

吉林工程技术师范学院

2021-2022-2 学期

课程名称：_____ 职业教育大数据分析实训 _____

考试性质：_____ ☐补考 ☐重修 ☒期末考试 _____

考核形式：_____ ☐论文 ☒报告 ☐作品 ☐其他 _____

任课教师：_____ 苏顺亭、王忠文 _____

小组成员：_____ 吴泞宇、杨济铨 _____

班 级：_____ 大数据 1941 班 _____

成 绩：_____

吉林工程技术师范学院教务处印制

吉林工程技术师范学院 2021--2022 学年第 2 学期论文考试要求

课程名称: 职业教育大数据分析 课程所在学院: 数据科学与人工智能学院

考查班级: 大数据 1941 命题人: 苏顺亭、王忠文

论文要求:

1、论文题目(范围) 关于中职教育发展现状的分析与预测;

2、论文要求 手写 ☐ 打印 ☒, 字数范围为 6298 字;

3、论文上交时间: 2022 年 06 月 17 日;

4、写作要求(请任课教师详细说明本次论文考试的内容要求)

严禁剽窃、抄袭等作弊行为!

1) 确立题目后, 首先进行相关文献阅读, 至少阅读三篇文献

2) 论文标明数据来源

3) 给出数据处理方法和相关结果, 解读运算结果的含义

4) 结合问题对结果进行分析, 总结。

5) 注明参考文献

鼓励同学自学新的统计方法, 鼓励创新!

论文上交电子版和纸质版。

教研室主任意见:

签字: _____ 年 月 日

学院负责人意见:

签字: _____ 年 月 日

注: 此表一式两份, 一份于考前交到考试中心, 一份随学生论文装订成册

目录

关于中职教育发展现状的分析与预测	1
一、研究背景和意义	1
二、理论知识	1
(一) 数据处理	1
(二) 数据分析	2
(三) 预测模型	3
三、应用分析	3
(一) 中职教育发展总体性分析	3
(二) 中职教育分科发展现状分析	7
(三) 中职教育分省发展现状分析	8
(四) 特征预测模型网页应用开发	9
四、结果分析	13
五、参考文献	16

关于中职教育发展现状的分析与预测

【摘要】

中职教育是我国高中阶段教育的重要组成部分,担负着培养高素质职业劳动者的重要任务,是我国经济发展的重要基础。当前我国中职教育发展缓慢,严重制约我国城镇化进程。因此,本文依托于大数据分析技术,对中职教育发展历程展开多个维度的分析与特征值预测,来探讨中职教育发展现状以及所要面临的现实问题。

针对于数据环节。本文采用动态网页采集技术,对来源于国家统计局官网的数据,展开数据采集工作,并将数据存储到了本地文件系统。然后,将经过数据清洗的原生数据集,集成为总体、分学科及分省三个层次的数据集,并存储到 Mysql 数据库中。

针对于分析环节。本文依托 Pandas_profiling (分析报表生成技术) 及 Pyecharts (数据可视化技术), 利用描述性分析、特征关联性分析以及特征对照分析等方法, 从总体、分学科、分省三个维度出发, 展开了中职教育发展现状分析实践。

针对于中职教育关键特征值预测环节。本文依托于,三个搭载有 Ubuntu 系统(基于 Linux 内核)的虚拟机,搭建了伪分布式 Hadoop 集群及 Spark 集群, 其中 HDFS 为网页应用提供了分布式存储服务, Spark 集群为网页应用提供了分布式运算服务。此外, 本环节又选用 PySpark 技术, 实现了线性回归模型 (LinearRegression) 的开发; 并依赖于 Streamlit 技术, 实现了关键特征值预测模型网页应用的开发与部署。

总而言之, 从总体、分学科及分省三个维度出发, 本文通过详细分析中职教育近 14 年来发展历程的同时, 挖掘出了中职教育发展所要面临的现实问题。最终, 向公众科普中职教育发展历程以及现实意义的同时, 也希望引起公众对中职教育发展的重视和思考。

【关键词】

中职教育、大数据分析技术、伪分布式 Spark 集群、PySpark、Streamlit

一、研究背景和意义

中职教育是我国高中阶段教育的重要组成部分，担负着培养高素质职业劳动者的重要任务，是我国经济社会发展的重要基础。当前，我国中职教育发展相对缓慢，是整个教育中的薄弱环节。部分地区甚至对发展中职教育存在误解，把发展高中阶段教育片面理解为就是发展普通高中，在经费投入、资源配置等方面往往忽视中职教育。

公众对中职教育发展的普遍忽视，将制约我国走新型工业化道路、解决“三农”问题和城镇化建设的进程。因此，本文希望通过大数据分析技术，对中职教育近年来的发展现状，展开多个维度的量化分析与特征值预测。探讨中职教育发展现状以及所面临的现实问题。向公众科普中职教育发展历程的同时，也希望引起公众对中职教育发展的广泛关注和思考。

二、理论知识

针对于数据采集、数据预处理、数据存储及数据分析环节，开发环境为：Windows10、Anaconda3-jupyter、Python3.7 及 Mysql。针对于预测模型网页应用开发环节，开发环境为：Ubuntu16.04.2 虚拟机系统（基于 Linux 内核）、Anaconda3-jupyter、Python3.7、MobaXtermP 及 VMware Workstation Pro。本文结合上述开发环境，从数据处理、数据分析及预测模型三个环节出发，展开理论知识部分的必要叙述。

（一）数据处理

1. 数据采集

本次职业教育分析实践，数据来源于国家统计局官网。该网站采

用主流网页信息加载模式：**Ajax** 异步同驱技术，它根据用户的点击行为，动态加载相应数据。因此，针对于动态网页而言，数据采集环节的实施，核心问题在于：了解用户行为驱动的请求表单构造及功能。

2. 数据预处理

前述环节采集到的中职教育相关数据集，存在：表格数量多、表格信息冗余、数据缺失及数据单位不统一等问题。考虑到中职教育发展现状分析环节的工作效率，数据存储到 **Mysql** 数据库之前，需要进行必要的**数据预处理（数据清洗、归一化及数据集成等）。

3. 数据存储

Mysql 数据库，常用于强调实时性处理数据集的情景，具有：运行速度快、使用成本低及可移植性强的特点。因此，本文选用 **Mysql** 数据库，作为集成数据的存储方式。

（二）数据分析

数据分析环节，本文依赖于集成数据集，从总体、分学科及分省三个维度出发，利用描述性分析、特征关联性分析及特征对照分析方法，展开中职教育发展现状的量化分析。

描述性分析，从数据类型、数据分布及数据特征角度出发，展开数据集的描述性刻画。特征关联性分析，利用肯德尔距离(**Kendall's T**)，来反映特征之间的正负相关性。特征对照分析，利用可视化技术，动态展示不同维度中的数据分布、变化趋势及关联性。

（三）预测模型

针对于环境搭建,本文依托于三个搭载有 Ubuntu 系统(基于 Linux 内核)的虚拟机节点,搭建了伪分布式 Hadoop 集群及 Spark 集群。其中分布式文件系统(Hadoop Distributed File System)为网页应用提供了分布式存储服务,Spark 集群提供分布式运算服务。

针对于模型构建,本文采用 PySpark 技术,实现了线性回归预测模型(LinearRegression)的开发,并选用均方误差及 R² 判定系数,作为模型泛化性能的度量方法。针对于网页应用开发,本文采用 Streamlit 技术,实现了特征值预测模型网页应用的开发与部署。

三、应用分析

本环节围绕着“中职教育发展现状分析及预测”这一主题,利用大数据分析技术及前述相关分析理论,从总体、分学科及分省三个维度出发,展开应用分析叙述。值得注意的是,转型期(2010 至 2014 年间)是本文重点关注的时间节点,这一阶段中职教育处于衰退期拐点。因此,在分学科及分省应用分析环节,将以转型期为切入点,探讨中职教育在社会关注冷落期及发展衰退期间,发生了哪些有趣的改变。

（一）中职教育发展总体性分析

1. 描述性分析

本文采用 pandas_profiling 模块,针对于中职教育发展总体性数据集,生成数据分析报告(网页版),展开了描述性分析。完整描述性分析结果,详见文件 AnalysisReport.html。

从数据规模的角度展开,如图 3.1 所示。总体性数据集存在 7 个

特征（年份、招生数、在校生数、毕业生数、获职业资格认证毕业生数、教职工数及学校数），数据类型皆为数值型，且无缺失值。

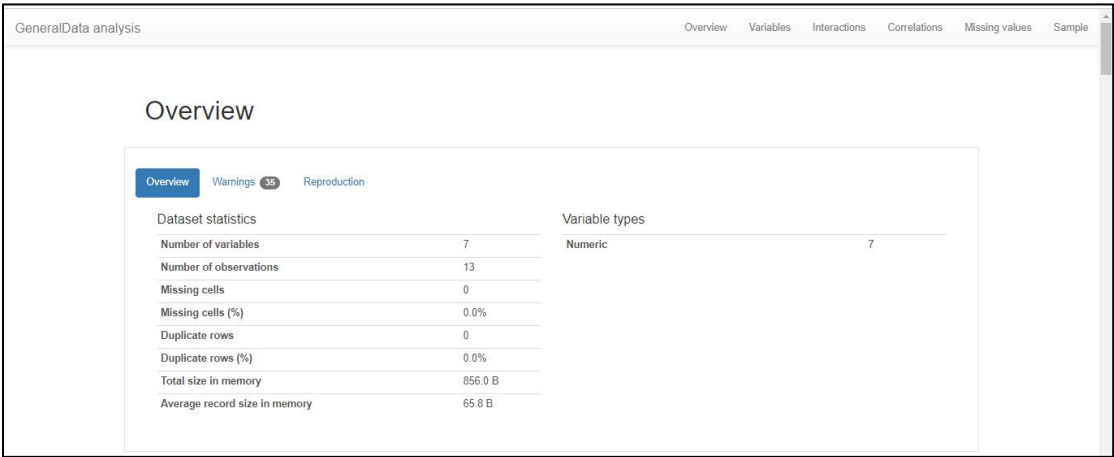


图 3.1 数据规模 Overview

从特征值分布的角度展开，以招生数为例，如图 3.2 所示。2008 至 2020 年间，最大招生数约 712 万人，最小招生数约 429 万人，十三年中有七年的招生数小于 500 万人。

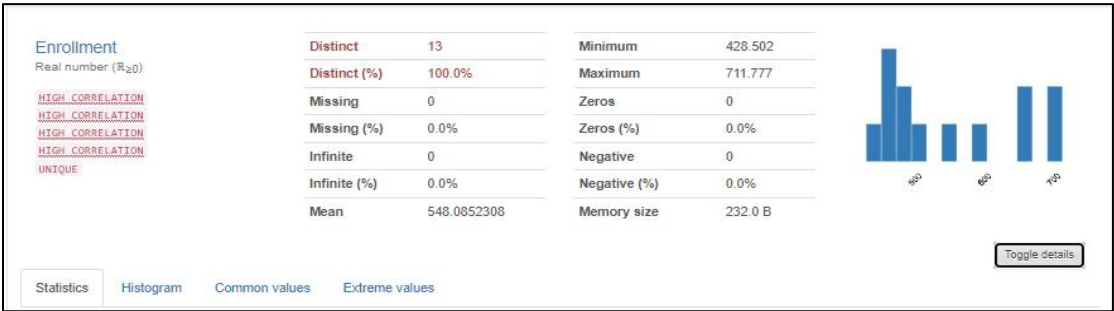
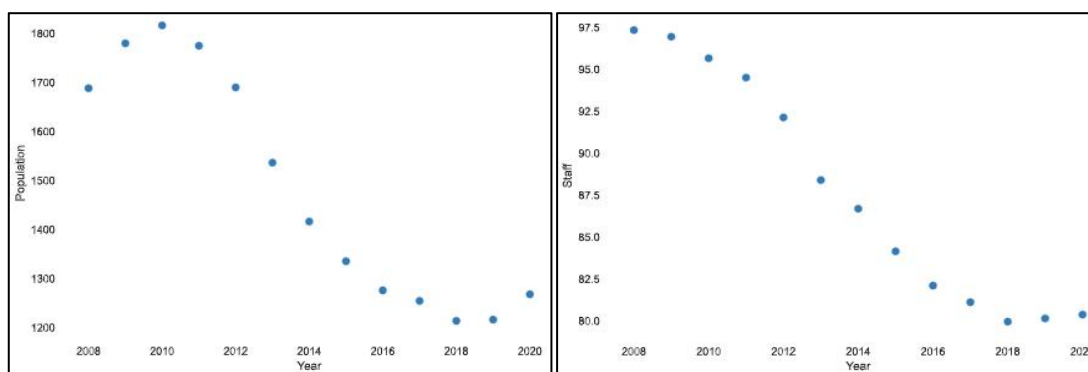


图 3.2 招生数特征分布（节选）

从特征变化趋势的角度展开。以在校生数为例，如图 3.3（a）所示。中职教育院校在校生数在 2010 年达到顶峰，约 1800 万人。在 2010 至 2018 年间，在校生规模出现了大幅度衰减。及至 2018 年，中职教育在校生规模才出现了小幅度回弹。值得注意的是，以教职工数为例，如图 3.3（b）所示。教职工数量自 2008 年以来，一直处于大幅度缩水的状态；及至 2018 年以后，教职工数量才出现受限的小幅度增长。而中职院校数变化趋势，同教职工数变化趋势基本一致。



(a) 在校生数变化趋势

(b) 教职工数变化趋势

图 3.3 关键特征变化趋势 Interactions (节选)

2. 特征关联性分析

从特征关联性的角度展开，以图 3.4 为例。时间特征与其它特征都呈显著的负相关，体现了中职教育整体上处于衰退的状态。招生数、在校生数、毕业生数三者密切相关，其中招生数的变化会滞后影响在校生数及毕业生数。而教职工数及学校数与招生数、在校生数及毕业生数之间存在一定程度上的关联性，体现了学校数及教职工数是制约中职教育规模的关键条件，符合客观事实。

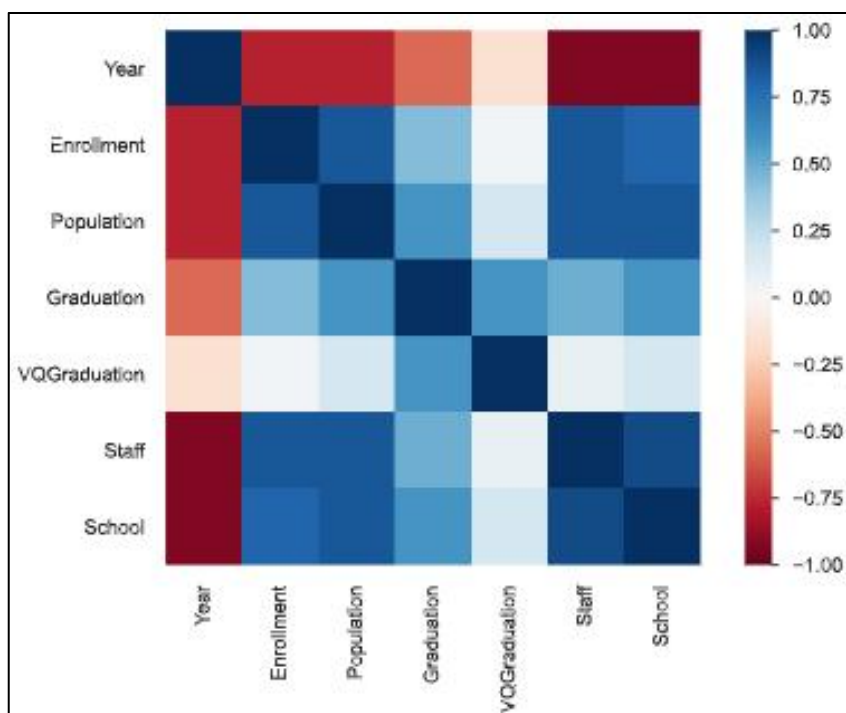


图 3.4 特征关联性可视化 (Kendall's T)

3. 特征对照性分析

特征对照性分析所依赖的可视化图表，详见文件：中职教育发展总体性数据可视化.html。

如图 3.5 所示，及至 2013 年，获得职业资格证书毕业生数占毕业生总数的比例大幅提高，中职院校职业技能教培水平显著提高。



图 3.5 两特征对照一览图

如图 3.6 所示，2018 年以来，随着招生数量的增加，获得职业资格证书毕业生数占毕业生总数的比例有所下降。推测是中职院校数量及教职工数量的相对缺乏，导致了学校教培水平出现了不利波动。当然，也可以合理推测出，中职教育生源质量并没有获得改善。



图 3.6 三特征对照一览图

（二）中职教育分科发展现状分析

本环节从招生数、在校生数、毕业生数、获得职业资格证毕业生数及教职工数这五个特征出发，依托于动态玫瑰图，展现不同学科在转型期间（2010至2014年间）的中职教育分科发展状况。中职教育分科发展现状分析所依赖的动态可视化图表，详见文件：中职教育学科发展数据可视化.html。

分科类在校生数占比，一定程度上体现了，中职教育的分科发展现状。因此，如图 3.7 所示。可以发现如下现象：其一是转型期期间中职教育的发展方向更加集中在信息技术、制造业及财经商贸类；其二是中职教育的学科类培养开始趋向于多元化，前述的三大类专业总体比重有所下降；其三是中职教育开始紧跟时代发展需要，比如农林类专业占比有所下降，教育及医药卫生类专业占比出现了大幅的增长。

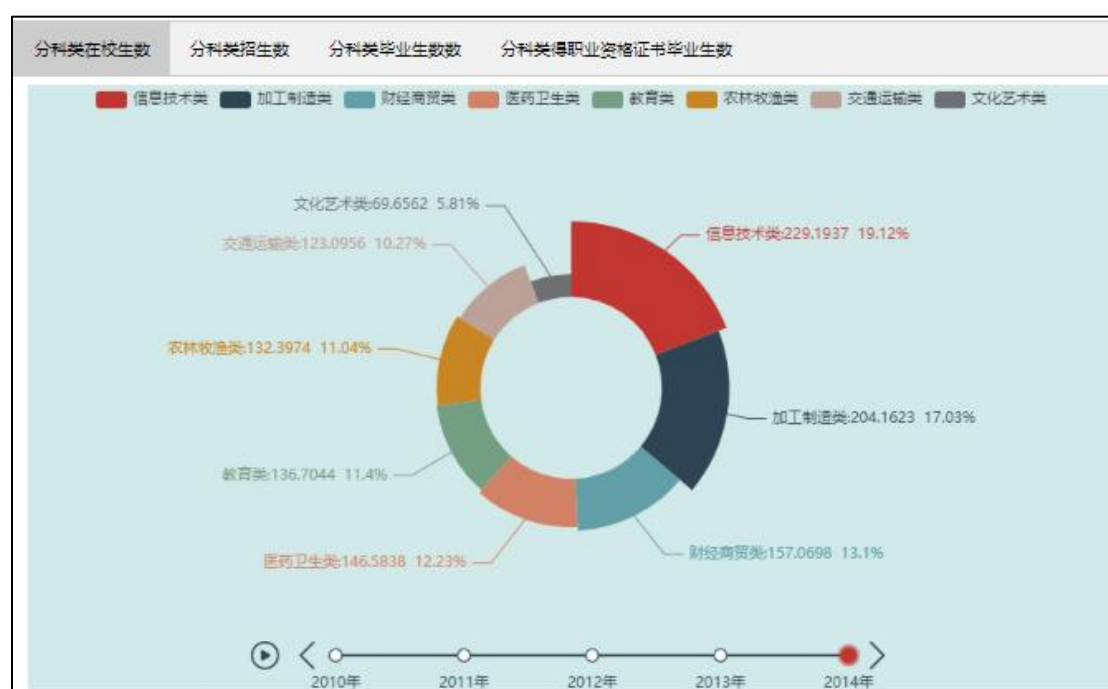


图 3.7 中职教育分科发展（转型期）

（三）中职教育分省发展现状分析

本环节同样从招生数、在校生数、毕业生数、获得职业资格证毕业生数及教职工数这五个特征出发，依托于动态热点地图，展现中职教育院校分布情况。中职教育分省发展现状分析所依赖的动态可视化图表，详见文件：中职教育分省发展数据可视化.html。

如图 3.8 所示，当时间节点固定在转型期末期的 2014 年时，中职教育主要分布在华北、广东及四川一带。这样特殊的分布，与中职教育热点地区自身特点密切相关。例如：在华北一带，人口稠密且重工业发达；沿海地带，商贸发达；而在广东及四川一带，轻工业发达。



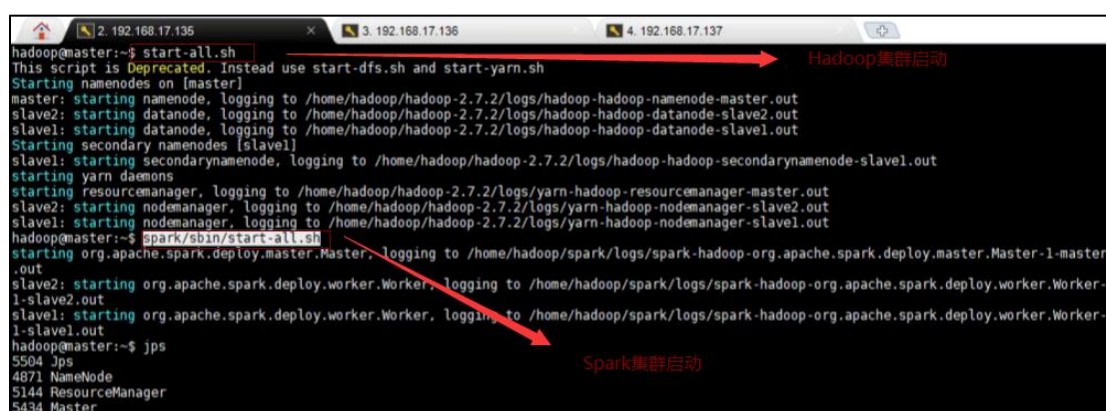
图 3.8 中职教育发展分布（转型期 2014 年）

（四）特征预测模型网页应用开发

本文关于中职教育关键特征预测模型网页应用的开发，涉及伪分布式 Hadoop 集群及 Spark 集群搭建、数据上传、特征工程、线性模型训练、性能评估、特征值预测及网页应用开发等环节。

1. 伪分布式环境

伪分布式 Hadoop 集群所属的分布式文件系统（HDFS），为预测模型网页应用提供数据存储服务。伪分布式 Spark 集群，为预测模型网页应用提供分布式运算服务。Anaconda3-jupyter 为预测模型网页应用开发，提供了编辑环境。由于伪分布式环境搭建过程太过于复杂，本文不再详细叙述，而是聚焦于伪分布式环境的使用。Hadoop 集群及 Spark 集群的启动方式，如图 3.9 所示。



```
hadoop@master:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [master]
master: starting namenode, logging to /home/hadoop/hadoop-2.7.2/logs/hadoop-hadoop-namenode-master.out
slave2: starting datanode, logging to /home/hadoop/hadoop-2.7.2/logs/hadoop-hadoop-datanode-slave2.out
slave1: starting datanode, logging to /home/hadoop/hadoop-2.7.2/logs/hadoop-hadoop-datanode-slave1.out
Starting secondary namenodes [slave1]
slave1: starting secondarynamenode, logging to /home/hadoop/hadoop-2.7.2/logs/hadoop-hadoop-secondarynamenode-slave1.out
starting yarn daemons
starting resourcemanager, logging to /home/hadoop/hadoop-2.7.2/logs/yarn-hadoop-resourcemanager-master.out
slave2: starting nodemanager, logging to /home/hadoop/hadoop-2.7.2/logs/yarn-hadoop-nodemanager-slave2.out
slave1: starting nodemanager, logging to /home/hadoop/hadoop-2.7.2/logs/yarn-hadoop-nodemanager-slave1.out
hadoop@master:~$ spark/sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /home/hadoop/spark/logs/spark-hadoop-org.apache.spark.deploy.master.Master-1-master.out
slave2: starting org.apache.spark.deploy.worker.Worker, logging to /home/hadoop/spark/logs/spark-hadoop-org.apache.spark.deploy.worker.Worker-1-slave2.out
slave1: starting org.apache.spark.deploy.worker.Worker, logging to /home/hadoop/spark/logs/spark-hadoop-org.apache.spark.deploy.worker.Worker-1-slave1.out
hadoop@master:~$ jps
5504 jps
4871 NameNode
5144 ResourceManager
5434 Master
```

图 3.9 集群启动

2. 数据上传

依赖于软件 MobaXtermP，可在 Windows 环境下登录虚拟机主节点 Master，实现 Windows 系统中数据的上传。上传至 HDFS 成功后，可通过分布式文件系统管理界面查看数据文件，如图 3.10 所示。

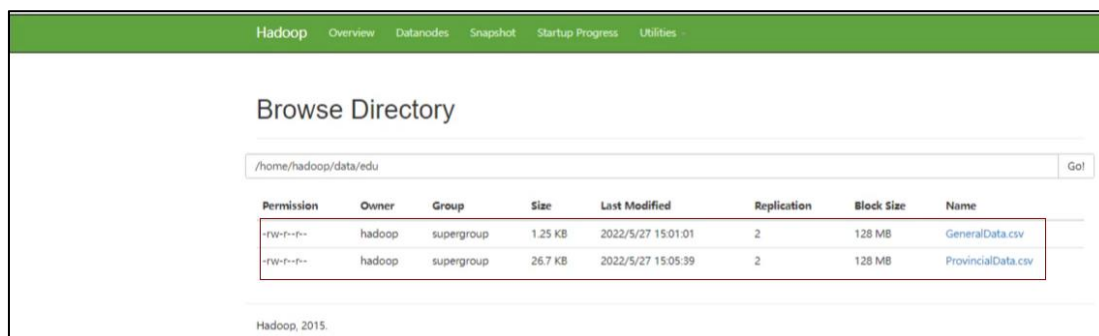


图 3.10 HDFS 管理界面

3. 特征工程及模型训练

PySpark 提供的线性回归算法，要求输入值类型为向量，因此需要引入特征关系性分析的经验指导，来选择适宜的特征线性组合（特征构造），并将它们合并为一个向量变量 **features**。如图 3.11 所示。

```
In [6]: from pyspark.ml.linalg import Vector
        from pyspark.ml.feature import VectorAssembler

        vec = VectorAssembler(inputCols=['教职工总数', '普通中等专业学校数'], outputCol='features')
        features_df = vec.transform(df)
        features_df.printSchema()
```

图 3.11 特性向量化

此外，还需要对存储在 HDFS 中的数据文件 GeneralData.csv，按照 7:3 比例随机划分为训练集 **train_df** 及测试集 **test_df**，用于模型训练及预测。考虑到特征值类型皆为连续型数值，故选择应用较为广泛的多元线性回归模型算法 **LinearRegression**，该模型属于有监督学习。在模型构建时，需要人为标定影响因素向量 **features** 及预测标签 **labelCol**。模型构建及拟合训练代码，如图 3.12 所示。

```
In [10]: from pyspark.ml.regression import LinearRegression

        lin_reg = LinearRegression(featuresCol='features', labelCol='招生数')
        lr_model = lin_reg.fit(train_df)
        lr_model.coefficients
```

图 3.12 模型构建及拟合训练

4. 特征值预测及性能评估

以选用教职工数及学校数作为输入影响因子 **features**，招生数为预测标签 **labelCol** 为例。针对于事先划分好的预测集 **test_df**，其预测结果如图 3.13 所示。

```
In [13]: #预测结果与真实结果对比
predictions = lr_model.transform(test_df)
print (predictions.collect())
print (predictions.show())
```

[Row(features=DenseVector([79.9593, 3322.0]), 招生数=428.5024, prediction=436.43624680944015), Row(features=DenseVector([92.1332, 3681.0]), 招生数=597.0785, prediction=603.6432342457803), Row(features=DenseVector([95.6631, 3938.0]), 招生数=711.3957, prediction=536.3080737150805)]

features	招生数	prediction
[79.9593, 3322.0]	428.5024	436.43624680944015
[92.1332, 3681.0]	597.0785	603.6432342457803
[95.6631, 3938.0]	711.3957	536.3080737150805

图 3.13 招生数特征预测

此外，本文选用均方误差及 **R2** 判定系数，作为模型泛化性能的度量工具。以 **R2** 判定系数为例，其取值范围为 **0-1**，值越接近于 **1**，则说明预测误差越小，泛化性能越强。如图 3.14 所示，训练集 **R2** 判定系数约为 **0.97**，说明教职工数及学校数是影响招生数的重要特征，这与前文特征关联性分析结果不谋而合。

```
In [12]: #通过R2判定系数，评估模型的拟合程度，其值越接近1说明模型有较高的价值
train_p = lr_model.evaluate(train_df)
print(' 0 0 '.format('训练集R2判定系数:', train_p.r2 ))
test_p = lr_model.evaluate(test_df)
print(' 0 0 '.format('预测集R2判定系数:', test_p.r2 ))
#通过均方误差meanSquaredError评估估值的准确性
print(' 0 0 '.format('均方误差:', train_p.meanSquaredError))
```

训练集R2判定系数: 0.9721166799817282

图 3.14 模型泛化性能评估

5. 网页应用开发

Streamlit 是专门为机器学习专业人士提供的，用于网页应用便捷开发及快速部署的 **Python** 第三方包。本环节依赖于编辑环境

Anaconda3-jupyter-linux，实现开发及部署。

由于虚拟主机选用的 Ubuntu 字符界面系统（基于 Linux 内核），不提供可视网页交互界面。因此需要在主节点 Master 中输入 `streamlit run Main.py` 命令来启动网页应用服务的同时，复制指定网址（虚拟机 IP 地址+网页应用端口号）到 Windows 环境下的浏览器中，实现跨操作系统登录。启动 Streamlit 网页应用服务，如图 3.15 所示。

```
hadoop@master:~$ cd /home/hadoop/jupyternotebook/PysparkLR/SteamlitApp
hadoop@master:~/jupyternotebook/PysparkLR/SteamlitApp$ streamlit run Main.py
2022-06-06 20:43:21.746 INFO matplotlib.font_manager: generated new fontManager
er

You can now view your Streamlit app in your browser.

Local URL: http://localhost:8501
Network URL: http://192.168.17.135:8501

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hadoop/spark/jars/slf4j-log4j12-1.7.30.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hadoop/hadoop-2.7.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
22/06/06 20:44:30 WARN NativeCodeLoader: Unable to load native-hadoop library fo
```

图 3.15 启动 Streamlit 网页应用服务

网页应用界面，如图 3.16 所示。

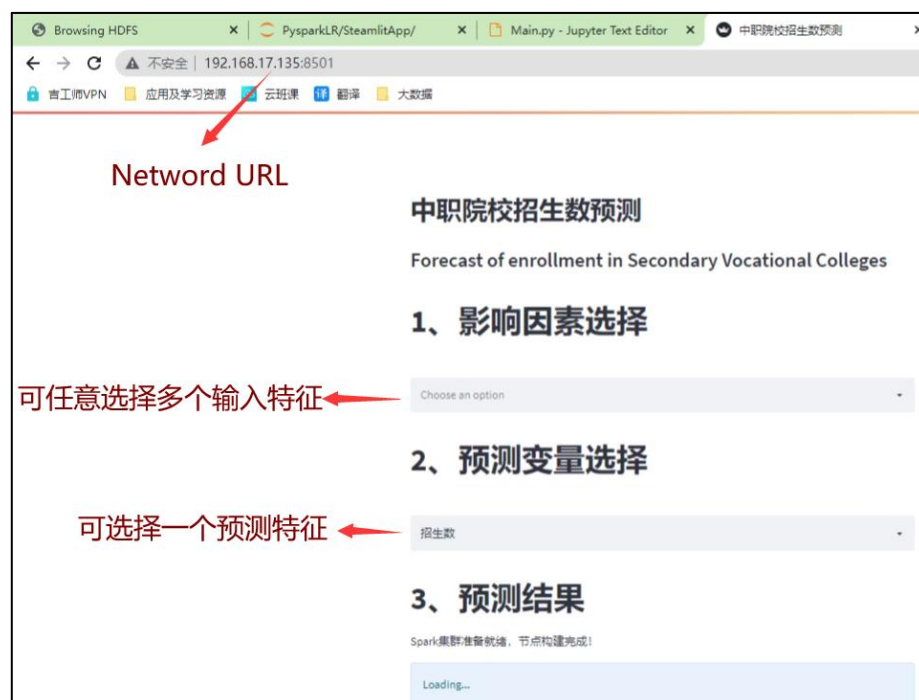


图 3.16 网页应用界面

针对于中职教育特征值预测模型网页应用，用户可选任意多个输入特征为影响因素，一个特征为预测变量。点击运行按钮后，网页应用将运行封装好的模型程序，依赖于 Spark 集群环境，完成指定特征值的预测，并将结果存储到 HDFS 的同时，显示输出到网页界面中。网页应用测试，如图 3.17 所示。



图 3.17 网页应用测试

四、结果分析

针对于中职教育发展现状，经过前述总体、分学科、分省三个维度的应用分析论述后，本环节总结出了中职教育发展历程时间线，如图 4.1 所示。依托于中职教育发展历程时间线，本环节就中职教育发展在近十余年来的跌宕起伏，给出一个总结：2008 至 2010 年属于后增长时期，虽然中职教育规模处于增长状态，但是教职工数量已经开始衰减，中职教育已经显现衰落迹象。因此，2010 年成为了中职教

育发展的最后辉煌。而 2010 至 2018 年可以认为是“遗失的八年”。及至 2010 年以后，中职教育的招生数、在校生数、毕业生数、教职工数及学校数开始整体性的衰减，中职教育发展正式步入衰退期。其中，2013 至 2015 年属于中职教育发展转型期，是中职教育规模衰退的拐点，也是前述应用分析环节的重点关注时期。在转型期期间，中职教育“集中力量”在资源优势地区保持住了规模的同时，学科培养体系也逐渐向国家经济发展需要靠拢。及至 2018 年，中职教育在新时代教育改革的大背景下，其规模水平有所回升。然而，总体上来说，中职教育仍然存在大量问题，制约其发展。

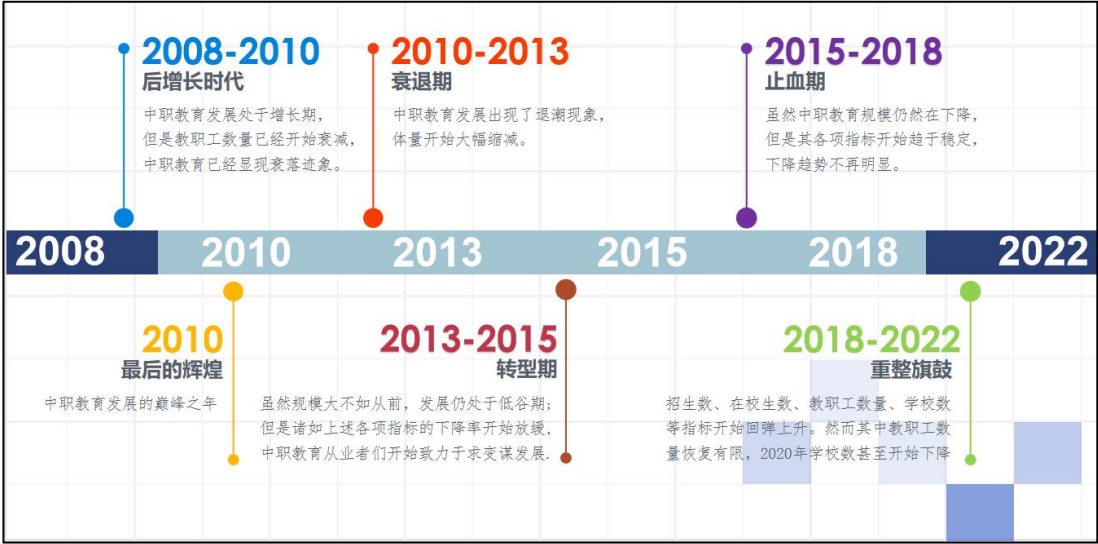


图 4.1 中职教育发展历程时间线

制约中职教育发展的关键问题，可以总结为如下几点：

其一是教育资源匮乏。纵观中职教育发展历程，可以发现中职教育关于教职工需求的巨大缺口，是制约中职教育转型发展的关键因素。这也反映了，中职教育在教育资金、办学环境及政策方面的资源匮乏。

其二是社会认可度低。中职教育作为高中阶段教育的重要组成部分，长期饱受社会的质疑和忽视，社会认可度普遍较低。它带来了生

源质量普遍较差等一系列的连锁问题。

其三是地区发展失衡。中职院校集中在具有人口聚集、贸易及重工业发达等特征的地区。其它地区并没有结合自身特点，发展因地制宜的中职教育。中职教育在不同地区的发展出现了严重失衡现象。

其四是培养体系不完善。中职教育遇冷的关键在于，中职院校本身在学生职业能力培养、升学深造及职业资格认证等方面存在大量问题，无法保证毕业生的就业质量。

造成上述四个问题的本质原因，是我国教育资源的总体性缺乏及发展不均衡，中职教育本身在部分地区也常常被忽视和牺牲。值得注意的是，近年来国际贸易市场的动荡，使得国家愈加希望通过拉动内需来刺激经济的持续增长。老生常谈的城镇化，实际上就是刺激内需的关键所在。而在城镇化的建设过程中，本身就需要大量高素质职业劳动者，中职教育在其中扮演着不可忽视的作用。因此，中职教育应当在新时代教育改革的大背景下，获得更多的关注和投入。

五、参考文献：

- [1] 路宝利.新世纪十年中国职业教育的发展困境与思考[J].开放教育研究，2012(05).
- [2] 林海.基于大数据下的 Spark 快速大数据分析[J].现代工业经济和信息化，2019(10).
- [3] 吴喜之.多元统计分析 R 与 Python 的实现[M].北京：中国人民大学出版社，1999.
- [4] 周志华.机器学习及其应用[M].北京：清华大学出版社，2007.