# 房价估值大数据平台案例教学指导手册

**教学适用的课程：**《大数据架构与设计》

**教学适用的专业：**电子信息

**教学目的与用途：**理解实际生活中的大数据应用；掌握大数据预处理、数据分析相关知识点；掌握大数据处理与分析的基本流程；掌握数据可视化的相关知识点。

**教学内容：**

（1）理论依据：本案例基于大数据技术根据房屋各项属性对价格进行预测。案例整合了常用的数据挖掘与机器学习方法，进行数据预处理，预测了各种房屋类型及其状况下的价格，为决策问题做出相应指导。

（2）涉及知识点：数据挖掘；数据预处理；回归预测；数据可视化。

（3）分析路径：首先分析了房价的实际意义，介绍了房价问题在社会生活中的重要性。随后通过数据挖掘，得到了房屋的各项基本属性及房价，展示并说明了数据的基本信息。再对挖掘到的数据进行预处理，主要包括筛选有效数据、离散映射房屋价格、面积等属性及数据编码。然后使用决策树算法对房价进行预测，并对算法结果与实际房价进行了误差和准确性测试。最后可视化算法的预测结果，形成房价估值大数据平台。

**启发思考题：**

（1）如果决策树算法的输入数据维度过高，可以采用哪种办法降低数据维度？（参考答案：分析各属性与目标值的关联性或使用例如PCA等数据降维方法。）

（2）如果算法的测试误差较大或过小，应该怎么定位问题并解决？（参考答案：若误差较大，则可能说明数据集较少，算法不能调整到相对较好的参数，可以考虑加大数据集容量；若误差过小，则可能说明算法对于数据过拟合，可以考虑添加算法惩罚项。）

**建议课堂计划：**

1. 时间安排：3课时
2. 学生学习准备：了解常见的数据挖掘相关知识、数据预处理相关方法、分类预测方法的理论知识。
3. 分组及讨论内容：分组讨论启发思考题。
4. 案例开场白：房价是社会生活中的一个普遍而又重要的话题，它关系着我们当中的每一个人。作为购房者的我们，如何尽量挑选满足自己需求同时价格又相对低廉的房子是最重要且关键的问题。这节课我们将学习如何根据市场信息预测房屋价格，可能可以为我们在购房时提供有效的建议。
5. 结束总结：这节课我们以房价大数据为例，进行了数据挖掘并预处理；通过使用决策树模型，预测具有不同情况的房屋价格；同时，使用了数据可视化的相关技术，将我们得到的数据结果较为直观地展现出来，以供我们后续分析购房问题。在生活中，我们经常可以遇到这样的例子。如果同学们有这些棘手的问题，不妨试一试我们学习到的相关知识来解决它，我相信你会感受到它的实际应用价值。
6. 案例引导建议：在教学过程中，教师可以适度融入其他的机器学习算法对该问题进行分析处理。

**参考文献：**

[1] SHVACHKO K, KUANG H, RADIA S, et al. The Hadoop Distributed File System; proceedings of the IEEE Symposium on Mass Storage Systems & Technologies, F, 2010 [C].

[2] DEAN J, GHEMAWAT S. MapReduce: Simplified Data Processing on Large Clusters [J]. USENIX Association, 2004.

[3] ZAHARIA M, CHOWDHURY M, DAS T, et al. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing; proceedings of the Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation, F, 2012 [C].

[4] TAYLOR R C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics [J]. Bmc Bioinformatics, 2010, 11 Suppl 12(S12): S1.

[5] 刘小虎， 李生. 决策树的优化算法 [J]. 软件学报, 1998, 9(10): 4.

[6] 格林贝格. Flask Web开发 [M]. Flask Web开发, 2015.

[7] 王子毅, 张春海. 基于ECharts的数据可视化分析组件设计实现 [J]. 微型机与应用, 2016, 35(014): 46-8.

**其他教学支持材料：**

1. 本案例提供配套的PPT、视频、数据集与代码等，发布于Github，链接为：https://github.com/Wanghui-Huang/CQU\_bigdata。

2. 本案例涉及到数据预处理以及多种机器学习算法，建议使用python语言进行编写，推荐的工具包有pandas（数据读取与预处理库），scikit-learn（机器学习算法库），Flask和ECharts（可视化绘图库）。