研究内容：

本文的研究内容分为5个章节，各章节主要内容如下：

第 1 章 绪论。解析与本研究相关的背景知识及国内外研究现状，阐述本研究的研究目的、理论意义及实践意义。从国内外相关研究的数据来源、研究方法及工具、研究内容方面介绍糖尿病视网膜病变当前研究现状。最后，简析本文的研究方法、拟解决的关键问题、遵循的研究思路及技术路线。

第 2 章相关理论、方法与工具。本章首先介绍了医疗大数据相关的理论基础，包括知识发现、信息链、决策支持理论，并简述了其嵌入在数据挖掘流程的具体环节及发挥的作用。概述了与数据挖掘过程相关的数据降维、过度拟合的概念，辨析了数据挖掘与知识发现概念的联系和差别。最后，简要介绍了较为常用的数据挖掘算法的工作原理，以及三种数据挖掘工具的来源、功能及优缺点。

第 3 章 疾病预测模型构建策略。本章首先解析了医疗大数据背景下，疾病预测模型构建过程中数据集成与规范化、数据填补与降维、模型构建与评价三个环节的关键问题。然后，针对问题制定了对应的数据集成与规范方案、缺失数据填补方案、数据降维方法及模型构建与评价方案，制定了疾病预测模型构建策略。

第 4 章糖尿病视网膜病变预测模型构建实证研究。本章首先对数据进行集成与清洗，接着判断不同性别在患病结果的差异显著性、同一性别在不同年龄段患病结果的差异显著性，据此对目标数据集中的缺失数据进行分层均值填补。然后，分别采用主成分分析提取特征根大于 1 的因子、提取累计贡献率大于 85%的因子、Logistic 回归提取患病结果差异有显著性的因子，据此对数据进行降维。最后，通过机器学习确定原始数据基线精度，再利用决策树算法构建未降维与三种降维方法的疾病预测模型，并检验三种降维方法结果差异显著性。通过评价指标和对照实验模型对所建模型进行评估，并分析对照试验模型与所建模型结果差异显著性。

第 5 章 结论与展望。根据研究所做的各项工作，得出研究相关结论。探讨缺失数据填补方案、不同降维方法、不同机器算法对疾病预测模型构建结果的影响，以及该策略的优势及面临的挑战。

拟待解决的问题：

* 缺失值

缺失数据会增加分析任务的难度和复杂性，影响数据挖掘工作的效率的同时，更容易造成分析结果的偏倚。运用常规统计方法填补缺失数据，即使不能代替整体，但合适的填补方法也能降低结果的偏倚。常用的数据缺失处理方法有均值填补法(Mean)、期望值最大化法(expectation maximization，EM)、回归填补法(Regression)和多重填补法(multiple imputation，MI)。均值填补法以简单便捷、快速高效的优势，已被广泛应用到缺失数据填补过程。因此，选择统计学中缺失值处理方法中的分层均值插补，将其应用到预测模型缺失值处理环节。有文献提出性别和年龄是糖尿病视网膜病变的关联因素之一，据此，对目标数据集按照性别、年龄段、是否患病分为若干子集。首先，将总体划分为男女两部分，并检验两组是否患病之间统计学差异显著性，然后再将不同性别的两组按照年龄段划分，再检验各组之间是否有统计学差异。判断各组之间是否具有统计学差异，进一步证明分组划分的合理性，再据此对每层分别求平均值进行填补。

* 维度

本研究中选择三种方法对原始数据进行降维，即根据主成分分析提取特征根大于 1 的因子、根据主成分分析提取特征累计贡献率大于 85%的因子、根据逻辑回归提取差异有显著性的因子。第一种降维方法，从每个属性特征出发，根据每个属性在数据集整体中的特征值根降维，特征根的大小代表了矩阵正交化之后所对应特征向量对于整个矩阵的贡献程度，在此以特征根是否大于 1 作为分界点，提取特征根大于 1 的属性。第二种降维方法，是从数据整体特征出发，提取能够代表数据特征 85%以上的属性进行降维。第三种降维方法，采用逻辑回归方法，分析每个属性在分类属性上是否有显著差异，筛选差异有显著性的属性进行降维。

* 分布不平衡
* 过拟合

目前国际上通过大型的队列研究得到风险预测模型。包括美国弗兰明翰心脏研究中心建立的冠心病10年风险预警模型，英国国家健康与临床优化研究所对其修订的QRISK2风险评估，建立心血管疾病10年预警模型，整合了传统的弗兰明翰研究中心风险评估所包含的年龄，收缩压，血脂值，性别，吸烟等情况，也增加了种族，心血管疾病家族史，体重指数，社会剥脱指数以及和心血管危险相关疾病（高血压，糖尿病，房颤，肾脏病，类风湿性关节炎等），临床实践证明，该风险模型更适用于欧洲人群冠心病危险评估。

2007年里德发布了Reynolds风险评分，通过观察血压，抽烟状态，饮酒，绝经状态，体重，身高，胆固醇治疗，种族，年龄，家族早发的心肌梗死病史，运动频度以及最近服用多种维生素等35种可能的风险因子，预测Reynolds风险预警模型，相比与前两者，该风险评估新增了心肌梗死家族史和高敏性急性反应蛋白

本文的创新点主要体现在两个方面，首先，传统的心血管疾病的研究大多都即将于实验设计，选定人群和探索性发现已经设定可能风险因子建立风险预警模型，这对风险因子的发现有一定的局限性。而本文研究不同与传统医学的方法，而是建立在深圳市历史医疗大数据基础上，以高血压人群作为研究对象