# Data Mining:
# Analyzing and Predicting the Quality of Product online

*COMP3503 Project*
*Lei Xie*
*123800*

# Introduction

- Web changed people’s behaviour
    - Shopping, leaving comments
    - Amazon, Best Buy
- Data Analysis
    - Discovering useful information
    - Helping decision making
- Data Mining
    - Process of discovering patterns data
    - Automatic
    - Lead to advantage

# Approach

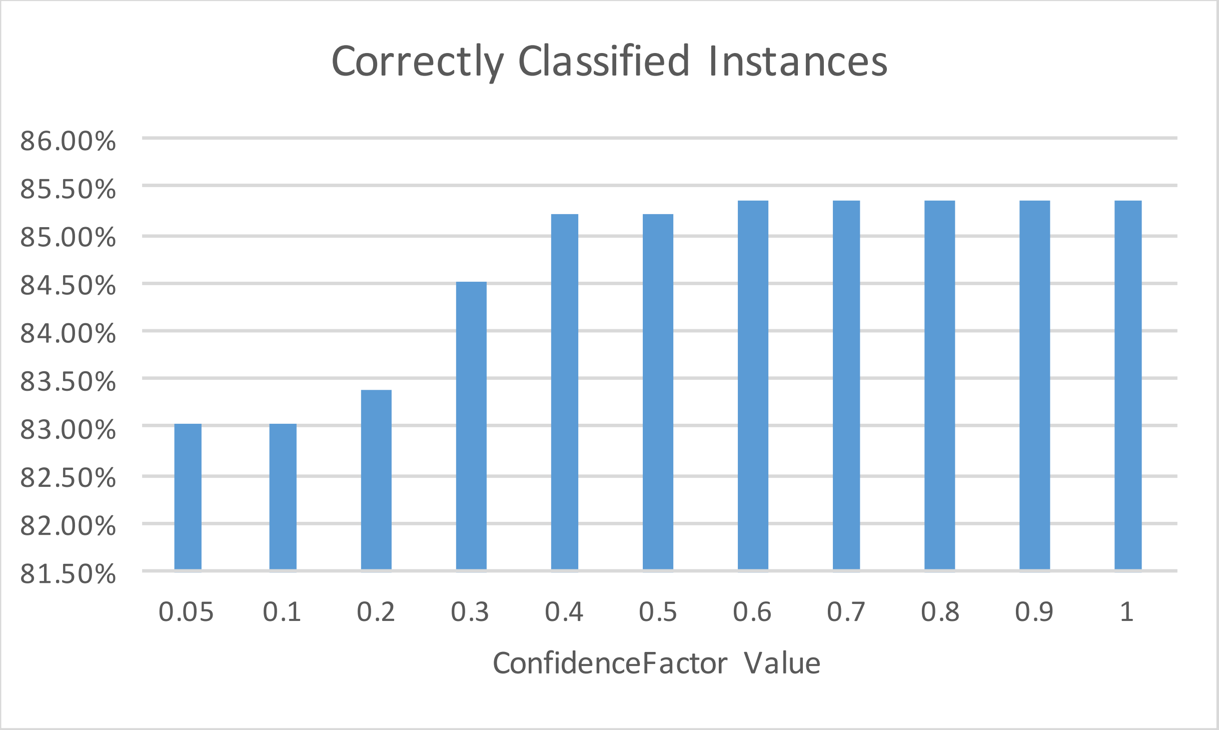
- Project Plan
  - Cross Industry Standard Process Data mining (CRISP-DM) model
  - Hierarchical process including six Steps: Business understanding, Data understanding, Data preparation, Modelling, Evaluation, Deployment
- Solution
  - C program to deal with original data
  - WEKA
  - Two models: J48 classifier, NaiveBayesMultinmial classifier
- Experimental Design
  - Adjust Parameters
  - J48: -C -M value
  - NaiveBayesMultinmial: IDFTransform, TFTransform

# Results

- J48
  - ConfidenceFactor Value
  - minNumObj Value

# Results

- -M = 2
- -C = 0.05 ~ 1.0

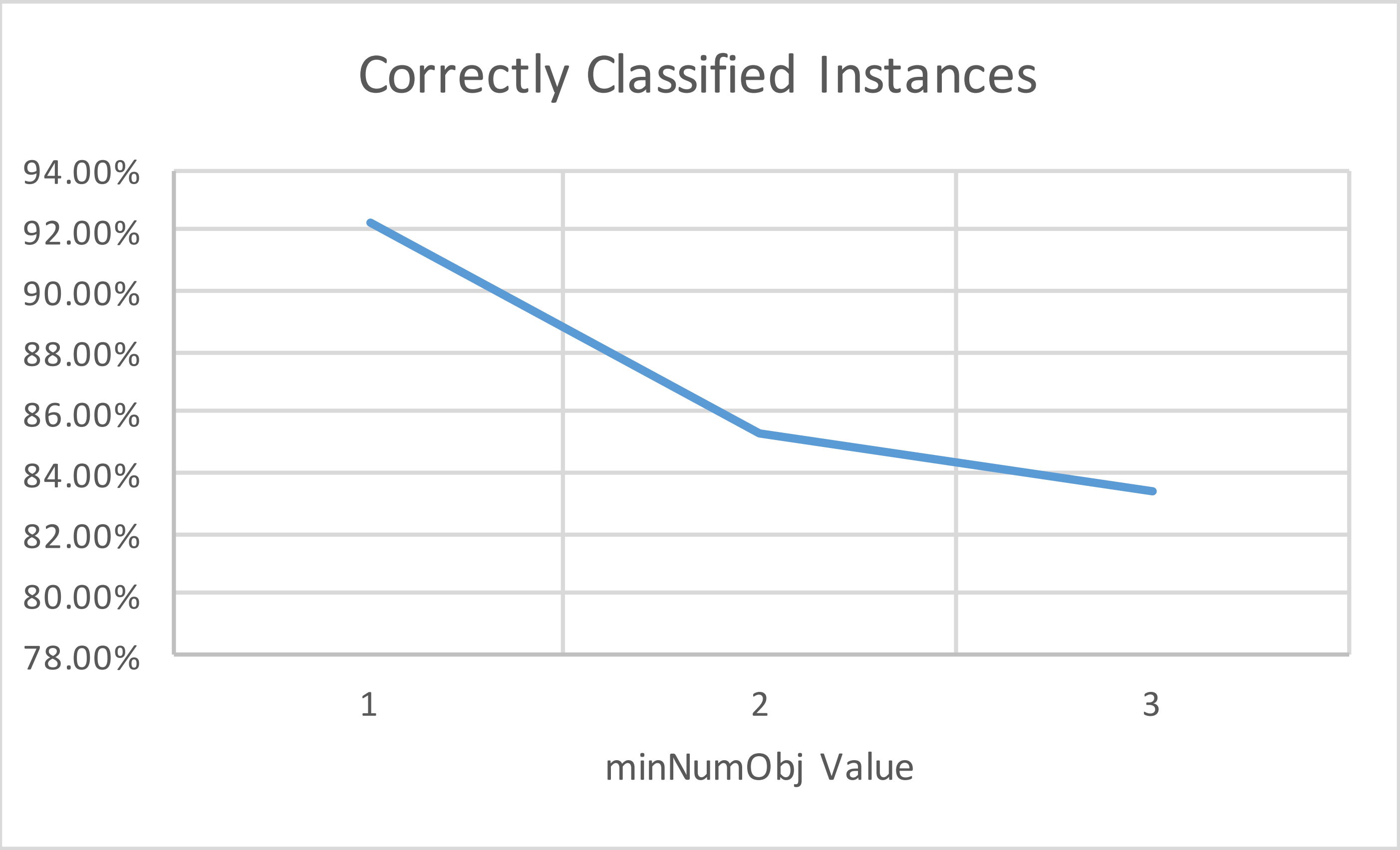


| ConfidenceFactor Value | Correctly Classified Instances |
|---|---|
| 0.05 | 83.05% |
| 0.1 | 83.05% |
| 0.2 | 83.40% |
| 0.3 | 84.50% |
| 0.4 | 85.20% |
| 0.5 | 85.20% |
| 0.6 | 85.35% |
| 0.7 | 85.35% |
| 0.8 | 85.35% |
| 0.9 | 85.35% |
| 1.0 | 85.35% |

# Results

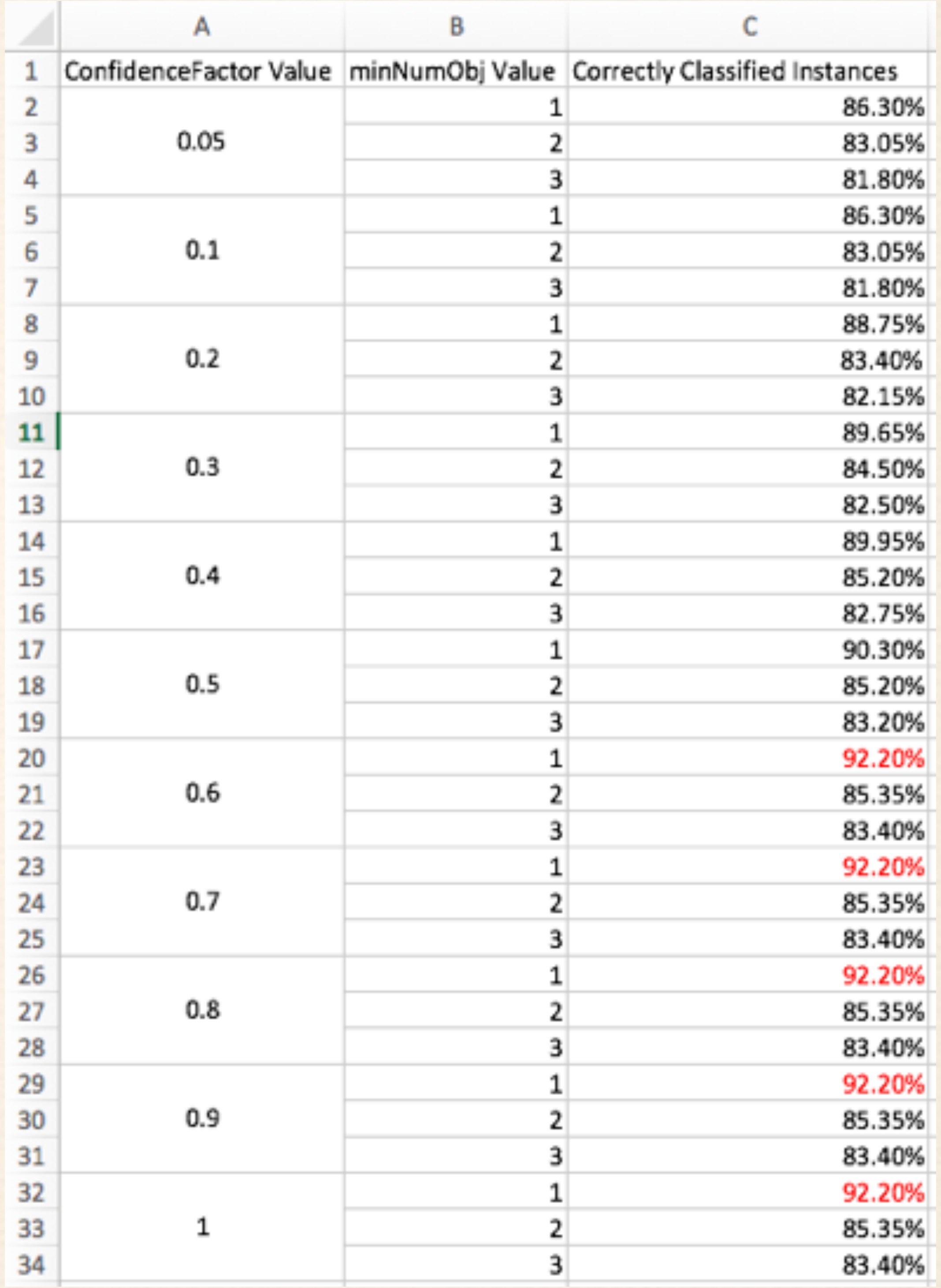
- -C = 0.6
- -M = 1, 2, 3

| minNumObj Value | Correctly Classified Instances |
|---|---|
| 1 | 92.20% |
| 2 | 85.35% |
| 3 | 83.84% |



# Results

- Best result for J48
- Correct Rate: 92.20%



|  | A | B | C |
|---|---|---|---|
| 1 | ConfidenceFactor Value | minNumObj Value | Correctly Classified Instances |
| 2 | 0.05 | 1 | 86.30% |
| 3 |  | 2 | 83.05% |
| 4 |  | 3 | 81.80% |
| 5 | 0.1 | 1 | 86.30% |
| 6 |  | 2 | 83.05% |
| 7 |  | 3 | 81.80% |
| 8 | 0.2 | 1 | 88.75% |
| 9 |  | 2 | 83.40% |
| 10 |  | 3 | 82.15% |
| 11 | 0.3 | 1 | 89.65% |
| 12 |  | 2 | 84.50% |
| 13 |  | 3 | 82.50% |
| 14 | 0.4 | 1 | 89.95% |
| 15 |  | 2 | 85.20% |
| 16 |  | 3 | 82.75% |
| 17 | 0.5 | 1 | 90.30% |
| 18 |  | 2 | 85.20% |
| 19 |  | 3 | 83.20% |
| 20 | 0.6 | 1 | 92.20% |
| 21 |  | 2 | 85.35% |
| 22 |  | 3 | 83.40% |
| 23 | 0.7 | 1 | 92.20% |
| 24 |  | 2 | 85.35% |
| 25 |  | 3 | 83.40% |
| 26 | 0.8 | 1 | 92.20% |
| 27 |  | 2 | 85.35% |
| 28 |  | 3 | 83.40% |
| 29 | 0.9 | 1 | 92.20% |
| 30 |  | 2 | 85.35% |
| 31 |  | 3 | 83.40% |
| 32 | 1 | 1 | 92.20% |
| 33 |  | 2 | 85.35% |
| 34 |  | 3 | 83.40% |

# Results

- NaiveBayesMultinomial Model
- Test 1:
  - IDFTransform = False
  - TFTransform = False
- Test 2:
  - IDFTransform = True
  - TFTransform = True

| NaiveBayesMultinomial model | Correctly Classified Instances |
|---|---|
| TEST 1 | 95.4% |
| TEST 2 | 97.3% |

# Discussion

- J48
  - best correct rate = 92.20%
- NaiveBayesMultinomial
  - best correct rate = 97.3%

# Conclusion

- Problem Statment
  - Big data
  - Prediction
- Approach
  - CRISP-DM
- Testing
  - Adjusting parameters
- Final model
  - NaiveBayesMultinoimal

# Bibliography

What is the CRISP-DM methodology? (n.d.). Retrieved November 29, 2015 from http://www.sv-europe.com/crisp-dm-methodology/

Chapman, P, & Clinton, J. (n.d.). CRISP-DM 1.0 *Step-by-step data mining guide*. Retrieved November 29, 2015, from http://www-staff.it.uts.edu.au/~paulk/teaching/dmkdd/ass2/readings/methodology/CRISPWP-0800.pdf

*Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Elsevier.*

*Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proceedings of the 14th International Conference on World Wide Web (pp. 342–351). New York, NY, USA: ACM. http://doi.org/10.1145/1060745.1060797*

*“Thanks and Questions?”*