_estimators	max_iter	kernel
20	20	100	500	linear

Table 9: 我们实验中使用的分类和回归模型的超参数设置。

统计模式和来自预训练 LLM 的语义理解。这创造了自然的弹性：当提取的依赖关系存在缺陷时，LLM 的语言理解可以弥补这些缺陷；当语言模式不清晰时，显式的依赖关系提供了指导。这种平衡的方法确保了 GraDe 在无需专家干预的情况下有效工作，同时对依赖提取过程中出现的不完美具有鲁棒性。

我们在此提供实验证明中识别和使用的 FDs 完整列表：

```
Bird
['lon', 'lat'] -> ['state_code', 'bird']
['lat'] -> ['lat_zone']
['state_code'] -> ['bird']
```

```
Sick
['TSH'] -> ['TSH_measured', 'TBG_measured', 'TBG']
['TSH', 'TT4'] -> ['hypopituitary']
['TSH', 'TT4', 'FTI'] -> ['I131_treatment', 'psych']
['TSH', 'TT4', 'FTI', 'T4U', 'query_on_thyroxine'] -> ['Class']
['TSH', 'TT4', 'FTI', 'age'] -> ['T4U']
['TSH', 'TT4', 'FTI', 'tumor', 'age', 'sex'] -> ['referral_source']
['TSH', 'TT4', 'FTI', 'age', 'T3_measured', 'sex'] -> ['referral_source']
['TSH', 'TT4', 'FTI', 'T3'] -> ['thyroid_surgery', 'lithium']
['TSH', 'TT4', 'FTI', 'T3',
```

Method	Model	# Parameters	# Trainable
GraDe	GPT-2	124.5M	124.5M
GraDe-Light	GPT-2	124.5M	28.4M
GraDe	GPT-2 Medium	355.0M	355.0M
GraDe-Light	GPT-2 Medium	355.0M	100.9M

Table 10: GraDe 和 GraDe-Light 在不同基础模型中的参数比较。# Parameters 和 # Trainable 分别表示总参数数目和可训练参数。参数计数是近似值，并包括动态图引导的注意力模块。

Dataset	Domain	# Samples	# Features	Task	# Classes	# Train [Pos %]	# Test [Pos %]
Bird (?)	Ecology	16,087	5	Classification	30	-	-
Sick (?)	Medical	3772	30	Classification	2	6.4 %	5.0 %
Income (?)	Social	32,561	15	Classification	2	24.2 %	23.6 %
Diabetes (From Kaggle ¹)	Medical	768	9	Classification	2	34.7 %	35.7 %
Housing (?)	Real Estate	20,640	10	Regression	-	-	-
Loan (From Kaggle ²)	Financial	252,000	13	Classification	2	12.3 %	12.4 %
MIC (?)	Medical	1,700	124	Classification	2	3.16 %	3.24 %

¹ <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

² <https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior>

Table 11: 我们实验中使用的真实世界数据集的统计信息。前五个数据集用于我们的主要实验。阴影部分的 Loan 和 MIC 数据集用于我们的附加实验 (§7.1, §??)。# Samples, # Features 和 # Classes 分别表示表格数据集中样本、特征和类别的数量。对于二元分类任务，我们报告训练集和测试集中正样本的百分比，以突出类别不平衡问题。

```

'query_on_thyroxine'] -> ['T4U']
['TSH', 'TT4', 'FTI', 'referral_source'] ->
['thyroid_surgery', 'lithium']
['TSH', 'TT4', 'FTI', 'T3_measured',
'referral_source', 'sex',
'query_on_thyroxine'] -> ['T4U']
['TSH', 'TT4', 'FTI', 'referral_source',
'sex', 'query_on_thyroxine', 'Class'] ->
['T4U']
['TSH', 'TT4', 'FTI', 'sex'] ->
['thyroid_surgery', 'lithium']
['TSH', 'TT4', 'FTI', 'on_thyroxine'] ->
['thyroid_surgery']
['TSH', 'TT4', 'FTI', 'T3_measured'] ->
['thyroid_surgery', 'Class']
['TSH', 'TT4', 'T4U'] -> ['I131_treatment',
'lithium']
['TSH', 'TT4', 'T4U', 'age'] -> ['FTI',
'psych', 'FTI_measured', 'Class']
['TSH', 'TT4', 'T4U', 'age', 'tumor', 'sex']
-> ['referral_source']
['TSH', 'TT4', 'T4U', 'age', 'T3_measured',
'sex'] -> ['referral_source']
['TSH', 'TT4', 'T4U', 'T3'] ->
['query_on_thyroxine', 'thyroid_surgery',
'psych']
['TSH', 'TT4', 'T4U', 'referral_source'] ->
['psych']
['TSH', 'TT4', 'T4U', 'referral_source',
'query_hypothyroid'] -> ['thyroid_surgery']
['TSH', 'TT4', 'T4U', 'sex'] ->
['thyroid_surgery']
['TSH', 'TT4', 'T4U', 'sex',
'query_on_thyroxine'] -> ['Class']
['TSH', 'TT4', 'T4U', 'sex', 'T3_measured']
-> ['Class']
['TSH', 'TT4', 'T4U', 'on_thyroxine',
'query_hypothyroid'] -> ['thyroid_surgery']
['TSH', 'TT4', 'T4U', 'query_on_thyroxine',
'sick'] -> ['Class']
['TSH', 'TT4', 'T4U', 'sick', 'T3_measured']
-> ['Class']
['TSH', 'TT4', 'goitre', 'T4U'] -> ['psych']
['TSH', 'TT4', 'T4U', 'T3_measured'] ->
['query_on_thyroxine', 'thyroid_surgery',
'FTI_measured']

['TSH', 'TT4', 'T4U', 'Class'] ->
['query_on_thyroxine']
['TSH', 'TT4', 'age'] ->
['query_on_thyroxine', 'I131_treatment',
'lithium']
['TSH', 'TT4', 'age', 'T3'] -> ['psych',
'T4U_measured', 'FTI_measured']
['TSH', 'TT4', 'age', 'T3', 'sex', 'sick']
-> ['referral_source']
['TSH', 'TT4', 'age', 'referral_source'] ->
['psych']
['TSH', 'TT4', 'age', 'referral_source',
'thyroid_surgery'] -> ['T4U_measured',
'FTI_measured']
['TSH', 'TT4', 'age', 'referral_source',
'query_hypothyroid'] -> ['T4U_measured',
'FTI_measured']
['TSH', 'TT4', 'age', 'referral_source',
'Class'] -> ['T4U_measured', 'FTI_measured']
['TSH', 'TT4', 'age', 'sex',
'thyroid_surgery'] -> ['T4U_measured',
'FTI_measured']
['TSH', 'TT4', 'age', 'sex',
'query_hypothyroid'] -> ['T4U_measured',
'FTI_measured']
['TSH', 'TT4', 'age', 'sex', 'Class'] ->
['T4U_measured', 'FTI_measured']
['TSH', 'TT4', 'age', 'thyroid_surgery'] ->
['Class']
['TSH', 'TT4', 'age', 'query_hypothyroid']
-> ['thyroid_surgery', 'Class']
['TSH', 'TT4', 'age', 'T4U_measured'] ->
['thyroid_surgery', 'FTI_measured', 'Class']
['TSH', 'TT4', 'FTI_measured', 'age'] ->
['thyroid_surgery', 'Class']
['TSH', 'TT4', 'Class', 'age'] ->
['thyroid_surgery']
['TSH', 'TT4', 'T3'] -> ['Class']
['TSH', 'TT4', 'T3', 'referral_source'] ->
['lithium']
['TSH', 'TT4', 'T3', 'referral_source',
'sex'] -> ['psych']
['TSH', 'TT4', 'T3', 'referral_source',
'query_hypothyroid'] -> ['thyroid_surgery']
['TSH', 'TT4', 'T3', 'referral_source',
'tumor'] -> ['thyroid_surgery']

```

```

['TSH', 'TT4', 'T3', 'referral_source',
'T4U_measured'] -> ['psych']
['TSH', 'TT4', 'T3', 'referral_source',
'FTI_measured'] -> ['psych']
['TSH', 'TT4', 'T3', 'sex'] ->
['thyroid_surgery', 'I131_treatment',
'lithium', 'T4U_measured', 'FTI_measured']
['TSH', 'TT4', 'T3', 'sex',
'on_antithyroid_medication'] -> ['psych']
['TSH', 'TT4', 'T3', 'on_thyroxine'] ->
['thyroid_surgery']
['TSH', 'TT4', 'referral_source', 'sex'] ->
['lithium']
['TSH', 'TT4', 'T3_measured',
'referral_source', 'sex',
'query_hypothyroid'] -> ['thyroid_surgery']
['TSH', 'TT4', 'T3_measured',
'referral_source', 'sex', 'T4U_measured']
-> ['thyroid_surgery']
['TSH', 'TT4', 'T3_measured',
'referral_source', 'sex', 'FTI_measured']
-> ['thyroid_surgery']
['TSH', 'TT4', 'T3_measured',
'referral_source', 'sex', 'Class'] ->
['thyroid_surgery']
['TSH', 'TT4', 'query_hyperthyroid',
'referral_source', 'T3_measured',
'on_thyroxine', 'T4U_measured'] ->
['thyroid_surgery']
['TSH', 'TT4', 'query_hyperthyroid',
'referral_source', 'T3_measured',
'on_thyroxine', 'FTI_measured'] ->
['thyroid_surgery']
['TSH', 'TT4', 'query_hyperthyroid',
'referral_source', 'T3_measured',
'on_thyroxine', 'Class'] ->
['thyroid_surgery']
['TSH', 'TT4', 'query_hyperthyroid',
'referral_source', 'T3_measured',
'query_hypothyroid'] -> ['thyroid_surgery']
['TSH', 'TT4', 'query_hyperthyroid',
'tumor', 'referral_source', 'T3_measured',
'T4U_measured'] -> ['thyroid_surgery']
['TSH', 'TT4', 'query_hyperthyroid',
'tumor', 'referral_source', 'T3_measured',
'FTI_measured'] -> ['thyroid_surgery']
['TSH', 'TT4', 'query_hyperthyroid', 'tumor',
'referral_source', 'T3_measured', 'Class'] ->
['thyroid_surgery']
['TSH', 'TT4', 'psych', 'sex'] ->
['lithium']
['TSH', 'TT4', 'T4U_measured',
'T3_measured'] -> ['FTI_measured']
['TSH', 'TT4', 'FTI_measured'] ->
['T4U_measured']
['TSH', 'FTI'] -> ['hypopituitary']
['TSH', 'FTI', 'T4U'] -> ['lithium']
['TSH', 'FTI', 'T4U', 'age'] ->
['query_on_thyroxine', 'psych', 'Class']
['TSH', 'FTI', 'T4U', 'age', 'T3',
'TT4_measured'] -> ['TT4']
['TSH', 'FTI', 'T4U', 'age', 'tumor', 'sex',
'TT4_measured'] -> ['referral_source']
['TSH', 'FTI', 'T4U', 'age', 'T3_measured',
'sex', 'TT4_measured'] -> ['referral_source']
['TSH', 'FTI', 'T4U', 'age', 'sex',
'TT4_measured'] -> ['TT4']
['TSH', 'FTI', 'T4U', 'age', 'on_thyroxine',
'TT4_measured'] -> ['TT4']
['TSH', 'FTI', 'T4U', 'age', 'T3_measured',
'TT4_measured'] -> ['TT4']
['TSH', 'FTI', 'T4U', 'T3'] ->
['query_on_thyroxine', 'thyroid_surgery',
'psych', 'Class']
['TSH', 'FTI', 'T4U', 'referral_source'] ->
['psych']
['TSH', 'FTI', 'T4U', 'referral_source',
'query_hyperthyroid'] -> ['thyroid_surgery']
['TSH', 'FTI', 'T4U', 'referral_source',
'T3_measured'] -> ['query_on_thyroxine']
['TSH', 'FTI', 'T4U', 'referral_source',
'Class'] -> ['query_on_thyroxine']
['TSH', 'FTI', 'T4U', 'sex',
'query_hyperthyroid'] -> ['thyroid_surgery']
['TSH', 'FTI', 'T4U', 'on_thyroxine',

```

```

'query_hyperthyroid'] -> ['thyroid_surgery']
['TSH', 'FTI', 'T4U', 'query_hyperthyroid',
'T3_measured'] -> ['query_on_thyroxine',
'thyroid_surgery']
['TSH', 'FTI', 'T4U', 'query_hyperthyroid',
'Class'] -> ['query_on_thyroxine']
['TSH', 'FTI', 'T4U', 'psych',
'T3_measured'] -> ['query_on_thyroxine']
['TSH', 'FTI', 'T4U', 'psych', 'Class'] ->
['query_on_thyroxine']
['TSH', 'FTI', 'age'] -> ['lithium',
'T4U_measured']
['TSH', 'FTI', 'age', 'T3'] -> ['psych']
['TSH', 'FTI', 'age', 'T3', 'sex'] ->
['referral_source']
['TSH', 'FTI', 'age', 'T3', 'sex',
'on_thyroxine'] -> ['T4U']
['TSH', 'FTI', 'age', 'T3', 'sex',
'on_thyroxine', 'TT4_measured'] -> ['TT4']
['TSH', 'FTI', 'age', 'T3', 'sex',
'on_antithyroid_medication'] -> ['T4U']
['TSH', 'FTI', 'age', 'T3', 'sex',
'TT4_measured', 'on_antithyroid_medication']
-> ['TT4']
['TSH', 'FTI', 'age', 'T3', 'sex',
'query_hypothyroid'] -> ['T4U']
['TSH', 'FTI', 'age', 'T3', 'sex',
'TT4_measured', 'query_hypothyroid'] ->
['TT4']
['TSH', 'query_hyperthyroid', 'FTI', 'age',
'T3', 'sex'] -> ['T4U']
['TSH', 'query_hyperthyroid', 'FTI', 'age',
'T3', 'sex', 'TT4_measured'] -> ['TT4']
['TSH', 'FTI', 'goitre', 'age', 'T3',
'on_thyroxine'] -> ['T4U']
['TSH', 'FTI', 'goitre', 'age', 'T3',
'on_thyroxine', 'TT4_measured'] -> ['TT4']
['TSH', 'FTI', 'goitre', 'age', 'T3',
'on_antithyroid_medication'] -> ['T4U']
['TSH', 'FTI', 'goitre', 'age', 'T3',
'TT4_measured', 'on_antithyroid_medication']
-> ['TT4']
['TSH', 'FTI', 'goitre', 'age', 'T3',
'query_hypothyroid'] -> ['T4U']
['TSH', 'FTI', 'goitre', 'age', 'T3',
'TT4_measured', 'query_hypothyroid'] ->
['TT4']
['TSH', 'query_hyperthyroid', 'FTI',
'goitre', 'age', 'T3'] -> ['T4U']
['TSH', 'query_hyperthyroid', 'FTI',
'goitre', 'age', 'T3', 'TT4_measured'] ->
['TT4']
['TSH', 'FTI', 'age', 'referral_source'] ->
['psych', 'Class']
['TSH', 'FTI', 'age', 'referral_source',
'sex'] -> ['I131_treatment']
['TSH', 'FTI', 'age', 'referral_source',
'on_thyroxine'] -> ['I131_treatment']
['TSH', 'FTI', 'age', 'referral_source',
'TT4_measured'] -> ['I131_treatment']
['TSH', 'FTI', 'age', 'sex'] ->
['query_on_thyroxine', 'Class']
['TSH', 'FTI', 'age', 'sex', 'on_thyroxine']
-> ['I131_treatment']
['TSH', 'FTI', 'age', 'sex', 'T3_measured']
-> ['psych']
['TSH', 'FTI', 'age', 'on_thyroxine',
'query_hypothyroid'] -> ['I131_treatment']
['TSH', 'FTI', 'age', 'on_thyroxine',
'query_hyperthyroid'] -> ['I131_treatment']
['TSH', 'FTI', 'age', 'pregnant',
'query_hyperthyroid'] ->
['query_on_thyroxine']
['TSH', 'FTI', 'age', 'pregnant', 'tumor']
-> ['query_on_thyroxine']
['TSH', 'FTI', 'age', 'T3_measured'] ->
['query_on_thyroxine']
['TSH', 'FTI', 'T3'] -> ['I131_treatment']
['TSH', 'FTI', 'T3', 'referral_source'] ->
['lithium', 'Class']
['TSH', 'FTI', 'T3', 'referral_source',
'on_thyroxine'] -> ['thyroid_surgery']
['TSH', 'FTI', 'T3', 'on_thyroxine', 'sick']
-> ['thyroid_surgery']
['TSH', 'query_hyperthyroid', 'FTI', 'T3']

```

```

-> ['Class']
['TSH', 'FTI', 'referral_source',
'thyroid_surgery', 'query_hyperthyroid']
-> ['lithium']
['TSH', 'TT4_measured', 'FTI',
'referral_source'] -> ['T4U_measured']
['TSH', 'FTI', 'Class', 'referral_source']
-> ['lithium']
['TSH', 'TT4_measured', 'FTI', 'sex'] ->
['T4U_measured']
['TSH', 'FTI', 'query_hypothyroid', 'goitre',
'TT4_measured'] -> ['T4U_measured']
['TSH', 'FTI', 'goitre', 'TT4_measured',
'Class'] -> ['T4U_measured']
['TSH', 'T4U', 'age'] -> ['thyroid_surgery',
'lithium']
['TSH', 'T4U', 'age', 'T3'] ->
['I131_treatment']
['TSH', 'T4U', 'age', 'T3', 'sex'] ->
['referral_source', 'psych']
['TSH', 'T4U', 'age', 'referral_source'] ->
['FTI_measured', 'Class']
['TSH', 'T4U', 'age', 'referral_source',
'sex', 'on_thyroxine'] -> ['psych']
['TSH', 'T4U', 'age', 'referral_source',
'sex', 'sick'] -> ['psych']
['TSH', 'query_hyperthyroid', 'T4U', 'age',
'referral_source', 'sex'] -> ['psych']
['TSH', 'T4U', 'age', 'sex'] ->
['I131_treatment']
['TSH', 'T4U', 'age', 'sex', 'on_thyroxine',
'FTI_measured'] -> ['psych']
['TSH', 'T4U', 'age', 'sex', 'FTI_measured',
'sick'] -> ['psych']
['TSH', 'query_hyperthyroid', 'T4U', 'age',
'sex', 'FTI_measured'] -> ['psych']
['TSH', 'on_thyroxine', 'T4U', 'age'] ->
['I131_treatment']
['TSH', 'T4U', 'age', 'psych'] ->
['FTI_measured']
['TSH', 'TT4_measured', 'T4U', 'age'] ->
['I131_treatment']
['TSH', 'T4U', 'T3'] -> ['FTI_measured']
['TSH', 'T4U', 'T3', 'referral_source'] ->
['thyroid_surgery', 'lithium', 'psych']
['TSH', 'T4U', 'tumor', 'T3',
'referral_source', 'sex'] ->
['I131_treatment']
['TSH', 'T4U', 'T3', 'sex'] -> ['Class']
['TSH', 'T4U', 'tumor', 'T3', 'sex',
'on_thyroxine'] -> ['I131_treatment']
['TSH', 'T4U', 'T3', 'sex',
'thyroid_surgery'] -> ['lithium']
['TSH', 'T4U', 'T3', 'on_thyroxine',
'thyroid_surgery', 'query_hypothyroid']
-> ['lithium']
['TSH', 'query_hyperthyroid', 'T4U', 'T3']
-> ['Class']
['TSH', 'lithium', 'T4U', 'T3'] ->
['thyroid_surgery']
['TSH', 'T4U', 'referral_source'] ->
['hypopituitary']
['TSH', 'T4U', 'referral_source', 'sex',
'T3_measured'] -> ['FTI_measured']
['TSH', 'T4U', 'psych', 'referral_source']
-> ['lithium']
['TSH', 'T4U', 'referral_source',
'T3_measured', 'TT4_measured'] ->
['FTI_measured']
['TSH', 'T4U', 'T3_measured', 'psych',
'sex', 'on_thyroxine', 'thyroid_surgery',
'query_hypothyroid'] -> ['lithium']
['TSH', 'T4U', 'T3_measured', 'psych', 'sex',
'query_on_thyroxine'] -> ['FTI_measured']
['TSH', 'query_hyperthyroid', 'T4U',
'T3_measured', 'psych', 'sex'] ->
['FTI_measured']
['TSH', 'on_thyroxine', 'T4U'] ->
['hypopituitary']
['TSH', 'query_on_thyroxine', 'T4U'] ->
['hypopituitary']
['TSH', 'T4U', 'psych', 'T3_measured',
'TT4_measured', 'query_on_thyroxine'] ->
['FTI_measured']
['TSH', 'query_hyperthyroid', 'T4U',

```

```

'psych', 'T3_measured', 'TT4_measured']
-> ['FTI_measured']
['TSH', 'T4U', 'T3_measured'] ->
['hypopituitary']
['TSH', 'age'] -> ['hypopituitary']
['TSH', 'age', 'T3'] -> ['Class']
['TSH', 'age', 'T3', 'referral_source'] ->
['lithium', 'psych']
['TSH', 'age', 'T3', 'referral_source',
'on_thyroxine', 'query_hypothyroid'] ->
['thyroid_surgery']
['TSH', 'age', 'T3', 'referral_source',
'sick', 'query_hypothyroid'] ->
['thyroid_surgery']
['TSH', 'age', 'T3', 'sex'] ->
['thyroid_surgery', 'lithium']
['TSH', 'FTI_measured', 'age', 'T3'] ->
['T4U_measured']
['TSH', 'T4U_measured', 'age',
'referral_source'] -> ['FTI_measured']
['TSH', 'FTI_measured', 'age',
'referral_source'] -> ['T4U_measured']
['TSH', 'FTI_measured', 'age', 'sex'] ->
['T4U_measured']
['TSH', 'T4U_measured', 'age', 'psych'] ->
['FTI_measured']
['TSH', 'TT4_measured', 'FTI_measured',
'age'] -> ['T4U_measured']
['TSH', 'T3'] -> ['hypopituitary']
['TSH', 'T3', 'TT4_measured', 'FTI_measured',
'Class'] -> ['T4U_measured']
['TSH', 'T4U_measured', 'T3'] ->
['FTI_measured']
['TSH', 'referral_source', 'sex',
'T3_measured', 'T4U_measured'] ->
['FTI_measured']
['TSH', 'referral_source', 'T3_measured',
'TT4_measured', 'T4U_measured'] ->
['FTI_measured']
['TSH', 'T3_measured', 'psych', 'sex',
'query_on_thyroxine', 'T4U_measured'] ->
['FTI_measured']
['TSH', 'query_hyperthyroid', 'T3_measured',
'psych', 'sex', 'T4U_measured'] ->
['FTI_measured']
['TSH', 'psych', 'T3_measured',
'TT4_measured', 'query_on_thyroxine',
'T4U_measured'] -> ['FTI_measured']
['TSH', 'query_on_thyroxine', 'Class'] ->
['hypopituitary']
['TSH', 'query_hyperthyroid', 'psych',
'T3_measured', 'TT4_measured',
'T4U_measured'] -> ['FTI_measured']
['TT4'] -> ['TT4_measured', 'TBG_measured',
'TBG']
['TT4', 'FTI'] -> ['hypopituitary',
'T4U_measured']
['TT4', 'FTI', 'age'] ->
['query_on_thyroxine', 'thyroid_surgery',
'I131_treatment', 'lithium', 'Class']
['TT4', 'FTI', 'age', 'T3'] -> ['T4U',
'psych']
['TT4', 'FTI', 'age', 'T3',
'referral_source', 'sex', 'TSH_measured',
'query_hypothyroid'] -> ['TSH']
['TT4', 'FTI', 'age', 'T3_measured',
'referral_source', 'sex'] -> ['T4U']
['TT4', 'FTI', 'age', 'psych',
'referral_source', 'T3_measured'] -> ['T4U']
['TT4', 'FTI', 'age', 'sex'] -> ['psych']
['TT4', 'FTI', 'age', 'T3_measured', 'sex',
'sick'] -> ['T4U']
['TT4', 'FTI', 'age', 'psych', 'T3_measured',
'sick'] -> ['T4U']
['TT4', 'FTI', 'T3'] -> ['Class']
['TT4', 'FTI', 'T3', 'referral_source'] ->
['lithium']
['TT4', 'FTI', 'T3', 'referral_source',
'sex'] -> ['psych']
['TT4', 'FTI', 'T3', 'TSH_measured'] ->
['lithium']
['TT4', 'FTI', 'referral_source', 'sex',
'Class'] -> ['lithium']
['TT4', 'FTI', 'referral_source',
'on_thyroxine', 'Class', 'sick',

```

```

'I131_treatment'] -> ['lithium']
['query_hyperthyroid', 'TT4', 'FTI',
'referral_source', 'on_thyroxine', 'Class',
'I131_treatment'] -> ['lithium']
['TT4', 'FTI', 'referral_source',
'T3_measured', 'Class', 'sick',
'I131_treatment'] -> ['lithium']
['query_hyperthyroid', 'TT4', 'FTI',
'referral_source', 'T3_measured', 'Class',
'I131_treatment'] -> ['lithium']
['TT4', 'T4U'] -> ['hypopituitary']
['TT4', 'T4U', 'age'] -> ['thyroid_surgery',
'I131_treatment', 'lithium']
['TT4', 'T4U', 'age', 'T3'] -> ['FTI',
'query_on_thyroxine', 'psych']
['TT4', 'T4U', 'age', 'T3',
'referral_source', 'sex', 'TSH_measured',
'query_hypothyroid'] -> ['TSH']
['TT4', 'T4U', 'age', 'referral_source'] ->
['psych', 'Class']
['TT4', 'T4U', 'age', 'sex'] -> ['Class']
['on_thyroxine', 'TT4', 'T4U', 'age'] ->
['Class']
['TT4', 'T4U', 'age', 'query_hyperthyroid',
'T3_measured'] -> ['FTI_measured']
['TT4', 'T4U', 'age', 'TSH_measured'] ->
['FTI_measured']
['TT4', 'T4U', 'age', 'T3_measured'] ->
['query_on_thyroxine']
['TT4', 'T4U', 'T3'] -> ['Class']
['TT4', 'T4U', 'T3', 'referral_source'] ->
['lithium']
['TT4', 'T4U', 'T3', 'referral_source',
'sex'] -> ['psych']
['TT4', 'T4U', 'T3', 'referral_source',
'query_hyperthyroid'] -> ['psych']
['TT4', 'T4U', 'T3', 'sex'] -> ['lithium']
['TT4', 'T4U', 'T3', 'sex', 'on_thyroxine']
-> ['thyroid_surgery']
['TT4', 'T4U', 'T3', 'sex',
'query_hypothyroid'] -> ['thyroid_surgery']
['TT4', 'T4U', 'T3', 'query_on_thyroxine']
-> ['lithium']
['query_hyperthyroid', 'TT4', 'T4U', 'T3']
-> ['lithium']
['TT4', 'T4U', 'T3', 'TSH_measured'] ->
['query_on_thyroxine', 'lithium']
['TT4', 'T4U', 'psych', 'referral_source',
'sex', 'Class'] -> ['lithium']
['TT4', 'T4U', 'psych', 'referral_source',
'on_thyroxine', 'Class', 'sick',
'I131_treatment'] -> ['lithium']
['query_hyperthyroid', 'TT4', 'T4U', 'psych',
'referral_source', 'on_thyroxine', 'Class',
'I131_treatment'] -> ['lithium']
['TT4', 'T4U', 'psych', 'referral_source',
'T3_measured', 'Class', 'sick',
'I131_treatment'] -> ['lithium']
['query_hyperthyroid', 'TT4', 'T4U', 'psych',
'referral_source', 'T3_measured', 'Class',
'I131_treatment'] -> ['lithium']
['TT4', 'age'] -> ['hypopituitary']
['TT4', 'age', 'T3'] -> ['thyroid_surgery',
'Class']
['TT4', 'age', 'T3', 'referral_source'] ->
['psych']
['TT4', 'age', 'T3', 'referral_source',
'sex'] -> ['I131_treatment']
['TT4', 'age', 'T3', 'referral_source',
'sex', 'TSH_measured'] -> ['T4U_measured',
'FTI_measured']
['TT4', 'age', 'T3', 'sex'] ->
['query_on_thyroxine']
['TT4', 'age', 'T3', 'sex', 'on_thyroxine']
-> ['I131_treatment']
['TT4', 'age', 'T3', 'sex', 'T4U_measured']
-> ['I131_treatment']
['TT4', 'age', 'T3', 'sex', 'FTI_measured']
-> ['I131_treatment']
['TT4', 'age', 'T3', 'query_hypothyroid'] ->
['query_on_thyroxine']
['TT4', 'age', 'T3', 'TSH_measured'] ->
['query_on_thyroxine']
['TT4', 'age', 'T3', 'TSH_measured',
'referral_source', 'sex', 'on_thyroxine',

```

```

'thyroid_surgery'] -> ['I131_treatment']
['TT4', 'age', 'query_hyperthyroid',
'T3_measured', 'T4U_measured'] ->
['FTI_measured']
['TT4', 'age', 'TSH_measured',
'T4U_measured'] -> ['FTI_measured']
['TT4', 'FTI_measured', 'age'] ->
['T4U_measured']
['TT4', 'T3'] -> ['hypopituitary']
['TT4', 'FTI_measured', 'T3'] ->
['T4U_measured']
['TT4', 'query_on_thyroxine'] ->
['hypopituitary']
['FTI'] -> ['FTI_measured', 'TGB_measured',
'TGB']
['FTI', 'T4U'] -> ['hypopituitary']
['FTI', 'T4U', 'age'] -> ['lithium']
['FTI', 'T4U', 'age', 'T3'] ->
['I131_treatment']
['FTI', 'T4U', 'age', 'T3',
'referral_source', 'sex', 'on_thyroxine',
'query_hypothyroid', 'TSH_measured'] ->
['TSH']
['FTI', 'T4U', 'age', 'T3',
'referral_source', 'sex', 'on_thyroxine',
'TT4_measured'] -> ['TT4']
['FTI', 'T4U', 'age', 'referral_source',
'sex'] -> ['psych', 'Class']
['FTI', 'T4U', 'age', 'referral_source',
'sick'] -> ['psych', 'Class']
['FTI', 'T4U', 'age', 'referral_source',
'query_hypothyroid'] -> ['psych', 'Class']
['FTI', 'T4U', 'age', 'referral_source',
'query_hyperthyroid'] ->
['query_on_thyroxine']
['FTI', 'T4U', 'age', 'referral_source',
'psych'] -> ['Class']
['FTI', 'T4U', 'age', 'referral_source',
'T3_measured'] -> ['psych', 'Class']
['FTI', 'T4U', 'age', 'referral_source',
'TT4_measured'] -> ['psych', 'Class']
['FTI', 'T4U', 'age', 'referral_source',
'Class'] -> ['psych']
['FTI', 'T4U', 'age', 'sex',
'query_hyperthyroid'] ->
['query_on_thyroxine']
['on_thyroxine', 'FTI', 'T4U', 'age'] ->
['I131_treatment']
['FTI', 'T4U', 'age', 'query_hyperthyroid',
'psych'] -> ['query_on_thyroxine']
['FTI', 'T4U', 'age', 'query_hyperthyroid',
'TSH_measured'] -> ['query_on_thyroxine']
['FTI', 'T4U', 'age', 'query_hyperthyroid',
'T3_measured'] -> ['query_on_thyroxine']
['FTI', 'T4U', 'T3', 'referral_source'] ->
['lithium']
['FTI', 'T4U', 'T3', 'referral_source',
'sex', 'sick'] -> ['psych']
['FTI', 'T4U', 'T3', 'referral_source',
'sex', 'query_hypothyroid'] -> ['psych']
['FTI', 'T4U', 'T3', 'referral_source',
'sex', 'Class'] -> ['psych']
['FTI', 'T4U', 'T3', 'referral_source',
'sick'] -> ['Class']
['query_hyperthyroid', 'FTI', 'T4U', 'T3',
'referral_source', 'sick'] -> ['psych']
['FTI', 'T4U', 'T3', 'referral_source',
'query_hypothyroid'] -> ['Class']
['query_hyperthyroid', 'FTI', 'T4U', 'T3',
'referral_source', 'query_hypothyroid'] ->
['psych']
['query_hyperthyroid', 'FTI', 'T4U', 'T3',
'referral_source', 'Class'] -> ['psych']
['FTI', 'T4U', 'T3', 'referral_source',
'psych'] -> ['Class']
['FTI', 'T4U', 'T3', 'sex'] -> ['lithium']
['FTI', 'T4U', 'T3', 'sex', 'TSH_measured']
-> ['query_on_thyroxine']
['query_on_thyroxine', 'FTI', 'T4U', 'T3']
-> ['lithium']
['query_hyperthyroid', 'FTI', 'T4U', 'T3']
-> ['lithium']
['FTI', 'T4U', 'T3', 'query_hyperthyroid',
'TSH_measured'] -> ['query_on_thyroxine']
['FTI', 'T4U', 'T3', 'TSH_measured'] ->

```

```

['lithium']
['query_hyperthyroid', 'FTI', 'T4U', 'psych',
'referral_source', 'sex', 'Class'] ->
['lithium']
['query_hyperthyroid', 'FTI', 'T4U', 'psych',
'referral_source', 'on_thyroxine', 'Class',
'I131_treatment'] -> ['lithium']
['query_hyperthyroid', 'FTI', 'T4U', 'psych',
'referral_source', 'T3_measured', 'Class',
'I131_treatment'] -> ['lithium']
['FTI', 'age'] -> ['hypopituitary']
['FTI', 'age', 'T3'] ->
['query_on_thyroxine', 'thyroid_surgery',
'T4U_measured', 'Class']
['FTI', 'age', 'T3', 'referral_source'] ->
['I131_treatment']
['FTI', 'age', 'T3', 'referral_source',
'psych'] -> ['lithium']
['FTI', 'age', 'T3', 'sex'] ->
['I131_treatment']
['FTI', 'age', 'psych', 'referral_source',
'sex', 'T3_measured'] -> ['lithium']
['query_hyperthyroid', 'FTI', 'age',
'psych', 'referral_source', 'T3_measured'] ->
['lithium']
['FTI', 'sick', 'age', 'T3_measured'] ->
['T4U_measured']
['TT4_measured', 'FTI', 'sick', 'age'] ->
['T4U_measured']
['FTI', 'Class', 'age', 'T3_measured'] ->
['T4U_measured']
['TT4_measured', 'FTI', 'Class', 'age'] ->
['T4U_measured']
['FTI', 'T3'] -> ['hypopituitary']
['TT4_measured', 'FTI', 'T3'] ->
['T4U_measured']
['FTI', 'referral_source', 'sex'] ->
['hypopituitary']
['on_thyroxine', 'FTI', 'tumor',
'referral_source'] -> ['hypopituitary']
['FTI', 'Class', 'sex'] -> ['hypopituitary']
['on_thyroxine', 'FTI', 'Class'] ->
['hypopituitary']
['query_on_thyroxine', 'FTI'] ->
['hypopituitary']
['T4U'] -> ['T4U_measured', 'TBG_measured',
'TBG']
['T4U', 'age', 'T3'] -> ['FTI_measured']
['T4U', 'age', 'T3', 'referral_source'] ->
['lithium', 'psych']
['T4U', 'age', 'T3', 'sex'] -> ['lithium',
'Class']
['T4U', 'age', 'T3', 'sex', 'TSH_measured']
-> ['I131_treatment']
['T4U', 'age', 'T3', 'query_hypothyroid'] ->
['Class']
['T4U', 'age', 'referral_source'] ->
['hypopituitary']
['T4U', 'age', 'psych', 'referral_source',
'sex', 'query_hypothyroid'] -> ['lithium']
['query_hyperthyroid', 'T4U', 'age',
'T3_measured', 'referral_source', 'sex']
-> ['FTI_measured']
['T4U', 'age', 'referral_source', 'sex',
'TSH_measured'] -> ['FTI_measured']
['query_hyperthyroid', 'T4U', 'age',
'referral_source', 'T3_measured',
'TT4_measured'] -> ['FTI_measured']
['T4U', 'age', 'referral_source',
'TSH_measured', 'TT4_measured'] ->
['FTI_measured']
['T4U', 'age', 'sex'] -> ['hypopituitary']
['query_hyperthyroid', 'T4U', 'age', 'psych',
'T3_measured', 'sex'] -> ['FTI_measured']
['T4U', 'age', 'sex', 'psych',
'TSH_measured'] -> ['FTI_measured']
['query_on_thyroxine', 'T4U', 'age'] ->
['hypopituitary']
['T4U', 'age', 'TSH_measured'] ->
['hypopituitary']
['T4U', 'age', 'T3_measured'] ->
['hypopituitary']
['T4U', 'T3', 'referral_source',
'query_hypothyroid'] -> ['hypopituitary']
['T4U', 'T3', 'sex'] -> ['hypopituitary']

```

```

['query_on_thyroxine', 'T4U', 'T3'] ->
['hypopituitary']
['query_on_thyroxine', 'T4U',
'referral_source'] -> ['hypopituitary']
['query_on_thyroxine', 'T4U',
'TSH_measured'] -> ['hypopituitary']
['Class', 'T4U'] -> ['hypopituitary']
['age'] -> ['TBG_measured', 'TBG']
['age', 'T3'] -> ['hypopituitary']
['age', 'T3', 'referral_source',
'on_thyroxine', 'FTI_measured',
'thyroid_surgery'] -> ['T4U_measured']
['age', 'T3', 'sex', 'on_thyroxine',
'FTI_measured', 'thyroid_surgery'] ->
['T4U_measured']
['TT4_measured', 'FTI_measured', 'age',
'T3'] -> ['T4U_measured']
['T4U_measured', 'age', 'T3'] ->
['FTI_measured']
['age', 'referral_source', 'sex'] ->
['hypopituitary']
['age', 'referral_source', 'sex',
'TT4_measured', 'FTI_measured', 'sick']
-> ['T4U_measured']
['query_hyperthyroid', 'age', 'T3_measured',
'referral_source', 'sex', 'T4U_measured'] ->
['FTI_measured']
['age', 'referral_source', 'sex',
'TSH_measured', 'T4U_measured'] ->
['FTI_measured']
['age', 'referral_source', 'sex',
'TT4_measured', 'FTI_measured', 'Class']
-> ['T4U_measured']
['query_on_thyroxine', 'age',
'referral_source'] -> ['hypopituitary']
['query_hyperthyroid', 'age',
'referral_source', 'T3_measured',
'TT4_measured', 'T4U_measured'] ->
['FTI_measured']
['age', 'referral_source', 'TSH_measured',
'TT4_measured', 'T4U_measured'] ->
['FTI_measured']
['query_on_thyroxine', 'age', 'sex'] ->
['hypopituitary']
['query_hyperthyroid', 'age', 'psych',
'T3_measured', 'sex', 'T4U_measured'] ->
['FTI_measured']
['age', 'sex', 'psych', 'TSH_measured',
'T4U_measured'] -> ['FTI_measured']
['query_on_thyroxine', 'age',
'TSH_measured'] -> ['hypopituitary']
['Class', 'age'] -> ['hypopituitary']
['T3'] -> ['T3_measured', 'TBG_measured',
'TBG']
['Class', 'T3'] -> ['hypopituitary']
['referral_source'] -> ['TBG_measured',
'TBG']
['referral_source', 'sex',
'query_on_thyroxine', 'goitre', 'Class']
-> ['hypopituitary']
['referral_source', 'query_on_thyroxine',
'sick', 'goitre', 'Class'] ->
['hypopituitary']
['sex'] -> ['TBG_measured', 'TBG']
['sex', 'query_on_thyroxine', 'goitre',
'TSH_measured', 'Class'] -> ['hypopituitary']
['on_thyroxine'] -> ['TBG_measured', 'TBG']
['query_on_thyroxine'] -> ['TBG_measured',
'TBG']
['query_on_thyroxine', 'sick', 'goitre',
'TSH_measured', 'Class'] -> ['hypopituitary']
['TBG_measured', 'TBG']
['sick'] -> ['TBG_measured', 'TBG']
['pregnant'] -> ['TBG_measured', 'TBG']
['thyroid_surgery'] -> ['TBG_measured',
'TBG']
['I131_treatment'] -> ['TBG_measured',
'TBG']
['query_hypothyroid'] -> ['TBG_measured',
'TBG']
['query_hyperthyroid'] -> ['TBG_measured',
'TBG']
['lithium'] -> ['TBG_measured', 'TBG']
['goitre'] -> ['TBG_measured', 'TBG']

```

```

['tumor'] -> ['TBG_measured', 'TBG']
['hypopituitary'] -> ['TBG_measured', 'TBG']
['psych'] -> ['TBG_measured', 'TBG']
['TSH_measured'] -> ['TBG_measured', 'TBG']
['T3_measured'] -> ['TBG_measured', 'TBG']
['TT4_measured'] -> ['TBG_measured', 'TBG']
['T4U_measured'] -> ['TBG_measured', 'TBG']
['FTI_measured'] -> ['TBG_measured', 'TBG']
['Class'] -> ['TBG_measured', 'TBG']
['TBG_measured'] -> ['TBG']
['TBG'] -> ['TBG_measured']

```

Income

```

['fnlwgt', 'capital-gain', 'hours-per-week',
'capital-loss', 'education', 'workclass',
'relationship'] -> ['gender']
['fnlwgt', 'capital-gain', 'hours-per-week',
'capital-loss', 'education-num', 'workclass',
'relationship'] -> ['gender']
['fnlwgt', 'capital-gain', 'hours-per-week',
'age', 'education', 'occupation',
'workclass', 'relationship'] -> ['income']
['fnlwgt', 'capital-gain', 'hours-per-week',
'age', 'education-num', 'occupation',
'workclass', 'relationship'] -> ['income']
['fnlwgt', 'capital-gain', 'hours-per-week',
'age', 'occupation', 'relationship'] ->
['marital-status']
['fnlwgt', 'capital-gain', 'hours-per-week',
'education', 'workclass', 'relationship',
'race'] -> ['gender']
['fnlwgt', 'capital-gain', 'hours-per-week',
'education', 'workclass', 'relationship',
'income'] -> ['gender']
['fnlwgt', 'capital-gain', 'hours-per-week',
'education-num', 'workclass', 'relationship',
'race'] -> ['gender']
['fnlwgt', 'capital-gain', 'hours-per-week',
'education-num', 'workclass', 'relationship',
'income'] -> ['gender']
['fnlwgt', 'hours-per-week', 'age',
'education'] -> ['race']
['fnlwgt', 'hours-per-week', 'age',
'education', 'occupation', 'workclass']
-> ['capital-loss']
['fnlwgt', 'hours-per-week', 'age',
'education', 'occupation', 'relationship'] ->
['marital-status']
['fnlwgt', 'hours-per-week', 'age',
'education-num'] -> ['race']
['fnlwgt', 'hours-per-week', 'age',
'education-num', 'occupation', 'workclass']
-> ['capital-loss']
['fnlwgt', 'hours-per-week', 'age',
'education-num', 'occupation',
'relationship'] -> ['marital-status']
['fnlwgt', 'hours-per-week', 'age',
'occupation', 'workclass'] -> ['race']
['fnlwgt', 'hours-per-week', 'age',
'workclass', 'marital-status',
'relationship'] -> ['race']
['fnlwgt', 'hours-per-week',
'native-country', 'education',
'marital-status', 'relationship'] -> ['race']
['fnlwgt', 'hours-per-week',
'native-country', 'education-num',
'marital-status', 'relationship'] -> ['race']
['fnlwgt', 'hours-per-week',
'native-country', 'occupation', 'workclass',
'gender'] -> ['race']
['fnlwgt', 'hours-per-week',
'native-country', 'occupation',
'marital-status'] -> ['race']
['fnlwgt', 'hours-per-week',
'native-country', 'occupation',
'relationship'] -> ['race']
['fnlwgt', 'hours-per-week',
'native-country', 'occupation', 'gender',
'income'] -> ['race']
['fnlwgt', 'hours-per-week', 'education',
'marital-status', 'relationship', 'income']
-> ['race']
['fnlwgt', 'hours-per-week', 'education-num',

```

```

'marital-status', 'relationship', 'income']
-> ['race']
['fnlwgt', 'occupation', 'hours-per-week',
'relationship'] -> ['gender']
['fnlwgt', 'hours-per-week', 'workclass',
'marital-status', 'relationship'] ->
['gender']
['fnlwgt', 'age', 'native-country',
'education'] -> ['race']
['fnlwgt', 'age', 'native-country',
'education-num'] -> ['race']
['fnlwgt', 'age', 'native-country',
'occupation', 'workclass'] -> ['race']
['fnlwgt', 'age', 'native-country',
'workclass', 'marital-status'] -> ['race']
['fnlwgt', 'marital-status', 'age',
'education'] -> ['race']
['fnlwgt', 'age', 'education', 'income'] ->
['race']
['fnlwgt', 'marital-status', 'age',
'education-num'] -> ['race']
['fnlwgt', 'age', 'income', 'education-num']
-> ['race']
['fnlwgt', 'occupation', 'marital-status',
'age'] -> ['race', 'gender']
['fnlwgt', 'occupation', 'relationship',
'age'] -> ['race']
['fnlwgt', 'occupation', 'age', 'income'] ->
['race']
['fnlwgt', 'age', 'workclass',
'marital-status', 'relationship', 'income']
-> ['race']
['fnlwgt', 'relationship', 'age'] ->
['gender']
['fnlwgt', 'native-country', 'education',
'occupation', 'gender'] -> ['race']
['fnlwgt', 'native-country', 'education',
'workclass', 'marital-status',
'relationship', 'gender'] -> ['race']
['fnlwgt', 'native-country', 'education-num',
'occupation', 'gender'] -> ['race']
['fnlwgt', 'native-country', 'education-num',
'workclass', 'marital-status',
'relationship', 'gender'] -> ['race']
['fnlwgt', 'education', 'occupation',
'workclass', 'relationship'] -> ['gender']
['fnlwgt', 'occupation', 'relationship',
'education'] -> ['race']
['fnlwgt', 'education', 'occupation',
'relationship', 'income'] -> ['gender']
['fnlwgt', 'education', 'workclass',
'marital-status', 'relationship', 'gender',
'income'] -> ['race']
['fnlwgt', 'education-num', 'occupation',
'workclass', 'relationship'] -> ['gender']
['fnlwgt', 'occupation', 'relationship',
'education-num'] -> ['race']
['fnlwgt', 'education-num', 'occupation',
'relationship', 'income'] -> ['gender']
['fnlwgt', 'education-num', 'workclass',
'marital-status', 'relationship', 'gender',
'income'] -> ['race']
['fnlwgt', 'occupation', 'marital-status',
'relationship'] -> ['gender']
['education'] -> ['education-num']
['education-num'] -> ['education']

```

Diabetes

```

['DiabetesPedigreeFunction', 'BMI',
'Insulin'] -> ['Glucose', 'Age',
'SkinThickness', 'BloodPressure',
'Pregnancies', 'Diabetes']
['DiabetesPedigreeFunction', 'BMI',
'Glucose'] -> ['Insulin', 'Age',
'SkinThickness', 'BloodPressure',
'Pregnancies', 'Diabetes']
['DiabetesPedigreeFunction', 'BMI', 'Age']
-> ['Insulin', 'Glucose', 'SkinThickness',
'BloodPressure', 'Pregnancies', 'Diabetes']

```

```

['DiabetesPedigreeFunction', 'BMI',
'SkinThickness'] -> ['Insulin', 'Glucose',
'Age', 'BloodPressure', 'Pregnancies',
'Diabetes']
['DiabetesPedigreeFunction', 'BMI',
'BloodPressure'] -> ['Insulin', 'Glucose',
'Age', 'SkinThickness', 'Pregnancies',
'Diabetes']
['DiabetesPedigreeFunction', 'BMI',
'Pregnancies'] -> ['Insulin', 'Glucose',
'Age', 'SkinThickness', 'BloodPressure',
'Diabetes']
['DiabetesPedigreeFunction', 'BMI',
'Diabetes'] -> ['Insulin', 'Glucose',
'Age', 'SkinThickness', 'BloodPressure',
'Pregnancies']
['DiabetesPedigreeFunction', 'Insulin',
'Glucose'] -> ['BMI', 'Age', 'SkinThickness',
'BloodPressure', 'Pregnancies', 'Diabetes']
['DiabetesPedigreeFunction', 'Insulin',
'Age', 'Pregnancies'] -> ['SkinThickness']
['DiabetesPedigreeFunction', 'Insulin',
'BloodPressure'] -> ['SkinThickness',
'Diabetes']
['DiabetesPedigreeFunction', 'Glucose',
'Age'] -> ['BMI', 'Insulin', 'SkinThickness',
'BloodPressure', 'Pregnancies', 'Diabetes']
['DiabetesPedigreeFunction', 'Glucose',
'SkinThickness'] -> ['BMI', 'Insulin', 'Age',
'BloodPressure', 'Pregnancies', 'Diabetes']
['DiabetesPedigreeFunction', 'Glucose',
'BloodPressure'] -> ['Diabetes']
['DiabetesPedigreeFunction', 'Glucose',
'Pregnancies'] -> ['BMI', 'Insulin', 'Age',
'SkinThickness', 'BloodPressure', 'Diabetes']
['DiabetesPedigreeFunction', 'Diabetes',
'Glucose'] -> ['BloodPressure']
['DiabetesPedigreeFunction', 'Age',
'SkinThickness'] -> ['Insulin']
['DiabetesPedigreeFunction', 'Age',
'BloodPressure'] -> ['BMI', 'Insulin',
'Glucose', 'SkinThickness', 'Pregnancies',
'Diabetes']
['DiabetesPedigreeFunction', 'Age',
'Pregnancies'] -> ['Diabetes']
['DiabetesPedigreeFunction', 'SkinThickness',
'BloodPressure'] -> ['Insulin', 'Diabetes']
['DiabetesPedigreeFunction', 'SkinThickness',
'Pregnancies'] -> ['Insulin', 'Diabetes']
['DiabetesPedigreeFunction', 'BloodPressure',
'Pregnancies'] -> ['Insulin',
'SkinThickness', 'Diabetes']
['BMI', 'Insulin', 'Glucose', 'Age'] ->
['DiabetesPedigreeFunction', 'SkinThickness',
'BloodPressure']
['BMI', 'Insulin', 'Glucose',
'BloodPressure'] ->
['DiabetesPedigreeFunction', 'Age',
'SkinThickness', 'Pregnancies']
['BMI', 'Insulin', 'Glucose', 'Pregnancies'] ->
['DiabetesPedigreeFunction', 'Age',
'SkinThickness', 'BloodPressure', 'Diabetes']
['BMI', 'Insulin', 'Age',
'SkinThickness', 'Pregnancies'] ->
['DiabetesPedigreeFunction', 'Glucose',
'BloodPressure']
['BMI', 'Glucose', 'Age'] -> ['Pregnancies',
'Diabetes']
['BMI', 'Glucose', 'Age', 'SkinThickness'] ->
['DiabetesPedigreeFunction',
'BloodPressure']
['BMI', 'Glucose', 'Age', 'BloodPressure'] ->
['DiabetesPedigreeFunction', 'Insulin',
'SkinThickness']
['BMI', 'Glucose', 'SkinThickness'] ->
['Insulin']
['BMI', 'Glucose', 'SkinThickness',
'BloodPressure'] ->
['DiabetesPedigreeFunction', 'Age',
'Pregnancies']
['BMI', 'Glucose',
'SkinThickness', 'Pregnancies'] ->
['DiabetesPedigreeFunction', 'Age',
'BloodPressure', 'Diabetes']
['Diabetes', 'BMI', 'Glucose',

```

```

'SkinThickness'] ->
['DiabetesPedigreeFunction', 'Age',
'BloodPressure', 'Pregnancies']
['BMI', 'Glucose', 'BloodPressure'] ->
['Diabetes']
['BMI', 'Glucose',
'BloodPressure', 'Pregnancies'] ->
['DiabetesPedigreeFunction', 'Insulin',
'Age', 'SkinThickness']
['Diabetes', 'BMI', 'Glucose',
'Pregnancies'] -> ['Age']
['BMI', 'Age', 'SkinThickness',
'BloodPressure'] -> ['Insulin', 'Diabetes']
['BMI', 'Age', 'SkinThickness',
'BloodPressure', 'Pregnancies'] ->
['DiabetesPedigreeFunction', 'Glucose']
['BMI', 'Age', 'SkinThickness',
'Pregnancies'] -> ['Diabetes']
['BMI', 'SkinThickness', 'BloodPressure',
'Pregnancies'] -> ['Insulin']
['Insulin', 'Glucose',
'Age', 'BloodPressure'] ->
['DiabetesPedigreeFunction', 'BMI',
'SkinThickness', 'Pregnancies', 'Diabetes']
['Insulin', 'Glucose', 'Age', 'Pregnancies'] ->
['Diabetes']
['Glucose', 'Age', 'SkinThickness',
'BloodPressure'] ->
['DiabetesPedigreeFunction', 'BMI',
'Insulin', 'Pregnancies', 'Diabetes']
['Glucose', 'Age', 'SkinThickness',
'Pregnancies'] -> ['Diabetes']
['Glucose', 'Age',
'BloodPressure', 'Pregnancies'] ->
['DiabetesPedigreeFunction', 'BMI',
'Insulin', 'SkinThickness', 'Diabetes']
['Glucose', 'SkinThickness', 'BloodPressure',
'Pregnancies'] -> ['Insulin']
['Glucose', 'SkinThickness', 'BloodPressure',
'Pregnancies', 'Diabetes'] ->
['DiabetesPedigreeFunction', 'BMI', 'Age']

```

Housing

```

['median_income', 'total_rooms',
'median_house_value'] -> ['population',
'total_bedrooms', 'households', 'latitude',
'longitude', 'housing_median_age',
'ocean_proximity']
['median_income', 'total_rooms',
'population'] -> ['median_house_value',
'total_bedrooms', 'households', 'latitude',
'longitude', 'housing_median_age',
'ocean_proximity']
['median_income', 'total_rooms',
'total_bedrooms'] -> ['median_house_value',
'population', 'households', 'latitude',
'longitude', 'housing_median_age',
'ocean_proximity']
['median_income', 'total_rooms',
'households'] -> ['median_house_value',
'population', 'total_bedrooms', 'latitude',
'longitude', 'housing_median_age',
'ocean_proximity']
['median_income', 'total_rooms',
'latitude'] -> ['median_house_value',
'population', 'total_bedrooms', 'households',
'longitude', 'housing_median_age',
'ocean_proximity']
['median_income', 'total_rooms',
'housing_median_age'] ->
['median_house_value', 'population',
'total_bedrooms', 'households', 'latitude',
'longitude', 'ocean_proximity']
['median_income', 'median_house_value',
'population'] -> ['ocean_proximity']
['median_income', 'median_house_value',
'population', 'latitude'] -> ['total_rooms',
'total_bedrooms', 'households', 'longitude',
'housing_median_age']

```



```

'total_bedrooms', 'latitude', 'longitude',
'ocean_proximity']
['median_house_value', 'population',
'latitude'] -> ['ocean_proximity']
['median_house_value', 'population',
'longitude'] -> ['median_income',
'total_rooms', 'total_bedrooms',
'households', 'latitude',
'housing_median_age', 'ocean_proximity']
['median_house_value', 'total_bedrooms',
'households', 'latitude'] ->
['median_income', 'total_rooms',
'population', 'longitude',
'housing_median_age']
['median_house_value', 'total_bedrooms',
'households', 'longitude'] ->
['median_income', 'total_rooms',
'population', 'longitude',
'housing_median_age']
['housing_median_age', 'median_house_value',
'total_bedrooms', 'households'] ->
['median_income', 'total_rooms',
'population', 'latitude', 'longitude',
'ocean_proximity']
['housing_median_age', 'median_house_value',
'total_bedrooms', 'latitude'] ->
['ocean_proximity']
['housing_median_age', 'median_house_value',
'total_bedrooms', 'longitude'] ->
['median_income', 'total_rooms',
'population', 'households', 'latitude',
'ocean_proximity']
['median_house_value', 'households',
'latitude'] -> ['ocean_proximity']
['median_house_value', 'households',
'latitude', 'longitude'] -> ['median_income',
'total_rooms', 'population',
'total_bedrooms', 'housing_median_age']
['median_house_value', 'households',
'longitude'] -> ['ocean_proximity']
['housing_median_age', 'median_house_value',
'households', 'longitude'] ->
['median_income', 'total_rooms',
'population', 'total_bedrooms', 'latitude']
['population', 'total_bedrooms',
'households'] -> ['ocean_proximity']
['housing_median_age', 'population',
'total_bedrooms', 'households'] ->
['median_income', 'total_rooms',
'median_house_value', 'latitude',
'longitude']
['population', 'total_bedrooms', 'latitude']
-> ['median_income', 'total_rooms',
'median_house_value', 'households',
'longitude', 'housing_median_age',
'ocean_proximity']
['population', 'total_bedrooms',
'longitude'] -> ['median_income',
'total_rooms', 'median_house_value',
'households', 'latitude',
'housing_median_age', 'ocean_proximity']
['population', 'households', 'latitude'] ->
['ocean_proximity']
['population', 'households', 'latitude',
'longitude'] -> ['median_income',
'total_rooms', 'median_house_value',
'total_bedrooms', 'housing_median_age']
['housing_median_age', 'population',
'households', 'latitude'] ->
['median_income', 'total_rooms',
'median_house_value', 'total_bedrooms',
'longitude']
['population', 'households', 'longitude'] ->
['ocean_proximity']
['housing_median_age', 'population',
'households', 'longitude'] ->
['median_income', 'total_rooms',
'median_house_value', 'total_bedrooms',
'latitude']
['housing_median_age', 'population',
'longitude'] -> ['ocean_proximity']
['total_bedrooms', 'households', 'latitude',
'longitude'] -> ['median_income',
'total_rooms', 'median_house_value',
'population', 'housing_median_age']

```

```

['housing_median_age', 'total_bedrooms',
'households', 'latitude'] ->
['ocean_proximity']
['housing_median_age', 'total_bedrooms',
'households', 'longitude'] ->
['ocean_proximity']
['housing_median_age', 'total_bedrooms',
'latitude', 'longitude'] -> ['median_income',
'total_rooms', 'median_house_value',
'population', 'households']
['latitude', 'longitude'] ->
['ocean_proximity']

```

Loan

```

['Income', 'CITY'] -> ['House_Ownership',
'Married/Single', 'Car_Ownership']
['Income', 'CITY', 'Age'] -> ['Profession',
'Experience', 'CURRENT_JOB_YRS',
'CURRENT_HOUSE_YRS']
['Income', 'CITY', 'Profession'] ->
['Experience', 'CURRENT_JOB_YRS',
'CURRENT_HOUSE_YRS']
['Income', 'CITY', 'Experience'] ->
['Profession', 'CURRENT_JOB_YRS',
'CURRENT_HOUSE_YRS']
['CURRENT_HOUSE_YRS', 'Income',
'CITY'] -> ['Profession', 'Experience',
'CURRENT_JOB_YRS']
['Income', 'Age'] -> ['House_Ownership',
'Married/Single', 'Car_Ownership']
['Income', 'Age', 'Profession'] -> ['CITY',
'STATE', 'Experience', 'CURRENT_JOB_YRS',
'CURRENT_HOUSE_YRS']
['Income', 'Age', 'STATE'] ->
['CITY', 'Profession', 'Experience',
'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS']
['Income', 'Age', 'Experience'] -> ['CITY',
'Profession', 'STATE', 'CURRENT_JOB_YRS',
'CURRENT_HOUSE_YRS']
['CURRENT_HOUSE_YRS', 'Income', 'Age'] ->
['CITY', 'Profession', 'STATE', 'Experience',
'CURRENT_JOB_YRS']
['Income', 'Age', 'Risk_Flag'] -> ['CITY',
'Profession', 'STATE', 'Experience',
'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS']
['Income', 'Profession'] ->
['House_Ownership', 'Married/Single',
'Car_Ownership']
['Income', 'Profession', 'STATE'] ->
['CITY', 'Experience', 'CURRENT_JOB_YRS',
'CURRENT_HOUSE_YRS']
['Income', 'Profession', 'Experience']
-> ['CITY', 'STATE', 'CURRENT_JOB_YRS',
'CURRENT_HOUSE_YRS']
['CURRENT_HOUSE_YRS', 'Income',
'Profession'] -> ['CITY', 'STATE',
'Experience', 'CURRENT_JOB_YRS']
['Income', 'STATE'] -> ['House_Ownership',
'Car_Ownership']
['Income', 'Married/Single', 'STATE',
'Experience'] -> ['Profession',
'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS']
['CURRENT_HOUSE_YRS', 'Income', 'STATE']
-> ['CITY', 'Profession', 'Experience',
'CURRENT_JOB_YRS', 'Married/Single']
['Income', 'Married/Single', 'STATE'] ->
['CITY']
['Income', 'Experience'] ->
['House_Ownership']
['CURRENT_HOUSE_YRS', 'Income',
'Experience'] -> ['CITY', 'Profession',
'STATE', 'CURRENT_JOB_YRS', 'Married/Single',
'Car_Ownership']
['Income', 'CURRENT_JOB_YRS'] -> ['CITY',
'Profession', 'STATE', 'Experience',
'CURRENT_HOUSE_YRS', 'House_Ownership',
'Married/Single', 'Car_Ownership']
['CURRENT_HOUSE_YRS', 'Income',
'Car_Ownership'] -> ['House_Ownership']
['CITY'] -> ['STATE']
['CITY', 'Age', 'Profession', 'Experience',
'CURRENT_JOB_YRS'] -> ['House_Ownership']
['CITY', 'Age', 'Profession', 'Experience',

```

```
'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS'] ->
['Car_Ownership']
['CITY', 'Age', 'Profession', 'Experience',
'CURRENT_HOUSE_YRS'] -> ['House_Ownership',
'Married/Single']
['CITY', 'Age', 'Profession', 'Experience',
'CURRENT_HOUSE_YRS', 'Risk_Flag',
'Car_Ownership'] -> ['CURRENT_JOB_YRS']
['CITY', 'Age', 'Profession', 'Experience',
'Risk_Flag', 'Married/Single'] ->
['House_Ownership']
['Age', 'Profession', 'STATE', 'Experience',
'CURRENT_JOB_YRS', 'CURRENT_HOUSE_YRS',
'Risk_Flag', 'Car_Ownership'] ->
['Married/Single']
```