Summative Report

Big Data Analytics

University of York

Word Count = 2995

Table of Contents

| 1.1 Part 1a | 2 |
| --- | --- |
| 1.2 Part 1b | 3 |
| 1.2 Part 1c | 5 |
| 2.1 Relational Database | 6 |
| 2.2 Scaling | 7 |
| 3.1 Web-based Application | 8 |
| Reference List | 11 |
| Appendix A | 15 |
| Appendix B | 16 |
| Appendix C | 17 |
| Appendix D | 18 |
| Appendix E | 19 |

# 

# 

1.1 Part 1a

An initial inspection of the data showed that it required cleaning, which involved: checking for duplicate rows and missing values, standardising values and the removal of outliers. [[1]](https://www.zotero.org/google-docs/?BkeWrE) highlights the importance of the error detection and correction process. Notably, time should be invested heavily in the cleaning of data in order to mitigate inaccurate results and prevent the creation of unreliable decisions extracted from data. This is supported by [[2]](https://www.zotero.org/google-docs/?8pwsfk) and examples of the importance of this process are illustrated in the detection of ‘https’ in the ‘demand’ instances and the spelling of ‘no’ being inconsistent. Correction of raw data means that it is now accurate for analysis.

Bias in data presents issues for the accuracy of results. For example, the total instances of ‘yes’ in the demand attribute are significantly greater than the ‘no’ instances. Therefore, the ‘yes’ and ‘no’ instances were split evenly to create a new balanced training set. [[3]](https://www.zotero.org/google-docs/?D7WiUP) argues for this approach stating that imbalanced datasets can perform poorly on the minority class while the error rate is reduced for the majority class. However, [[4]](https://www.zotero.org/google-docs/?V8wLXP)’s research delineates how this method can accidentally remove potentially useful information. That being said, algorithms can be sensitive to unbalanced data so undersampling of the ‘yes’ instances was utilised.

Low demand properties are those with a demand attribute instance of ‘no’ and high demand properties with a ‘yes’ instance. Therefore, this is a binary classification task. Having said this, [[5]](https://www.zotero.org/google-docs/?7X8yL8) argues for a greater understanding of the discrete variables which is useful to determine if certain variables influence demand more than others. Fig 1.2 [Appendix A] delineates an observation of bedrooms having a strong positive correlation of 0.87464 with demand. This implies that properties with more bedrooms have a higher demand, but correlation does not not imply causation. Indeed, [[6]](https://www.zotero.org/google-docs/?GLtduO) illuminates how ML learning algorithms are not skillful in finding causation between variables and so conclusions and assumptions should not be assumed. Supporting this, Fig 1.1 [Appendix A] reveals that properties with fewer bedrooms generate greater demand. A similar strong correlation is observed in the number of bathrooms, and to a lesser extent allowing cats, contribute to whether a property will be classified as high demand. Furthermore, analysis of these results suggests that low demand properties will have a greater number of bedrooms and bathrooms.

A Random Forest (RF) algorithm was selected for classification because it offers high accuracy and reduces the risk of overfitting. This is supported by [[7](https://www.zotero.org/google-docs/?JPamX2)] who identified Weka’s Random Forest algorithm’s robustness to provide highly accurate results with multiple variables. On the other hand, logistic regression provides an alternative if the relationship between attributes and prediction outcomes are simple and linear. Given the variables and size of the dataset, a random forest is suited as it handles non-linear relationships [[8]](https://www.zotero.org/google-docs/?mjmSI2) more advantageously.

The random forest model was run on two sets of balanced data: Model One with all discrete variables and the Model Two with bedrooms, bathrooms and cats\_allowed. Both models performed well with a high degree of accuracy [Appendix B]. Indeed, both models returned 97.87% accuracy with 184 instances correctly identified. However, Model Two has higher accuracy with its mean absolute error, 0.119 compared to Model One’s 0.1354. Root mean squared error is also lower for Model Two 0.1871. Significantly, Model Two’s receiving Operator Characteristic and Precision-Recall Area Under the Curve are higher, indicating a better performance when attributes with high levels of correlation are selected.

1.2 Part 1b

A preliminary step in the analysis process is to gain a visual understanding of patterns and the potential relationships between variables [[9]](https://www.zotero.org/google-docs/?g3cuCf). [[10]](https://www.zotero.org/google-docs/?tJL7lR) suggests an approach of encoding categorical to numerical values. However, no distinct linear relationship or correlation patterns were identified.Importantly, this step provides insight into variable relationships and assists in forming assumptions for which learning model to choose. Indeed, [[11]](https://www.zotero.org/google-docs/?1jeUeo) highlights how this is an important step which enables efficient processing and exploration of data.

Two choices were considered for a ML model: Logistic Regression (LR) or a Decision Tree (DT). An advantage of LR is that it can show a correlation by how independent variables are associated with being in a category of the dependent variable [[12]](https://www.zotero.org/google-docs/?7cE2wI). However, it assumes a linear relationship between dependent and independent variables. Importantly, from the inspection of the graphs, the relationships between variables are not linear and LR is not the most suitable choice given the complex relationship between variables. Alternatively, a DT branches data into subsets and works well with non-linear data [[13]](https://www.zotero.org/google-docs/?c4CSyM). Indeed, feature importance can be used to identify positive or negative relationships between dependent and independent variables. Significantly, a DT can also provide nuanced insights which coefficients, such as Spearman and Pearson, cannot provide. For example, positive relationships can be inferred if the tree branches on a certain feature which leads to higher values on a target variable [[14]](https://www.zotero.org/google-docs/?H2lTjN). As the relationship is not linear a decision tree was selected.

Correlation filters indicated that the ‘type’ attribute is more significant than rent, with 0.218 compared to 0.127, using the CorrelationAttributeEval Fig 3.2 [Appendix C]. This finding is supported with the InfoGainAttributeEval with type at 0.0385 and rent at 0.022 Fig 3.1 [Appendix C]. Importantly, correlation analysis reveals patterns in data [[15]](https://www.zotero.org/google-docs/?ORMzSS) and these results indicate an order of importance to the selected attributes in relation to demand.

The model indicates that apartment instances have a positive correlation with demand which supports findings from the correlation filters. These results show that demand positively correlates with rent as observed in Fig 3.3 [Appendix C] where a higher rent leads to more instances of ‘yes’ in the demand attribute. Notably, the DT shows branches taken from both ‘rent’ and ‘type’ Fig 3.4 [Appendix C]. Indeed, for certain ranges of ‘rent’ high demand is predicted and for other ranges low demand. This indicates a complex and non-linear relationship when looking for a correlation between rent and type. Significantly, ‘type’ and ‘rent’ in conjunction show influences in ‘demand’. For example, a medium of 875 across all property types is associated with high demand on this model. All in all, a positive correlation of ‘rent’ is particularly associated with ‘demand’.

1.3 Part 1c

To identify if the size of a property has an optimal range for generating demand, the data was reduced to two attributes, ‘sqfeet’ and ‘demand’. The interquartile range was then used for the grouping of properties with similar sizes [[16]](https://www.zotero.org/google-docs/?KLUBhe). Importantly, this identifies both high and low demand based on size. Furthermore, it illustrates which size of property generates more demand as well as how demand varies across property size [[17]](https://www.zotero.org/google-docs/?BLfiLx). This provides a greater insight compared to averages, or totals, and indicates that properties in the 50th percentile have a positive correlation with demand instances of ‘yes’. As a point of comparison, the data was attempted to be balanced using the interquartile ranges but this could not be done evenly and indicates a requirement for sampling more properties from the 25th and 75th quartiles.

Native Bayes (NB) was selected to model ‘demand’ and ‘sqfeet’. Notably, NB can work well with categorical and continuous data alongside providing a probabilistic prediction for decision making. However, it is a black box model so the nuance of its decisions might not provide an understanding of relationships in the data. Furthermore, [[18]](https://www.zotero.org/google-docs/?JAertr) highlights the importance of requiring balanced data for a greater accuracy in results. Also, NB assumes feature independence which can result in missing connections between variables, which is why feature selection was used to mitigate this [[19]](https://www.zotero.org/google-docs/?EEtdUu). However, the model results from the unbalanced data indicate that NB is only acting as a yes indicator as proved by the TP and FP having a score of 0.0 for no, but a score of 1.0 for yes. When NB was run on the balanced data the results improved, with a TP score of 0.774 and a Kappa score of 0.7064 [Appendix D]. This indicates a requirement to try a different model to improve confidence.

Support Vector Machine (SVM) is an alternative model which can model non-linear data and generalise well on unseen data [[20]](https://www.zotero.org/google-docs/?CyOAx1). Moreover it can make use of Kernel functions for optimisation, but this can be a disadvantage as the Kernell choice can affect the model’s performance and generalisation ability [[21]](https://www.zotero.org/google-docs/?TpKzw5). Having said that, the results show a high level of confidence on the balanced dataset with the Kappa statistic suggesting a substantial agreement between model predictions and actual classifications, see Fig 4.2 [Appendix D]. It should be noted that the model does perform better on identifying in demand properties with a ‘yes’ instance as indicated by the TP rate on the balanced dataset.

In comparison, when the SVM model is applied to an unbalanced dataset it does not perform as well. Notably, the model’s Kappa statistic suggests the performance is no better than chance. Furthermore, the confusion matrix supports this as 705 instances of ‘no’ were classified as ‘yes’. Indeed, the model’s performance indicates that it is biassed towards predicting ‘yes’, and is similar to Native Bayes, highlighting the importance of using a balanced dataset.

In summary, use of the interquartile ranges shows that properties in the 50th percentile indicate a greater probability of being in demand. With balanced data this can also be predicted with a strong level of confidence.

2.1 Relational Database

Normalisation of databases improves data integrity, reduces redundancy and allows for data manipulation [[22]](https://www.zotero.org/google-docs/?a0jiyb). These are important aspects for database design as they can minimise duplication of data entries and improve accuracy through tables linked by relationship. An important aspect of database normalisation is that it is cumulative and so efficiency increases with each normalisation. Indeed, depending on the requirements, fourth and fifth Normal Form (NF) can be employed [[23]](https://www.zotero.org/google-docs/?gRghkt).

In order to change the current flat file database into Third NF the data must first be transformed into First NF. This is done by ensuring atomicity, no duplication of rows and there must be a primary key. After cleaning the flat file database for data analysis most of these prerequisites are already met. By selecting ‘id’ as the primary key to uniquely identify each record the database is now in finalised First NF. ‘id’ is selected as the primary key as each property has a unique id number and so this can be utilised as the primary key for First NF in the new database.

For a table to be in Second NF it must already be in First NF and have no partial key dependencies as this can lead to data redundancy [[24]](https://www.zotero.org/google-docs/?JKJfWx). To achieve this: composite primary keys and partial dependencies are identified, then removed, and relationships between tables are established with foreign keys. Next, the relationships between columns are analysed to identify any partial dependencies. This is done by identifying unique values across the columns and those with fewer unique values are candidates [[25]](https://www.zotero.org/google-docs/?fKNgOP) as they suggest a dependency on other attributes other than the primary key. For instance, ‘region’ and ‘region\_url’ have 83 unique values indicating a direct relationship can be normalised to a separate table with region as the primary key. Importantly, this removes the partial dependency of ‘region\_url’ on ‘region’ [Appendix E].

Lastly, to ensure the database is in Third NF compliant transitive dependencies must be removed and a review of the relationships between tables, such as reviewing foreign keys to preserve reference integrity [[26]](https://www.zotero.org/google-docs/?7Q89u6). Significantly, changing to Third NF is vital for data integrity and reduction of data redundancy. For instance, join tables are used (listing\_options, listing\_amenities) to model relationships with greater complexity, many to many, which cannot be done directly by relational databases [Appendix E]. Furthermore, SQL statements can now be used to retrieve the required data [Appendix E].

2.2 Scaling

Expanding internationally will entail a system which can handle large scale data processing and provide fast response times. A distributed computing approach will be considered in order to improve speed and manage data efficiently. Notably, NoSql database options are utilised as they ensure scalability across multiple servers and can handle larger volumes of traffic [[27]](https://www.zotero.org/google-docs/?gyEbWf). Furthermore, [[28]](https://www.zotero.org/google-docs/?Cse5RU) highlights how performance is improved through the ability to scale horizontally. Importantly, NoSQL suits the housing manager’s requirements as scalability is achieved and performance is increased for future business requirements.

Apache Cassandra is considered for the database storage and management as it is capable of handling large volumes of data across commodity servers which provide a high level of availability. A benefit is put forward by [[29]](https://www.zotero.org/google-docs/?qlyFTQ) who highlights how Apache Cassandra has no single point of failure. Significantly, this is vital for the reliability and availability of the database given the housing manager’s requirements. An alternative to Apache Cassandra, Amazon DynamoDB, is a fully managed database with built-in security. However, it is not open source meaning extra costs might be incurred [[30]](https://www.zotero.org/google-docs/?2LC0wz) and the housing company can end up in vendor lock-in which is a strong consideration for future services and business needs. Positively, vendor lock-in offers the benefits of a fully integrated service, such as security. On the other hand, unpredictability with pricing, data portability and data customisation highlight the greater control and flexibility offered by open source options, like Apache Cassandra. Justification for this decision is made by examining the attributes in the dataset, such as: ‘region’, ‘type’, ‘rent’, ‘sqfeet’ and ‘bedrooms’. Apache Cassandra can efficiently query, and manage, data from these attributes and find listings within a certain range [[31]](https://www.zotero.org/google-docs/?AHoP3J).

Batch and stream processing are considerations for data analysis. Indeed, a service such as Apache Spark offers the large scale data analysis required by the housing manager. For instance, Apache Spark is fast as it makes use of a Directed Acyclic Graph scheduler which offers an alternative to Hadoop’s MapReduce [[32]](https://www.zotero.org/google-docs/?wPx4DK) which offers an improvement in processing speed. This makes a strong argument for Apache Spark being well suited for real time data analysis and ML [[33]](https://www.zotero.org/google-docs/?KDglYo). For example, Apache Spark’s features allow for alerts to be set up and send messages based on real time data analysis. Indeed, an alert can be set for an increase in rental prices in a certain region and if these conditions are met automated alerts can be sent. Significantly, this enables responsive decisions to be made by business stakeholders and appropriate actions taken which is underpinned by Apache Spark’s processing power. An alternative to Apache Spark is Apache Flink, which is designed with stream processing in mind and which the architecture is designed to support [[34]](https://www.zotero.org/google-docs/?lKk2Md). However, Apache Spark can handle a greater range of data processing and analysing tasks [[35]](https://www.zotero.org/google-docs/?8PGhze). Importantly, Apache Spark is designed as a general purpose cluster computing framework and is so well suited for a variety of tasks compared to Apache Flink.

3.1 Web-based Application

Given the requirements of creating a public facing application, the intention to analyse the data and the move towards permanent storage three areas are identified as important: data security, law and regulation and data limitation.

Data security is vital for any application which handles personal data. Globally, data breaches, such as [[36]](https://www.zotero.org/google-docs/?XcA6bi), where 26 billion records were leaked, illustrates this priority. Moreover, evolving cyber threats emphasise the seriousness of the value of data security [[37]](https://www.zotero.org/google-docs/?kdOJZH) for both financial and reputational reasons. This underscores the importance of data security for the company in order to grow financially and increase its reputation. To mitigate potential security issues, the following solutions are put forward: end-to-end encryption and compliance with international standards, such as ISO 27001. Justification for the use of end-to-end encryption is that it provides protection for information throughout its lifecycle. For instance, [[38]](https://www.zotero.org/google-docs/?1uMav1) puts forward that data should be encrypted at rest, in transit and when in use. Indeed, this is important for securing the data being collected from the online form, storage of this data and future analysis. Furthermore, [[39]](https://www.zotero.org/google-docs/?fr4yk6) proposes the use of a novel security and privacy application in response to the emergence of Artificial Intelligence. This highlights the importance of encryption and the need to adapt to emerging cyber security threats. Moreover, this will enable the company to be compliant with a standard, such as ISO 27001, which will give reassurance to stakeholders that there are processes and procedures in place [[40]](https://www.zotero.org/google-docs/?ui7lX0). Significantly, this will ensure that if there is a data breach it is responded to appropriately, damage is mitigated and procedures updated.

Law and regulation is another consideration and as a global company will need to be considered in each region or country. Indeed, legal and ethical standards are evolving, such as in Europe and the General Data Protection Regulation (GDPR) which sets a high international standard. However, [[41]](https://www.zotero.org/google-docs/?T7THzi) argues that GDPR causes a negative impact by limiting data collection and the secondary research possibilities. But, [[42]](https://www.zotero.org/google-docs/?0TX8p9) counters that GDPR is a strategic business decision which encourages trust between consumers and business. Importantly, GDPR compliance can facilitate trust of data collected from the online form. For example, Article 25 of the GDPR means data must be protected by default [[43]](https://www.zotero.org/google-docs/?XHjtnC) which means this is a vitally important legal requirement for the scaling of the company and operating in the European Union. Moreover, ISO 27001 certification will support this by showing that collected data can be kept safe. Therefore, given that there are similar laws and regulations in other countries, this presents a very strong argument for being legally compliant and ensuring a secure design with privacy built into the company's data collection and use policies.

Lastly, data limitation is considered as it links both legality and security together. For instance, it is required by GDPR, Article 5 [[43]](https://www.zotero.org/google-docs/?QjI6EA), that data is collected in a limited way and for a specific purpose. This highlights the importance of ensuring the type of data which is collected is only what is essentially needed for the company’s use and beneficial for the types of analysis required.

Reference List

[[1] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, ‘Data Cleaning: Overview and Emerging Challenges’, in *Proceedings of the 2016 International Conference on Management of Data*, San Francisco California USA: ACM, Jun. 2016, pp. 2201–2206. doi: 10.1145/2882903.2912574.](https://www.zotero.org/google-docs/?wcgLlK)

[[2] *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011. doi: 10.1016/C2009-0-19715-5.](https://www.zotero.org/google-docs/?wcgLlK)

[[3] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, ‘Data imbalance in classification: Experimental evaluation’, *Inf. Sci.*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.](https://www.zotero.org/google-docs/?wcgLlK)

[[4] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, ‘On the Class Imbalance Problem’, in *2008 Fourth International Conference on Natural Computation*, Jinan, Shandong, China: IEEE, 2008, pp. 192–201. doi: 10.1109/ICNC.2008.871.](https://www.zotero.org/google-docs/?wcgLlK)

[[5] P. Schober, C. Boer, and L. A. Schwarte, ‘Correlation Coefficients: Appropriate Use and Interpretation’, *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.](https://www.zotero.org/google-docs/?wcgLlK)

[[6] P. Lecca, ‘Machine Learning for Causal Inference in Biological Networks: Perspectives of This Challenge’, *Front. Bioinforma.*, vol. 1, p. 746712, Sep. 2021, doi: 10.3389/fbinf.2021.746712.](https://www.zotero.org/google-docs/?wcgLlK)

[[7] R. R. K. AL-Taie, B. J. Saleh, A. Y. Falih Saedi, and L. A. Salman, ‘Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq’, *Int. J. Electr. Comput. Eng. IJECE*, vol. 11, no. 6, p. 5229, Dec. 2021, doi: 10.11591/ijece.v11i6.pp5229-5239.](https://www.zotero.org/google-docs/?wcgLlK)

[[8] ‘The random forest algorithm for statistical learning’. Accessed: Feb. 24, 2024. [Online]. Available: https://journals.sagepub.com/doi/epub/10.1177/1536867X20909688](https://www.zotero.org/google-docs/?wcgLlK)

[[9] A. G. D’yakonov, ‘Solution Methods for Classification Problems with Categorical Attributes’, *Comput. Math. Model.*, vol. 26, no. 3, pp. 408–428, Jul. 2015, doi: 10.1007/s10598-015-9281-2.](https://www.zotero.org/google-docs/?wcgLlK)

[[10] R. Sullivan, ‘Machine-Learning Techniques’, in *Introduction to Data Mining for the Life Sciences*, Humana Press, 2011.](https://www.zotero.org/google-docs/?wcgLlK)

[[11] P. Ferragina and G. Vinciguerra, ‘Learned Data Structures’, in *Recent Trends in Learning From Data. Studies in Computational Intelligence*, vol. 896, Springer, Cham, 2019, pp. 5–41. Accessed: Feb. 29, 2024. [Online]. Available: https://doi.org/10.1007/978-3-030-43883-8\_2](https://www.zotero.org/google-docs/?wcgLlK)

[[12] J. C. Stoltzfus, ‘Logistic Regression: A Brief Primer’, *Acad. Emerg. Med.*, vol. 18, no. 10, pp. 1099–1104, Oct. 2011, doi: 10.1111/j.1553-2712.2011.01185.x.](https://www.zotero.org/google-docs/?wcgLlK)

[[13] B. Gupta, A. Rawat, A. Jain, A. Arora, and N. Dhami, ‘Analysis of Various Decision Tree Algorithms for Classification in Data Mining’, *Int. J. Comput. Appl.*, vol. 163, no. 8, pp. 15–19, Apr. 2017, doi: 10.5120/ijca2017913660.](https://www.zotero.org/google-docs/?wcgLlK)

[[14] ‘Decision tree methods: applications for classification and prediction - Shanghai Carchives of Psychiatry’. Accessed: Mar. 02, 2024. [Online]. Available: https://shanghaiarchivesofpsychiatry.org/en/215044.html](https://www.zotero.org/google-docs/?wcgLlK)

[[15] S. Kumar and I. Chong, ‘Correlation Analysis to Identify the Effective Data in Machine Learning: Prediction of Depressive Disorder and Emotion States’, *Int. J. Environ. Res. Public. Health*, vol. 15, no. 12, p. 2907, Dec. 2018, doi: 10.3390/ijerph15122907.](https://www.zotero.org/google-docs/?wcgLlK)

[[16] I. Palomares Carrascosa, H. K. Kalutarage, and Y. Huang, Eds., *Data Analytics and Decision Support for Cybersecurity*. in Data Analytics. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-59439-2.](https://www.zotero.org/google-docs/?wcgLlK)

[[17] Z. Khan, M. Naeem, U. Khalil, D. M. Khan, S. Aldahmani, and M. Hamraz, ‘Feature Selection for Binary Classification Within Functional Genomics Experiments via Interquartile Range and Clustering’, *IEEE Access*, vol. 7, pp. 78159–78169, 2019, doi: 10.1109/ACCESS.2019.2922432.](https://www.zotero.org/google-docs/?wcgLlK)

[[18] R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, ‘Variable selection for Naïve Bayes classification’, *Comput. Oper. Res.*, vol. 135, p. 105456, Nov. 2021, doi: 10.1016/j.cor.2021.105456.](https://www.zotero.org/google-docs/?wcgLlK)

[[19] Y. Liu *et al.*, ‘Hierarchical Independence Thresholding for learning Bayesian network classifiers’, *Knowl.-Based Syst.*, vol. 212, p. 106627, Jan. 2021, doi: 10.1016/j.knosys.2020.106627.](https://www.zotero.org/google-docs/?wcgLlK)

[[20] Y. Lin, H. Guo, and J. Hu, ‘An SVM-based approach for stock market trend prediction’, in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, Dallas, TX, USA: IEEE, Aug. 2013, pp. 1–7. doi: 10.1109/IJCNN.2013.6706743.](https://www.zotero.org/google-docs/?wcgLlK)

[[21] T. Dai and Y. Dong, ‘Introduction of SVM Related Theory and Its Application Research’, in *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, Shenzhen, China: IEEE, Apr. 2020, pp. 230–233. doi: 10.1109/AEMCSE50948.2020.00056.](https://www.zotero.org/google-docs/?wcgLlK)

[[22] S. Kolahi and L. Libkin, ‘An information-theoretic analysis of worst-case redundancy in database design’, *ACM Trans. Database Syst.*, vol. 35, no. 1, pp. 1–32, Feb. 2010, doi: 10.1145/1670243.1670248.](https://www.zotero.org/google-docs/?wcgLlK)

[[23] ‘Study of Various Normal forms and Functional Dependency’, *Int. J. Emerg. Technol. Adv. Eng.*.](https://www.zotero.org/google-docs/?wcgLlK)

[[24] M. Levene and M. W. Vincent, ‘Justification for Inclusion Dependency Normal Form’.](https://www.zotero.org/google-docs/?wcgLlK)

[[25] T. Bläsius, T. Friedrich, and M. Schirneck, ‘The Parameterized Complexity of Dependency Detection in Relational Databases’, p. 13 pages, 527908 bytes, 2017, doi: 10.4230/LIPICS.IPEC.2016.6.](https://www.zotero.org/google-docs/?wcgLlK)

[[26] P. Vassiliadis, M.-R. Kolozoff, M. Zerva, and A. V. Zarras, ‘Schema evolution and foreign keys: a study on usage, heartbeat of change and relationship of foreign keys to table activity’, *Computing*, vol. 101, no. 10, pp. 1431–1456, Oct. 2019, doi: 10.1007/s00607-019-00702-x.](https://www.zotero.org/google-docs/?wcgLlK)

[[27] J. Bhogal and I. Choksi, ‘Handling Big Data Using NoSQL’, in *2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops*, Gwangiu, South Korea: IEEE, Mar. 2015, pp. 393–398. doi: 10.1109/WAINA.2015.19.](https://www.zotero.org/google-docs/?wcgLlK)

[[28] C. Roy, K. Barua, S. Agarwal, M. Pandey, and S. S. Rautaray, ‘Horizontal Scaling Enhancement for Optimized Big Data Processing’, in *Emerging Technologies in Data Mining and Information Security*, vol. 755, A. Abraham, P. Dutta, J. K. Mandal, A. Bhattacharya, and S. Dutta, Eds., in Advances in Intelligent Systems and Computing, vol. 755. , Singapore: Springer Singapore, 2019, pp. 639–649. doi: 10.1007/978-981-13-1951-8\_58.](https://www.zotero.org/google-docs/?wcgLlK)

[[29] A. Chebotko, A. Kashlev, and S. Lu, ‘A Big Data Modeling Methodology for Apache Cassandra’, in *2015 IEEE International Congress on Big Data*, New York City, NY, USA: IEEE, Jun. 2015, pp. 238–245. doi: 10.1109/BigDataCongress.2015.41.](https://www.zotero.org/google-docs/?wcgLlK)

[[30] S. S. Chawathe, ‘Data Modeling for a NoSQL Database Service’, in *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York City, NY, USA: IEEE, Oct. 2019, pp. 0234–0240. doi: 10.1109/UEMCON47517.2019.8992924.](https://www.zotero.org/google-docs/?wcgLlK)

[[31] A. Lakshman and P. Malik, ‘Cassandra: a decentralized structured storage system’, *ACM SIGOPS Oper. Syst. Rev.*, vol. 44, no. 2, pp. 35–40, Apr. 2010, doi: 10.1145/1773912.1773922.](https://www.zotero.org/google-docs/?wcgLlK)

[[32] A. Verma, A. H. Mansuri, and N. Jain, ‘Big data management processing with Hadoop MapReduce and spark technology: A comparison’, in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, Indore, Madhya Pradesh, India: IEEE, Mar. 2016, pp. 1–4. doi: 10.1109/CDAN.2016.7570891.](https://www.zotero.org/google-docs/?wcgLlK)

[[33] R. Guo, Y. Zhao, Q. Zou, X. Fang, and S. Peng, ‘Bioinformatics applications on Apache Spark’, *GigaScience*, Aug. 2018, doi: 10.1093/gigascience/giy098.](https://www.zotero.org/google-docs/?wcgLlK)

[[34] J. Karimov, T. Rabl, A. Katsifodimos, R. Samarev, H. Heiskanen, and V. Markl, ‘Benchmarking Distributed Stream Data Processing Systems’, in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, Paris: IEEE, Apr. 2018, pp. 1507–1518. doi: 10.1109/ICDE.2018.00169.](https://www.zotero.org/google-docs/?wcgLlK)

[[35] E. Shaikh, I. Mohiuddin, Y. Alufaisan, and I. Nahvi, ‘Apache Spark: A Big Data Processing Engine’, in *2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM)*, Manama, Bahrain: IEEE, Nov. 2019, pp. 1–6. doi: 10.1109/MENACOMM46666.2019.8988541.](https://www.zotero.org/google-docs/?wcgLlK)

[[36] V. Petkauskas, ‘Mother of all breaches reveals 26 billion records: what we know so far’, Cybernews. Accessed: Mar. 08, 2024. [Online]. Available: https://cybernews.com/security/billions-passwords-credentials-leaked-mother-of-all-breaches/](https://www.zotero.org/google-docs/?wcgLlK)

[[37] H. Tao *et al.*, ‘Economic perspective analysis of protecting big data security and privacy’, *Future Gener. Comput. Syst.*, vol. 98, pp. 660–671, Sep. 2019, doi: 10.1016/j.future.2019.03.042.](https://www.zotero.org/google-docs/?wcgLlK)

[[38] V. V. Galushka, A. R. Aydinyan, O. L. Tsvetkova, V. A. Fathi, and D. V. Fathi, ‘System of end-to-end symmetric database encryption’, *J. Phys. Conf. Ser.*, vol. 1015, p. 042003, May 2018, doi: 10.1088/1742-6596/1015/4/042003.](https://www.zotero.org/google-docs/?wcgLlK)

[[39] M. Kantarcioglu and F. Shaon, ‘Securing Big Data in the Age of AI’, in *2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, Los Angeles, CA, USA: IEEE, Dec. 2019, pp. 218–220. doi: 10.1109/TPS-ISA48467.2019.00035.](https://www.zotero.org/google-docs/?wcgLlK)

[[40] C. Hsu, T. Wang, and A. Lu, ‘The Impact of ISO 27001 Certification on Firm Performance’, in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, Koloa, HI, USA: IEEE, Jan. 2016, pp. 4842–4848. doi: 10.1109/HICSS.2016.600.](https://www.zotero.org/google-docs/?wcgLlK)

[[41] D. Peloquin, M. DiMaio, B. Bierer, and M. Barnes, ‘Disruptive and avoidable: GDPR challenges to secondary research uses of data’, *Eur. J. Hum. Genet.*, vol. 28, no. 6, pp. 697–705, Jun. 2020, doi: 10.1038/s41431-020-0596-x.](https://www.zotero.org/google-docs/?wcgLlK)

[[42] I. M. Lopes, T. Guarda, and P. Oliveira, ‘Implementation of ISO 27001 Standards as GDPR Compliance Facilitator’, *J. Inf. Syst. Eng. Manag.*, vol. 4, no. 2, Aug. 2019, doi: 10.29333/jisem/5888.](https://www.zotero.org/google-docs/?wcgLlK)

[[43] O. Radley-Gardner, H. Beale, and R. Zimmermann, Eds., *Fundamental Texts On European Private Law*. Hart Publishing, 2016. doi: 10.5040/9781782258674.](https://www.zotero.org/google-docs/?wcgLlK)

Appendix A

Fig1.1 cfSsubsetEval BestFirst

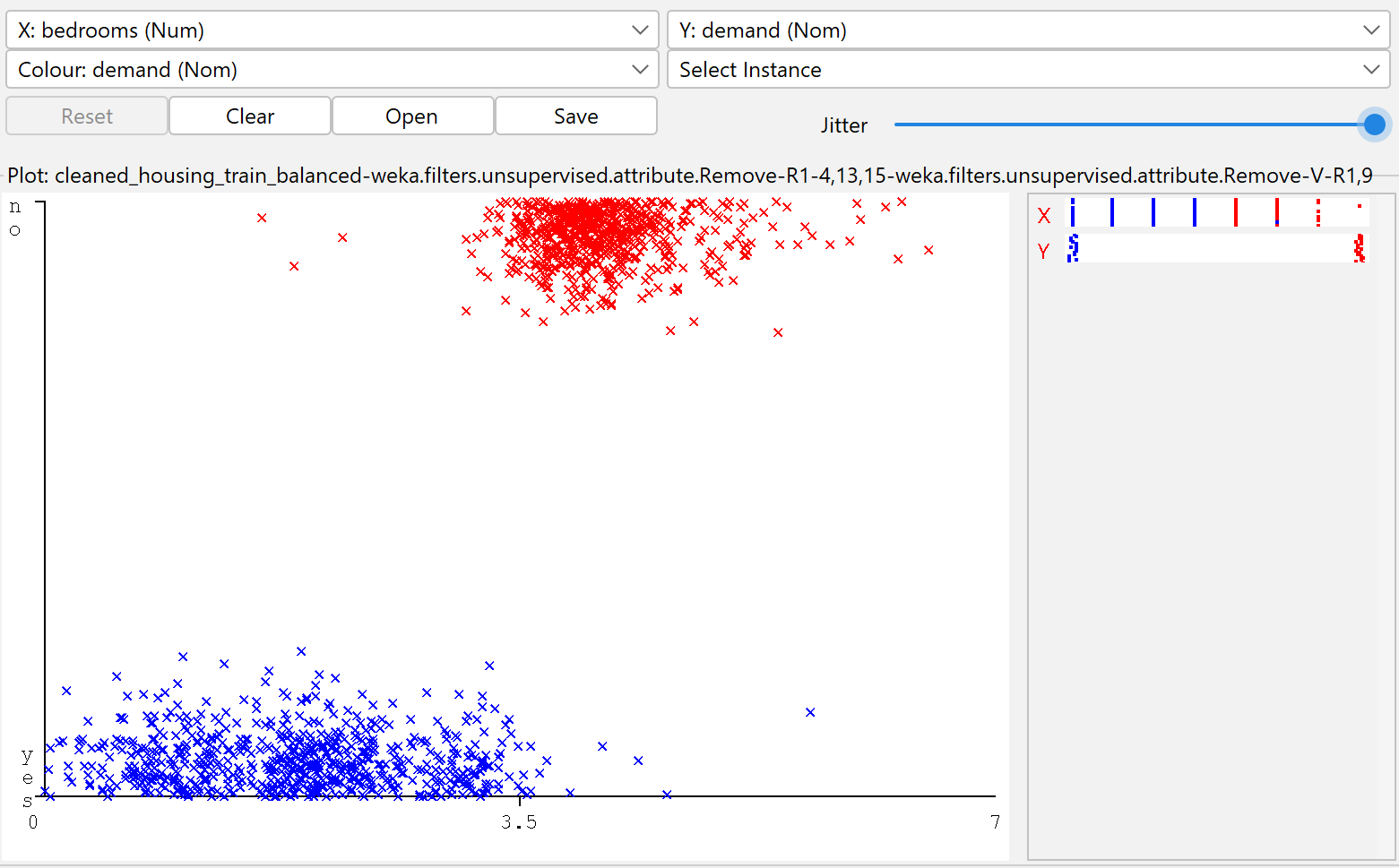
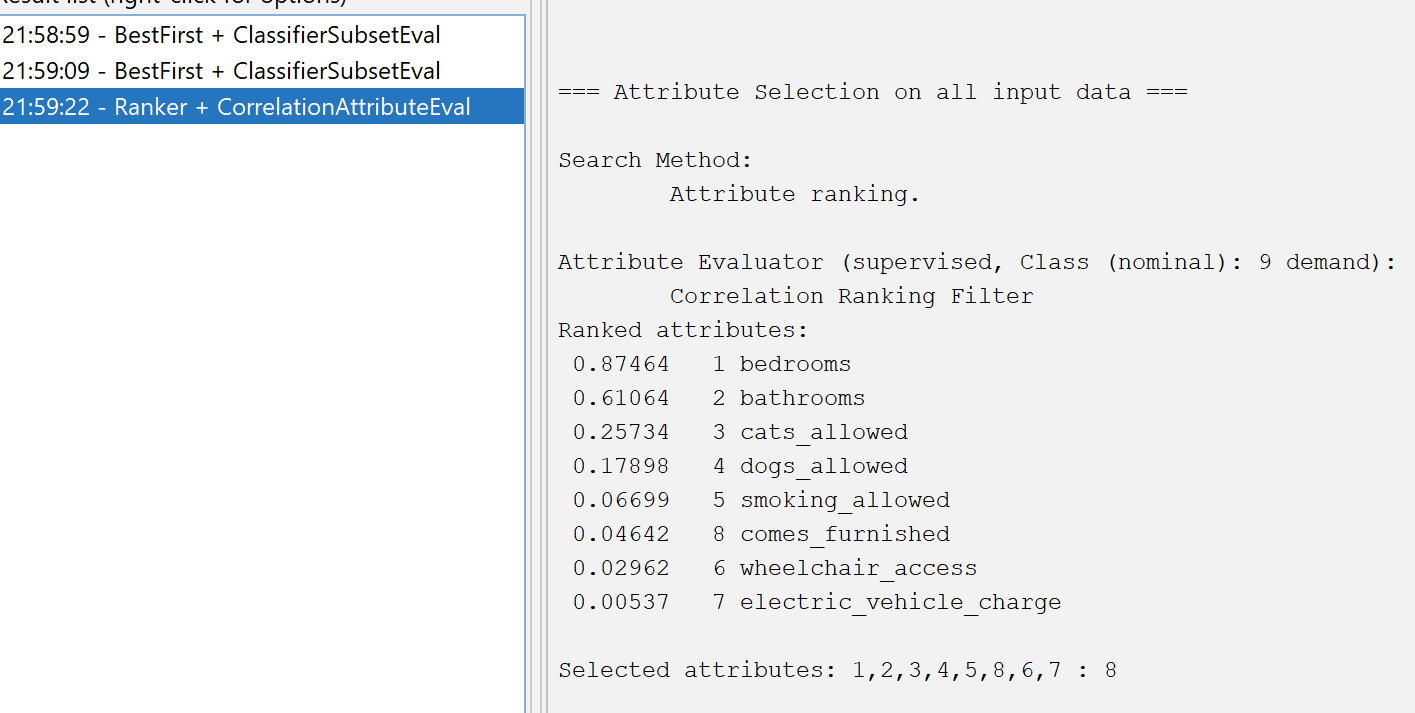


Fig 1.2 Correlation attributeEval + Ranker



Appendix B

Fig 2.1 Model 1: Four attributes - bedrooms, bathrooms and cats\_allowed discrete variables

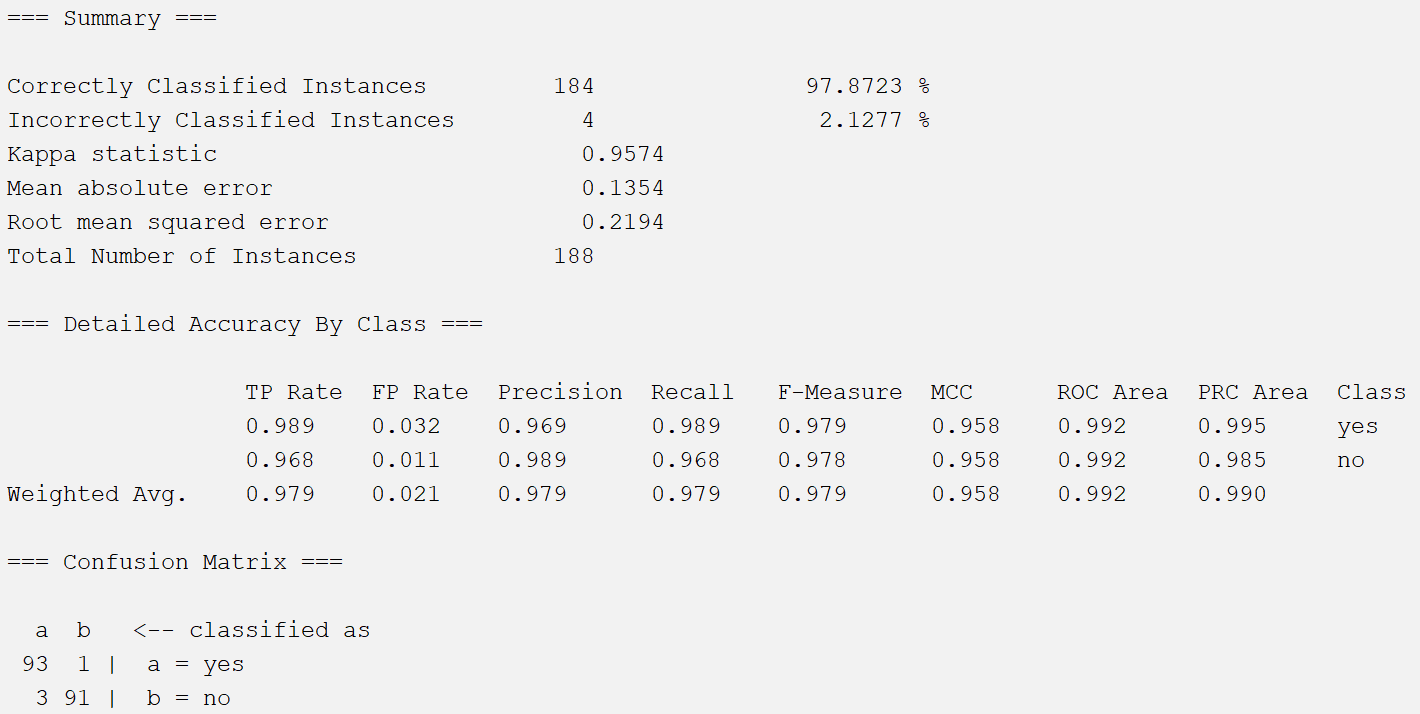
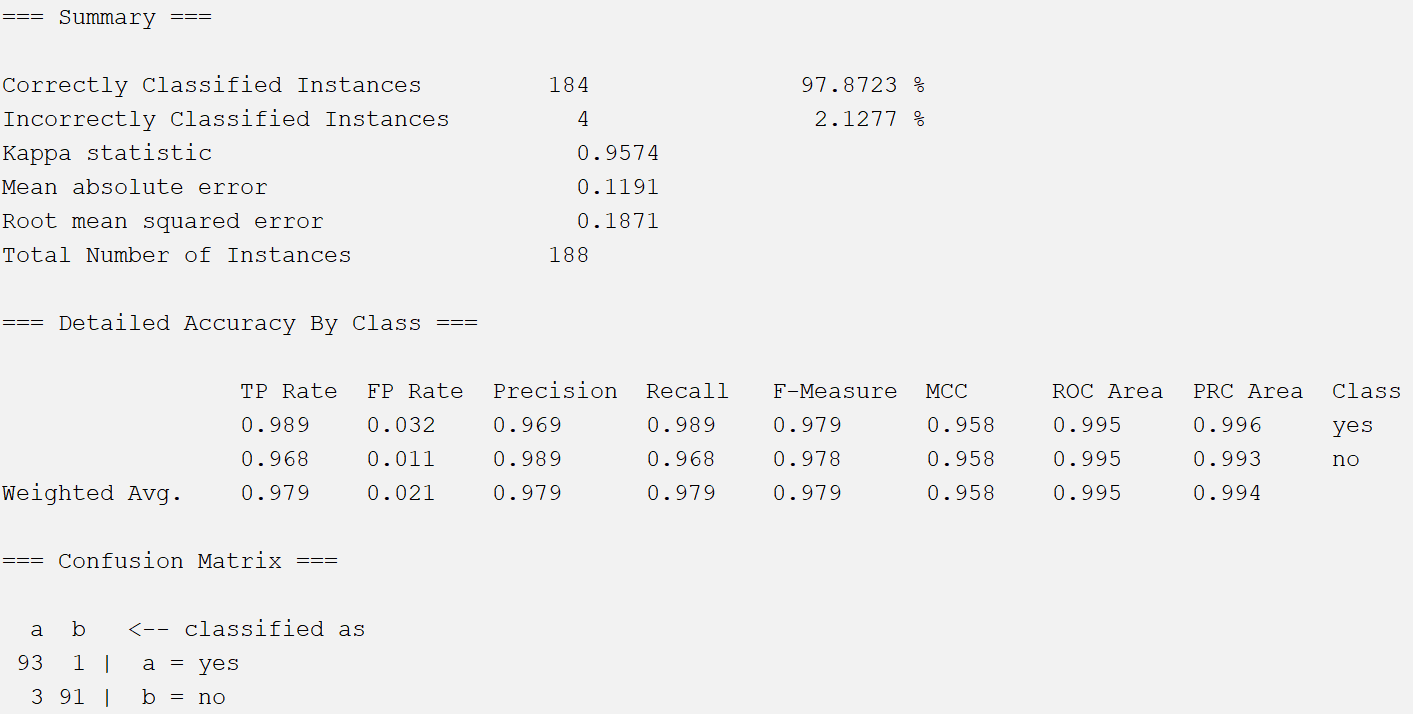


Fig 2.2 Model 2: Nine attributes - all discrete variables



Appendix C

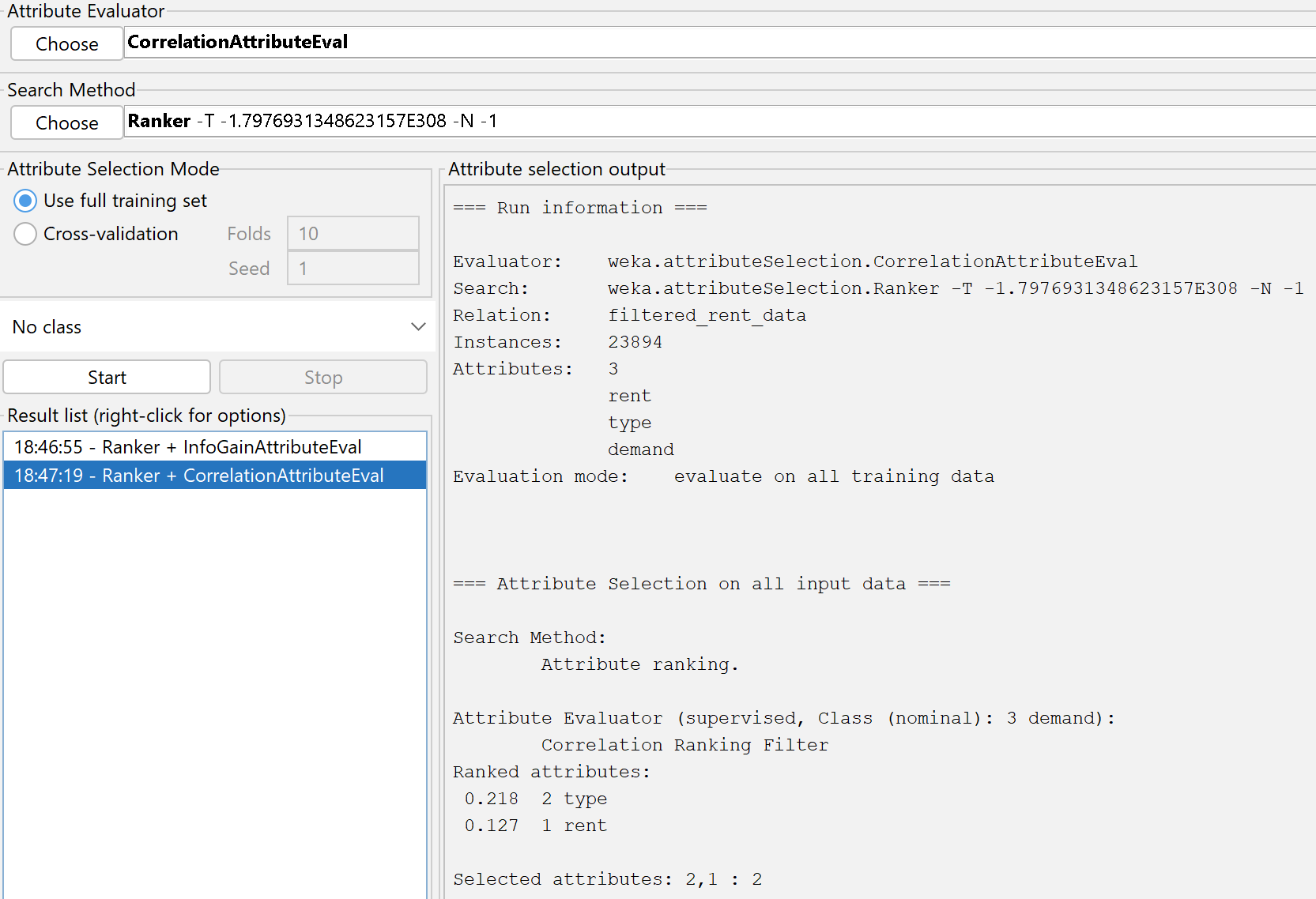
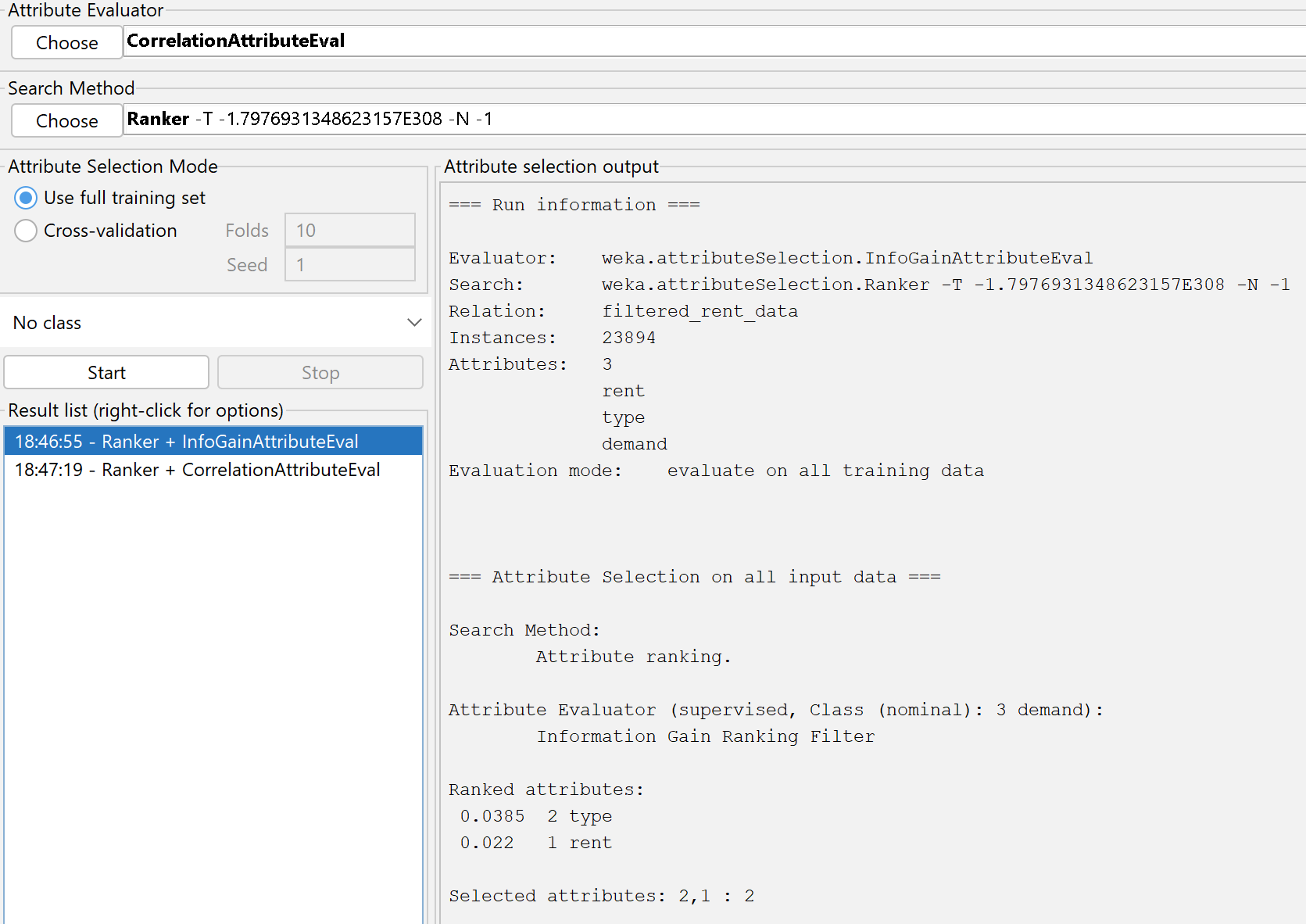
Fig 3.1 InfoGainAttributeEval Fig 3.2 CorrelationAttributeEval

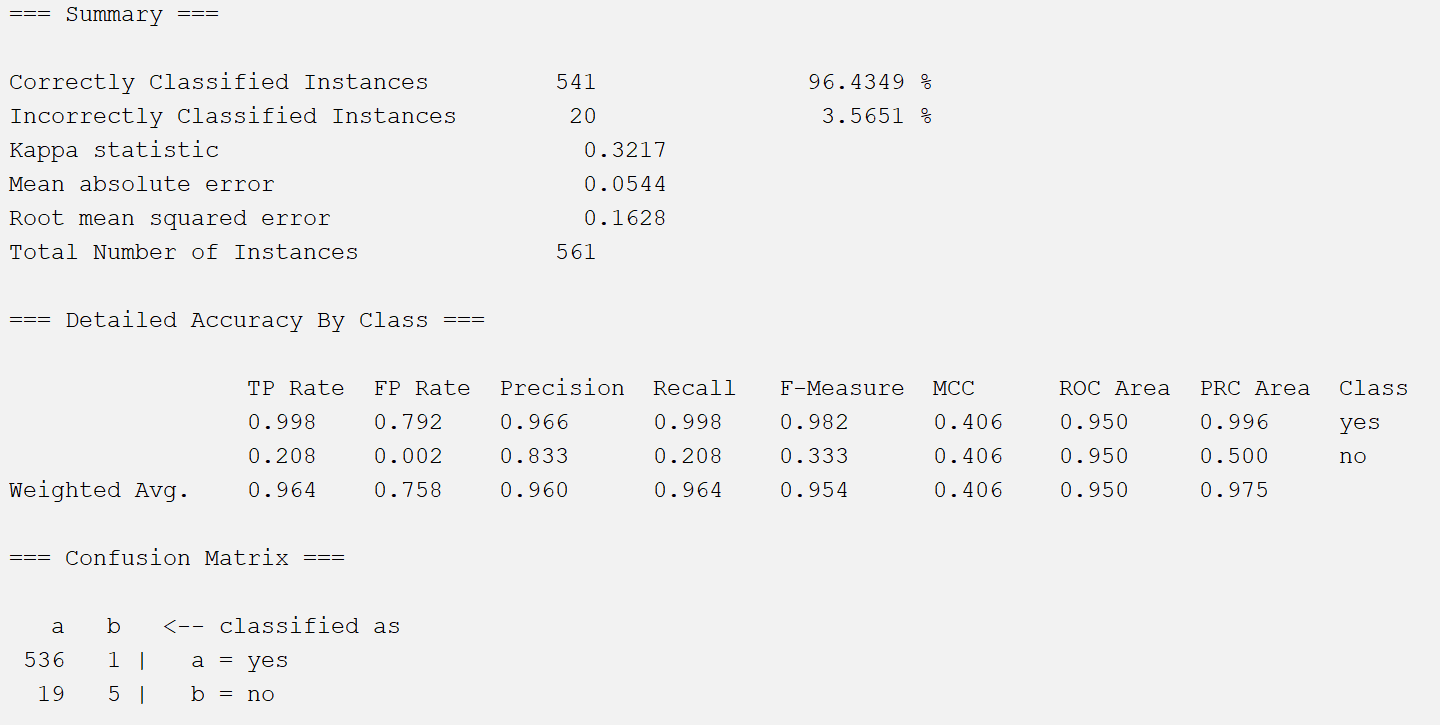
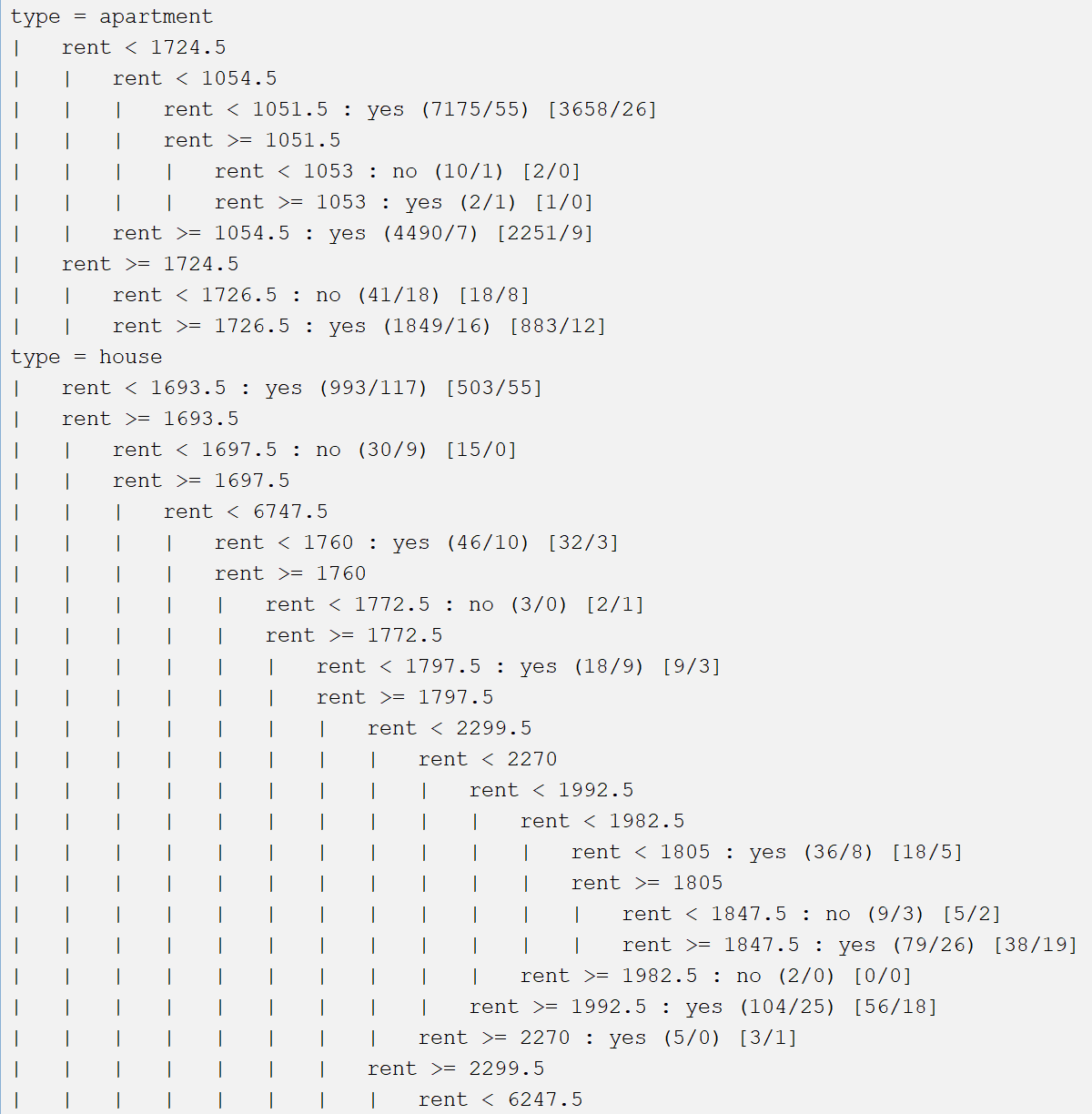
Fig 3.3 Decision Tree Results

Fig 3.4 Decision Tree Branching

Appendices D

Fig 4.1 Native Bayes Balanced Data Result

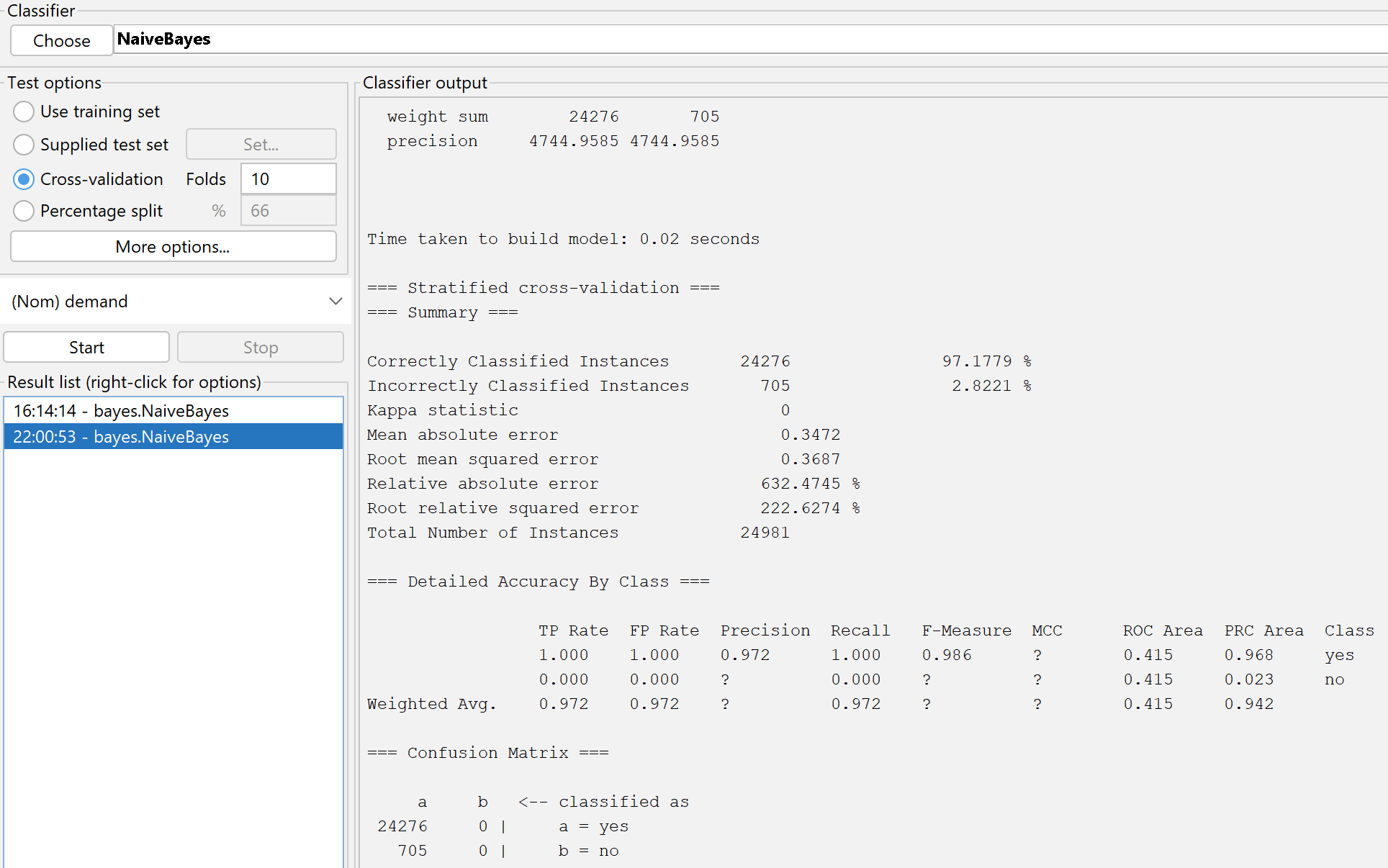
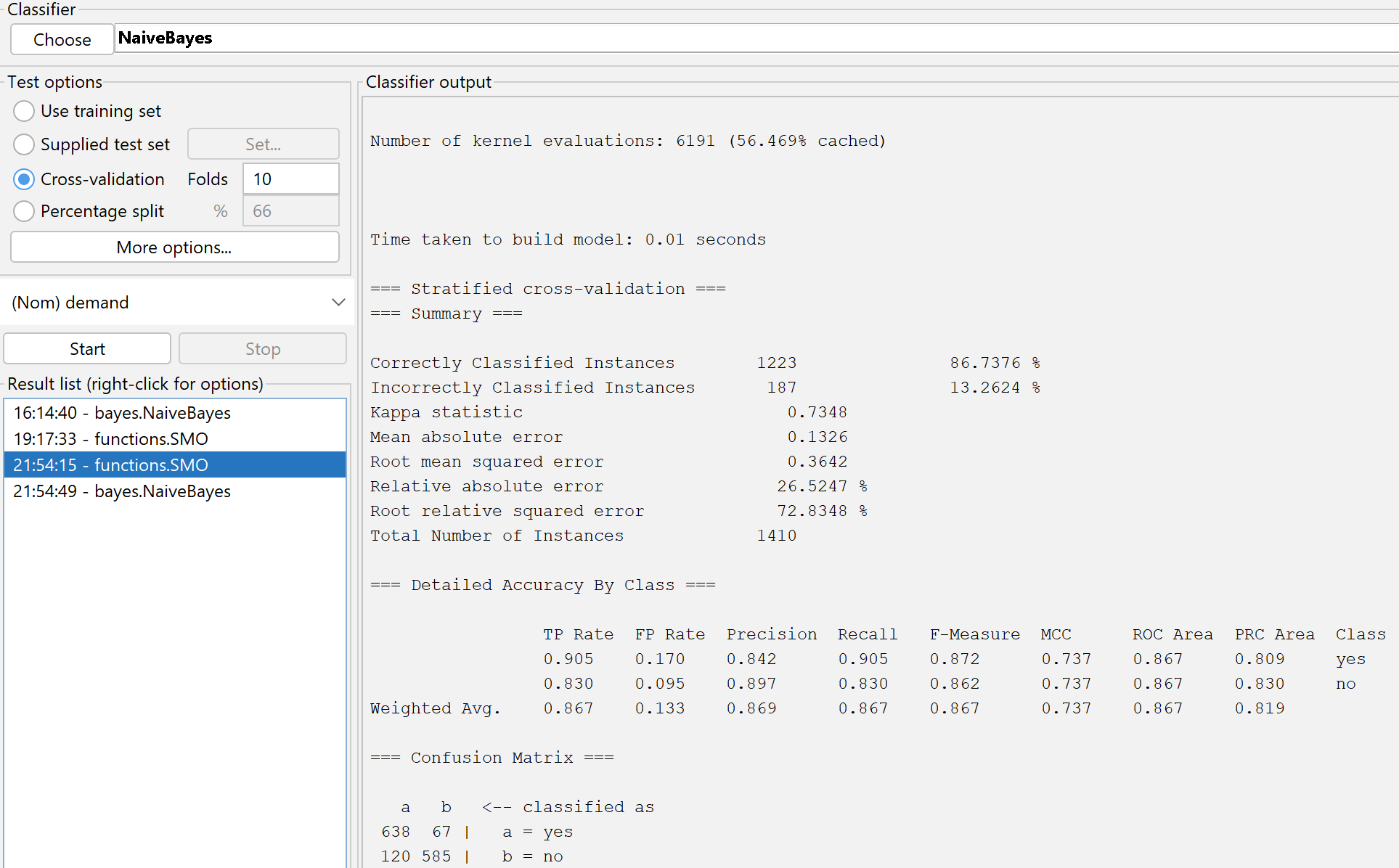
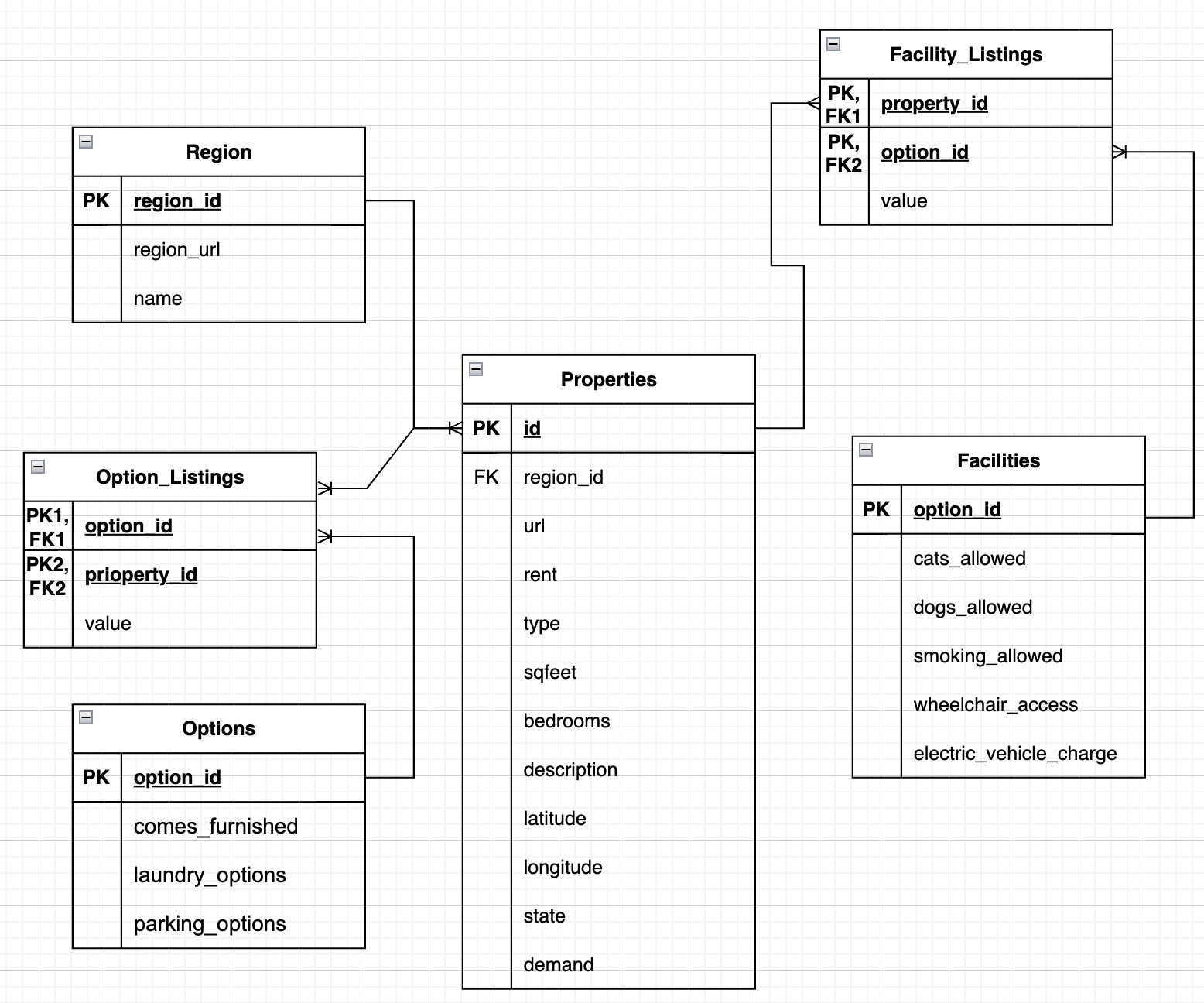


Fig 4.2 Support Vector Machine Balanced Data Results



Appendix E

Fig 5.1 Third Normal Form Database

# 

SQL Commands

i.

INSERT INTO Properties (id, url, region, rent, type, sqfeet, bedrooms, bathrooms, cats\_allowed, dogs\_allowed, smoking\_allowed, wheelchair\_access, electric\_vehicle\_charge, comes\_furnished, laundry\_options, parking\_options, demand, description, latitude, longitude, state)

ii.

SELECT description

FROM Properties

WHERE rent <= 1000 AND cats\_allowed = 1 AND dogs\_allowed = 1 AND state = 'ca';

iii.

SELECT state, AVG(rent) AS average\_rent

FROM Properties

GROUP BY state

ORDER BY average\_rent DESC;