University of York

Department of Computer Science

Analysis of Demand for Rental Properties using Weka and 'Big Data'

Including conversion to a relational database with example SQL and a discussion on privacy

Word Count: 2999 (including all text shown in tables)

2nd March 2024

Contents

1 Rental Demand Investigation

In this section, the Cross Industry Standard Process for Data Mining (CRISP-DM)[1] will be followed. The main benefits of this process are that it is software independent, and it does not follow any specific data analysis techinque[2]. Whilst other processes are available (e.g. SEMMA[3]), CRISP-DM remains the most popular[4].

The CRISP-DM model has six phases[5]:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

Each of these phases will discussed.

1.1 Business Understanding

The rental manager is concerned about a decrease in demand for rental properties. The referenced article discusses potential increases in supply or rent affordability issues as possible causes[6]. Therefore, the manager wishes to discover approaches to maximise demand.

1.2 Data understanding

The data was reviewed to understand each attribute's typical values. Diez et al. group variables into four categories: Continuous; discrete; nominal; and ordered [7]. Often though, only two types of data are used: nominal and ordinal with discrete variables included in both[8]. However, under this definition, all attributes for this dataset would be discrete which does not fit with requirements. Therefore, a mixture of both definitions was used to determine data types which can be seen in appendix A.

1.3 Data preparation

Tables 1.1 and 1.2 sumarise the data cleaning process. Updated csv files were imported to weka using the arff viewer and saved as arff files. The arff files were adjusted as follows:

- Nominal specification for 'type' attribute was standardised.

- Binary attributes changed from numeric to $\{1,0\}$ because discrete variables are typically nominal in practical situations and some classification processes can only use nominal attributes or might process numeric attributes unsatisfactorily[8].

The test data was cleaned in the same way as the training data[9].

Rule	Justification
DO NOT remove properties with 0 'sqfeet'.	5 out of the 7 properties with this figure showing as 'no' for demand. Error could be affecting demand? Removal could lead to incorrect conclusions and hence poor business strategy[10].
DO NOT remove instances with '.5' 'bathrooms'.	'.5' are rooms with toilets and wash basins only[11].
DO NOT remove instances with 0 'bedrooms'.	0 bedrooms denotes an open plan property[12].

Table 1.1 - Rules considered for application to CSV files

Rule	Justification
Remove all attributes except discrete variables shown in appendix A, 'type', 'sqfeet' and 'demand'.	Irrelevant data could cause problems with analysis[13]. Removal of redundant attributes can help to simplify models[2]
Remove entries with urls in 'demand'.	Incorrect data could cause problems with analysis[13].
Remove all instances with $0-9 rent.	Too low. Minimum wage in USA would allow $4,524 per year for rent [14] ($377 per month/$87 per week).
Remove all instances with 'type' apartment, 'sqfeet' greater than 650 and 'rent' less than $87.	Rent too low[14] for apartment not considered 'small'[15].
Remove instances with excessive 'rent' value ($21701907 and $90292 for apartments).	$22 million buys the most expensive estate in the Glen park Neighbourhood[16] - one of the best neighbourhoods to live in San Francisco[17].
Remove instance with excessive 'sqfeet' (8000000).	This is over 100 football pitches in size[18]
Correct entries in 'demand'.	Some show ' no', 'n' or 'n0' instead of 'no'. Easy to interpret what they should have been so can be corrected[19].
Check for missing entries.	As suggested[8] - none found

Table 1.2 - Rules applied to CSV files

1.4 Modelling

Initially, ZeroR and OneR models (where appropriate) were used for each requirement. This created benchmarks with which to compare performance[8].

Cross-validation was used when training to preserve the test data for final validation[20]. Cross validation was chosen as it makes good use of data and gives details of the performance[21]. For all models, default settings were used. Default settings are chosen sensibly to ensure results can be generated[8]. 2-3 models were tested for comparison purposes and the best performing model chosen[22].

Individual attributes were selected as required for each section in Weka's preprocessing tab (screenshot in appendix B).

1.4.1 Discrete Variables

Decision trees split between exact values so are good for classifying discrete data[23]. Support vector regression is designed to make predictions on discrete values[24]. Therefore, two decision trees and one support vector regression machine were chosen for analysing the discrete variables:

1. Random forest: Yields favourable error rates and has a degree of robustness to noise[25]
2. Sequential Minimal Optimization (SMO): Fast support vector machine[26]
3. J48: Finds and analyses patterns in data[27]

Other classifiers were considered including Naive Bayes but, this can produce poor results if there is a strong dependency among the attributes[28]. At the point of selecting classifiers, it was not known whether this was a feature of this dataset or not.

Ultimately, J48 was selected because of its ability to display the tree aiding with interpretation of the results (screenshot in appendix B).

1.4.2 Correlation

For correlation, regression tools were used. These attempt to find a good model for training points by reducing the prediction error[8].

'demand' was changed from nominal to binary using Weka's nominaltobinary filter to facilitate this[29] (screenshot in appendix B).

Three classifiers were chosen to test for correlation between 'demand' and a property's 'rent' and 'type':

- Sequential Minimal Optimization for Regression (SMOreg): Improved SMO model for regression[30].
- Multilayer Perceptron: Feed-forward neural network that continually repeats the learning phase, adjusting weights to reduce error[31].
- k-nearest neighbour classification (IBK): Nearest neighbour pattern classifier[32].

Overall, the IBK classifier performed best (see evaluation), so this model was chosen (screenshot in appendix B).

1.4.3 Property Size

Two tree classification tools were used to find an optimal range for 'demand' and 'sqfeet':

- J48: Finds and analyses patterns in data[27]
- REPTree: 'A fast decision tree learner'[33]

Tree learners were selected because they produce fast, easily understandable, accurate results[8]. Output from these trees was considered and values removed using Weka's RemoveWithValues filter[34].

Subsequently, the two tree classifiers were used again - this time without pruning - to help find the optimal range. The result from the J48 tool was easier to interpret, so this was selected.

The Simple Expectation Maximisation clustering tool (EM)[35] was used to verify the accuracy of the range - both with the full training data and test data.

Screenshots from this section are in appendix C.

1.5 Evaluation

In performing the cleaning, one area that was not considered in great detail was duplicate entries. Duplicate entries have a negative effect on the data, creating different results from machine learning tools[8]. However, one example of duplicate data was considered. There are two apartments listed with exactly the same attributes except for 'description' which was

written in different languages describing different things. Was this two listings for the same apartment? It is possible though, that in an apartment block, two apartments exist with exactly the same attributes (including 'latitude and 'long') but on different floors. Further work would need to be carried out to establish a method of truly identifying duplicates.

There are vastly more 'yes' values for 'demand' than there are 'no' values which might lead to unreliable results. Standard classifiers are not suited to learning with imbalanced datasets[36]. To improve the analysis, oversampling could have been performed which would have helped to deal with this imbalance in the data[37].

All models discussed were tested against the test data for evaluation. Separate arff files were created for each of the tasks by removing/adjusting attributes and saving to new files.

1.5.1 Discrete Variables

For 'yes' in 'demand', precision is high (0.999) for the training data. Whilst it is good that this is higher than for ZeroR benchmark figure (0.973), the question of overfitting must be considered. If the model fits too closely to the training data, then performance may be poor with unseen data[38].

For 'no' in 'demand', the precision figure is lower (0.804). ZeroR has no figure to compare with (it only checks for 'yes'). Using OneR as a benchmark (0.785), the model performed better again, in spite of the lower precision.

Comparing with the test data, for 'yes', the precision is identical, so overfitting has not occurred. However, for 'no', the precision drops to 0.760 for J48 with OneR at 0.748. Therefore,

J48 does not perform as well in predicting properties that would have low demand.

These results are confirmed in the confusion matrix with 0.10% and 0.05% of 'yes' incorrectly classified for the J48 and test data respectively. For 'no', the values are 19.64% and 24% exactly.

The J48 classifier's first split for the training and test data occurs on number of bedrooms being less than or equal to 3 with the vast majority of 'yes' results for 'demand' falling in this category. OneR also chooses to split on bedrooms at the same whole number of bedrooms.

See Appendix D for results.

1.5.2 Correlation

The correlation is weak for ZeroR, SMOreg and MultilayerPerceptron, and moderate for the selected model, IBK (0.4291)[39]. Converting mean absolute errors to percentages allows different models to be compared[40]. Comparing these Relative Absolute error percentages, IBK's value lies in the middle of SMOreg and MultilayerPerceptron's values, but all three are above 50% which is high[41].

When using the test data with J48, the correlation coefficient was lower than with cross-validation and the error percentages were higher.

Overall, these results show the correlation analysis cannot be relied upon.

See appendix D for results.

1.5.3 Property Size

The J48 classifier provided an upper boundary for 'sqfeet'as 1239. The majority of 'yes' values for demand fall at or below this figure.

For finding a lower boundary, the unpruned tree produced 175. Therefore, it could be concluded that the optimal range for square feet is 175 to 1239.

This is confirmed by the clustering tool. Using the tool, the mean values are approximately 744 and 873 square feet for the cluster with the highest proportion of 'yes' values for the training and test data respectively. These fall well within this range (screenshot in appendix C).

1.6 Deployment

The rental manager should note that properties with 3 bedrooms or fewer are likely to have high demand. This does not mean that properties with more than 3 bedrooms will have no demand. The OneR analysis confirms that properties with 7 bedrooms or more also have high demand.

The correlation results show that the manager should not consider 'rent' or 'type' for high demand due to the high absolute errors. However, there may be other machine learning tools that have not been tested that might show stronger correlation.

For 'sqfeet', the manager could expect properties with a value between 175 and 1239 to be in high demand. Predicting low demand properties would not be as reliable.

2 Storing data and scalable solutions

2.1 Database design

Initially, it was determined no repeating attributes were present in the dataset. Next, a check was made for potential composite attribute types, but none were identified. It was thus concluded the data was already in first normal form[42].

Normalisation of the data was then considered for 2NF[42]. A check was made for functional dependency of the attributes. The attributes 'state', 'region' and 'region_url' were not functionally dependent on the primary key for the data ('id'). These are in fact dependent on 'latitude' and 'long'. 'region_url' could also be derived from 'url'. A new entity was therefore created to contain these attributes with 'latitude' and 'long' being a new composite primary key.

Subsequently, conversion to 3NF was carried out[42]. 'state' and 'region_url'could be derived from 'region' as well as the composite primary key. Therefore, new entities were created for these.

The final entities can thus be sumarised as follows:

Property(id, url, 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, long*)

Location(latitude, long, *region*)

Region(region, region_url, state)

Relationships and cardinalities were established and a Unified Modelling Language standard Entity Relation diagram was created following models set out by Lemahieu[42] and OpenClassrooms.com[43] (see appendix E). The required SQL can be also found in appendix E.

2.2 Consideration of scaling

In considering scaling, assumptions about the scenario are required. Firstly, the data pertaining to each property will have the same attributes in each international location as the original dataset. Secondly, requirements for determining effects on demand will be the same. It is envisaged though that in time, the models will produce different characteristics for high and low demand properties than those for the original dataset - especially for international locations.

With the large amounts of data expected, the data should be stored in territories close to each international office to facilitate the rapid-response system required[44]. However, this approach would lead to isolated instances of each dataset for a territory which may not be desirable. Individual territories may have very different results when considering reasons for

high and low demand though, so having isolated databases may in fact be an advantage. Therefore, the best approach would be to keep the datasets separate with the facility to combine them as required.

In general, there are several options when it comes to spreading workload over multiple computers:

- Over the Network - effective for simultaneously processing information for multiple users[45].
- Message Passing Interface (MPI) - useful for situations where multiple computers can run a program in parallel[46].
- Specific Software such as VMware ESXI - works directly on a server, creating partitions in order to effectively maximise the server's resource usage [47].
- Cloud Computing - provides nearly unlimited processing and storage facilities over public or private networks[48].

In considering the above options, network bottlenecks could cause problems using MPI systems or specific software such as VMware ESXi in a rapid response system[49] [50] so would not be suitable to this application. A cloud based system such as Microsoft Azure[51] could be used to scale-up the business. Using Azure Virtual Machines to provide an 'Infrastructure as a Service' system within an Azure Region would not require much configuration[52]. It would remove the need for having large on-site premises to store the data with a large IT team to maintain the systems thus reducing the cost[52]. Cloud computing is highly scaleable[53], so should higher usage be required, this could easily be achieved using Azure virtual machines[54].

However, rapid response decision-making requirements cannot easily be met by cloud systems due to latency issues[55]. Instead, a hybrid model is proposed similar to Yang et al.'s edge-cloud system proposal for manufacturing[55]. The rapid response requirements could be met via utilisation of a physical, local area network in each international office. Functionality that does not require a rapid response - such as analysis required to find typical high demand properties - can be carried out using virtual machines in the cloud. Cloud providers have implementations to facilitate this - for example Azure Stack Edge which is a 'Hardware as a Service' system[56].

A possible drawback with this approach however might be the requirement to have a local copy of the dataset as well as one in the cloud. However, the dataset stored in each local office could be viewed as the current version and the cloud version as the backup version. Analysis for demand (and any other non rapid-response analysis) could be performed on virtual machines in the cloud (for example, in Azure[56]) as this would not need the most up-to-date information. Nightly backups of the current dataset to the cloud could be carried out from the on-site versions. This would remove the requirement for on-site backup systems, thus reducing cost[57]. International datasets could then be merged through the cloud for further analysis if required.

3 Considering public-facing application

The virtues of implementing privacy have been discussed since the 1990s[58], but this was usually implemented in the 'back-end' in software[59]. By 2014, organisations were considering it part of overall design[60]. Generally, privacy regulations now incorporate 'Privacy by Design'[61].

Some of the most stringent regulations are in the European Union (EU)[62] - The General Data Protection Regulations (GDPR). In the UK's version of this legislation, there are seven key aspects to consider[63]. There are key similarities with other legislation for privacy across the globe. For example, common elements exist between the GDPR and privacy regulations in Asia[64], and, despite key differences between GDPR and the California Consumer Privacy Act(CCPA), there are still similarities[65]. The three most salient aspects of international regulations for this application will be considered. All three are requirements of GDPR[63], CCPA[66] [67] and South Korea's Personal Information Protection Act - PIPA[68] which is included as another example of stringent privacy laws[69]. They are:

1. Personal data is stored securely. This includes prevention of fraudulent access and protection against loss.

International standard ISO/IEC27001 defines requirements for information security management[70]. When the application is implemented, it should conform to this standard. Risks need to be identified, and the level of acceptance determined[70]. Risks can be evaluated using the OWASP risk rating methodology[71] and appropriate treatments applied. One risk to consider is false positive authentication. This could be mitigated against through an appropriate authentication method. For the application here, users are likely to favour ease over security as few credentials will need to be stored. 0Auth could be a good method to choose[72]. There would be other risks to consider, but only by working through a specific risk assessment following ISO/IEC27001 could these risks and their appropriate treatments be established.

2. Notification of intended use of collected data.

   Users often do not read privacy policies[73]. However, they are more likely to read it if presented to them before they can continue than if they have to choose to read it[74]. Therefore, for this application, it should feature in a 'Terms and Conditions' page to which the user must agree before enrolling.

   It should be noted that notifying an individual of intention to sell their data would NOT comply with GDPR since it contravenes article 5.1(a, b and c) of the GDPR[63].

3. Right to opt-out of provision of personal data.

   Both GDPR and PIPA require that individuals opt-in to specific mechanisms designed by companies for using their personal data[75] [68]. The CCPA is less prescriptive because users must choose to opt-out of these features[66]. It is recommended that an opt-in approach be adopted by the new application. This would meet the stricter

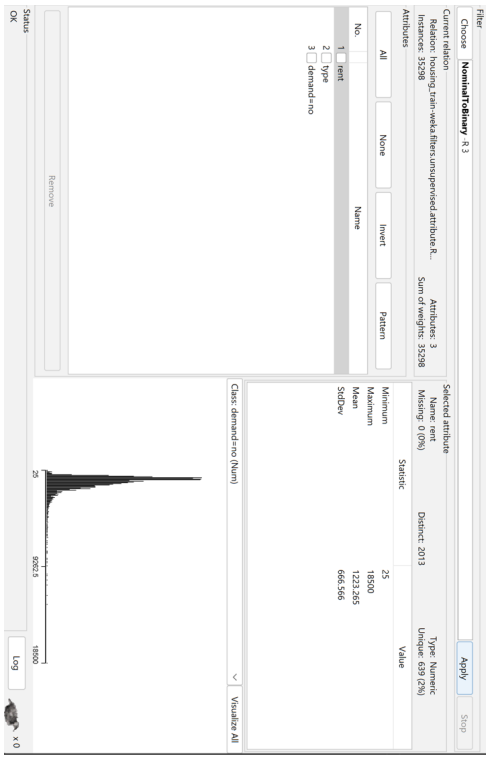
requirements of GDPR and PIPA and not go against requirements of CCPA. This could also be implemented under 'Terms and Conditions'.

Appendix A - Definitions of Data Types

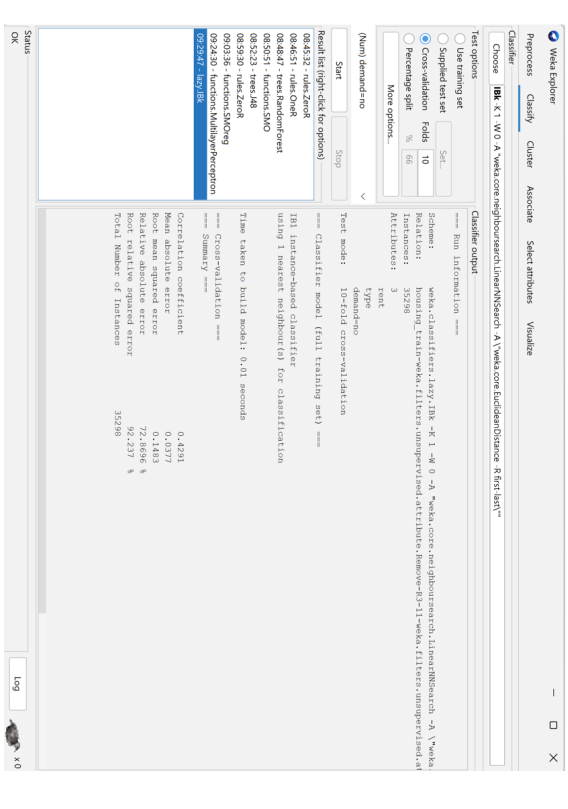
Attribute	Data Type	Attribute	Data Type
id	Numerical: Discrete	smoking_allowed	Numerical: Discrete
url	Categorical: Nominal	wheelchair_access	Numerical: Discrete
region	Categorical: Nominal	electric_vehicle_charge	Numerical: Discrete
region_url	Categorical: Nominal	comes_furnished	Numerical: Discrete
rent	Numerical: Discrete	laundry_options	Categorical: Nominal
type	Categorical: Nominal	parking_options	Categorical: Nominal
sqfeet	Numerical: Continuous	demand	Categorical: Nominal
bedrooms	Numerical: Discrete	description	Categorical: Nominal
bathrooms	Numerical: Discrete	latitude	Numerical: Continuous
cats_allowed	Numerical: Discrete	long	Numerical: Continuous
dogs_allowed	Numerical: Discrete	state	Categorical: Nominal

Appendix B - Screenshots for discrete attribute and correlation analysis

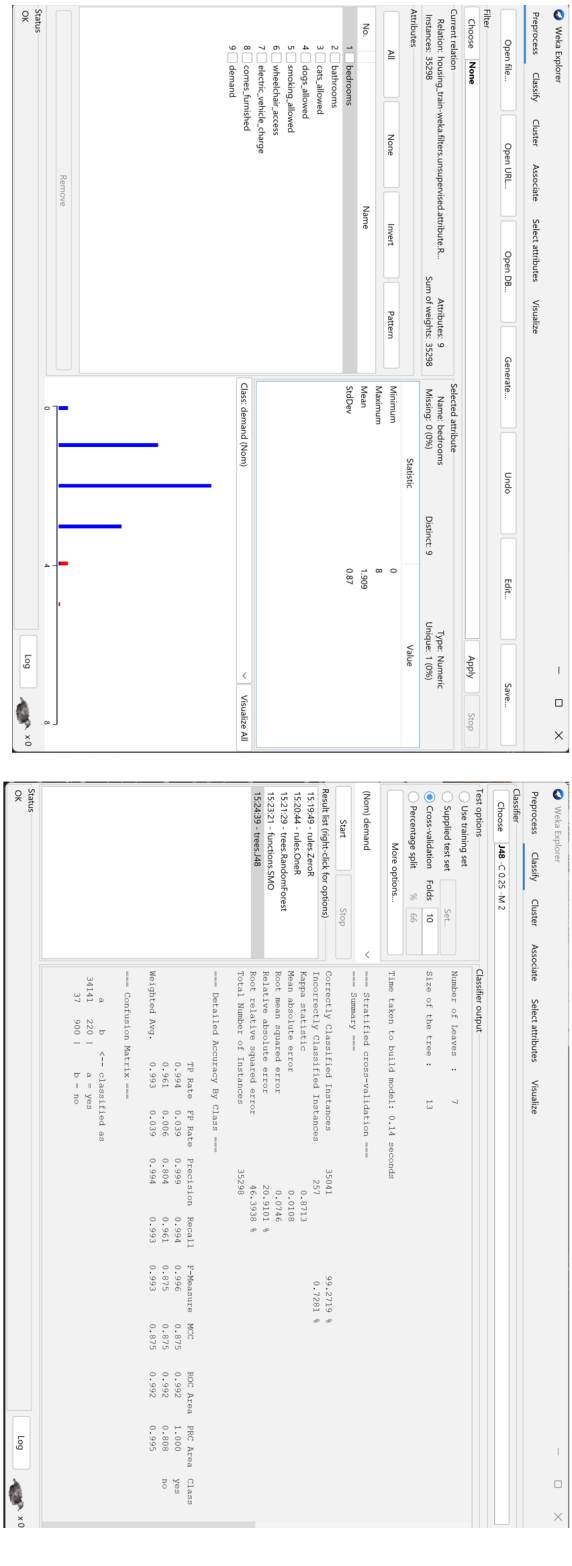
Screenshot for discrete attribute selection:



Changing demand from nominal to binary (includes correlation attribute selection):

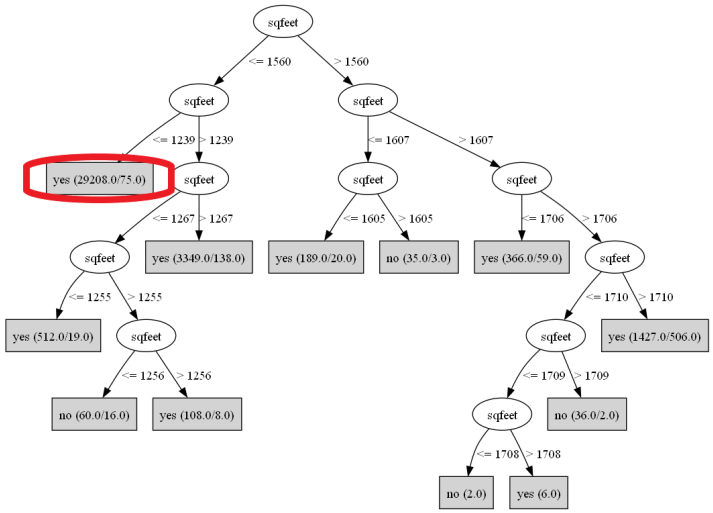


J48 - 10 Fold Cross Validation for discrete attributes:

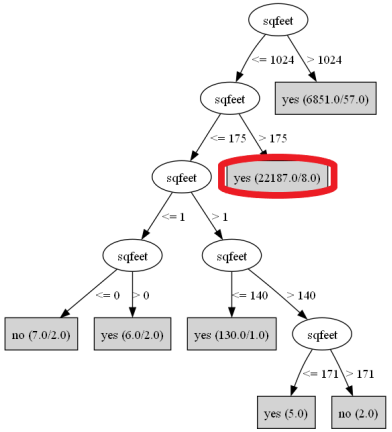


k-nearest neighbour classification (IBK) - 10 Fold Cross Validation for correlation analysis:

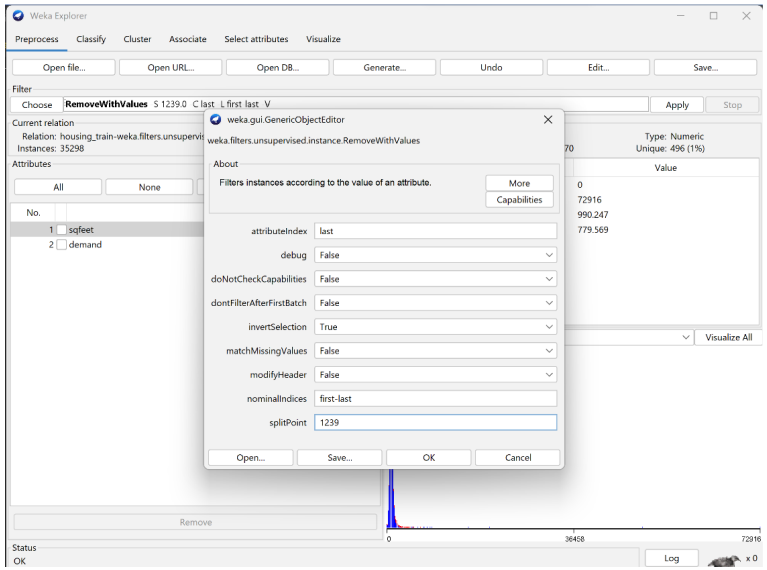
Initial J48 tree showing majority of instances less than or equal to 1239 feet:



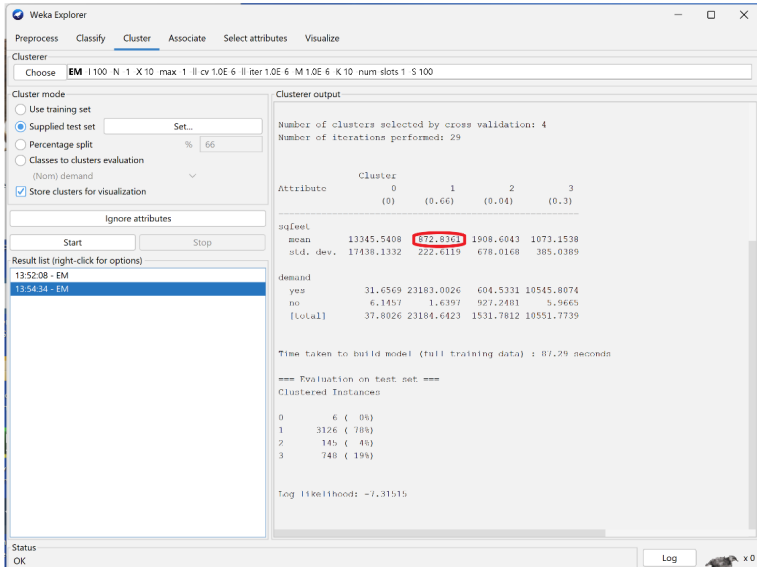
Second J48 tree showing majority of remaining instances greater than 175 feet:



Removal of instances with 'sqfeet' values above 1239 (also shows attribute selection):



Results from clustering tool when used against test data:



Specific results for discrete value analysis

Results for cross-validation when demand is 'no'

Classifier	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
ZeroR	0	0	?	0	?	?	0.499	0.026
OneR	0.994	0.007	0.785	0.994	0.877	0.88	0.993	0.78
Random Forest	0.943	0.006	0.807	0.943	0.87	0.869	0.994	0.809
SMO	0.995	0.008	0.783	0.995	0.876	0.879	0.994	0.779
J48	0.961	0.006	0.804	0.961	0.875	0.875	0.992	0.808

Results for cross-validation when demand is 'yes'

Classifier	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
ZeroR	1	1	0.973	1	0.987	?	0.499	0.973
OneR	0.993	0.006	1	0.993	0.996	0.88	0.993	1
Random Forest	0.994	0.057	0.998	0.994	0.996	0.869	0.994	1
SMO	0.992	0.005	1	0.992	0.996	0.879	0.994	1
J48	0.994	0.039	0.999	0.994	0.996	0.875	0.992	1

Results against test data when demand is 'no'

Classifier	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
ZeroR	0	0	?	0	?	?	0.5	0.029
OneR	1	0.01	0.748	1	0.856	0.861	0.995	0.748
Random Forest	0.948	0.008	0.769	0.948	0.849	0.849	0.997	0.836
SMO	1	0.01	0.748	1	0.856	0.861	0.995	0.748
J48	0.983	0.009	0.76	0.983	0.857	0.86	0.997	0.819

Results for Classifiers tested against test datawhen demand is 'yes'

Classifier	TP	FP	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
ZeroR	1	1	0.971	1	0.985	?	0.5	0.971
OneR	0.99	0	1	0.99	0.995	0.861	0.995	1
Random Forest	0.992	0.052	0.998	0.992	0.995	0.849	0.997	1
SMO	0.99	0	1	0.99	0.995	0.861	0.995	1
J48	0.991	0.017	0.999	0.991	0.995	0.86	0.997	1

Specific results for corrolation analysis between rent, type and demand

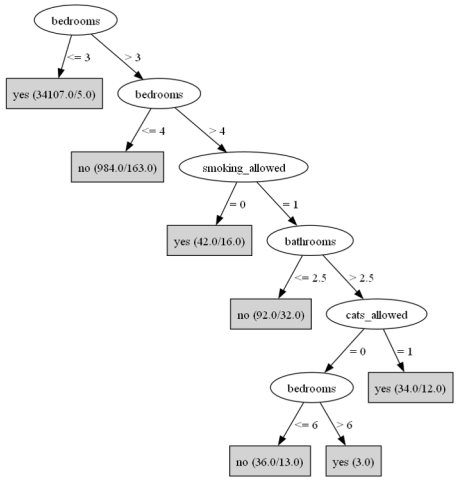
Results for cross-validation:

Clasifier	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Relative Root Mean Squared Error(%)
ZeroR	-0.0171	0.0517	0.1608	100	100
SMOreg	0.1698	0.0266	0.1629	51.5422	101.3091
MultilayerPerceptron	0.2839	0.0444	0.1546	85.921	96.1979
IBK	0.4291	0.0377	0.1483	72.8696	92.237

Results against test data:

Clasifier	Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Relative Root Mean Squared Error(%)
ZeroR	0	0.0538	0.1673	100	100
SMOreg	0.1916	0.0299	0.1695	55.6141	101.302
MultilayerPerceptron	0.2859	0.0628	0.1651	116.6237	98.7031
IBK	0.1869	0.0569	0.1933	105.689	115.531

J48 Tree for discrete values used against test data



i. SQL for new line of data:

New line of data to be added (It is assumed that this is a new region and that details for 'region', 'region_url' and 'state' have not been entered previously):

id	url	region
1	https://hes.craigslist.org/apa/d/heslington-huge-three-bedroom-among-the/1.html	Heslington

region_url	rent	type	sqfeet	bedrooms	bathrooms	cats_allowed	dogs_allowed
https://hes.craigslist.org/	1195	apartment	1908	3	2	1	1

smoking_allowed	wheelchair_access	electric_vehicle_charge	comes_furnished	laundry_options	parking_options
1	0	0	0	laundry on site	street parking

demand	description	latitude	long	state
yes	Apartments In Heslington York	53.944851	-1.048814	York

```
INSERT INTO Property
VALUES('1',
'https://hes.craigslist.org/apa/d/heslington-huge-three-bedroom-among-the/
1.html', '1195', 'apartment', '1908', '3', '2', '1', '1', '1', '0', '0', '0', 'laundry on site',
street parking', 'yes', 'Apartments In Heslington York', '53.944851', '-1.048814');

INSERT INTO Location
VALUES('53.944851', '-1.048814', 'Heslington');

INSERT INTO Region
Values('Heslington', 'https://hes.craigslist.org/', 'York');
```

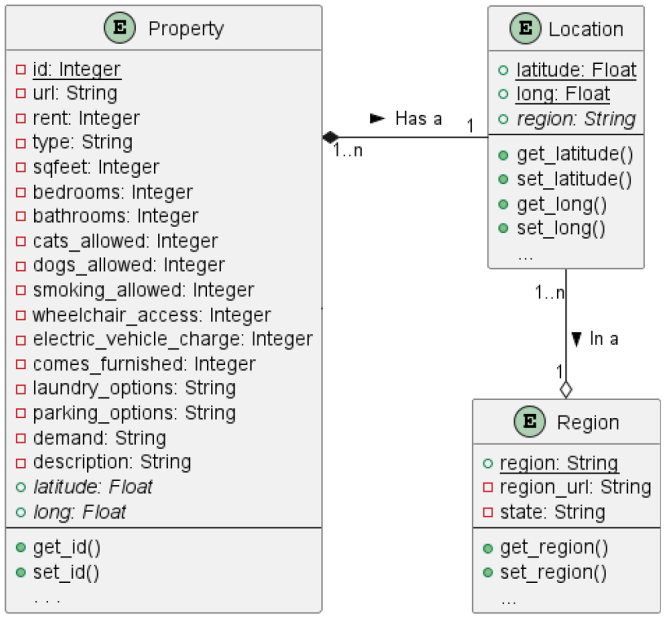
ii. SQL to Extract Description for properties with rent less than or equal to 1000:

```
SELECT A.description FROM Property AS A
JOIN Location AS C ON A.latitude = C.latitude AND A.long = C.long
JOIN Region AS B ON C.region = B.region
WHERE A.rent <= 1000 AND A.cats_allowed = 1
AND A.dogs_allowed = 1 AND B.state = 'ca';
```

iii. SQL to extract the average rental value for each state:

```
SELECT B.state, AVG(A.rent) AS AverageValue FROM Region AS B
JOIN Location AS C ON B.region = C.region
JOIN Property AS A ON C.latitude = A.latitude AND C.long = A.long
GROUP BY B.State;
```

Entity Relation Diagram:



References

[1] P. Chapman, 'Crisp-dm 1.0: Step-by-step data mining guide,' 2000. [Online]. Available: https://api.semanticscholar.org/CorpusID:59777418.

[2] J. D. Keleher and B. Tierney, *DATA SCIENCE*. Massachusetts Institute of Technology, 2018.

[3] S. I. Inc, *SAS Enterprise Miner 14.3: Reference Help. Cary, NC*. SAS Institute Inc, 2017.

[4] G. Mariscal, Ó. Marbán and C. Fernández, 'A survey of data mining and knowledge discovery process models and methodologies,' *The Knowledge Engineering Review*, vol. 25, no. 2, pp. 137–166, 2010. DOI: 10.1017/S0269888910000032.

[5] N. HOTZ, *What is crisp dm?* Accessed 2024-23-01. [Online]. Available: https://www.datascience-pm.com/crisp-dm-2/.

[6] R. Donnell, *Rental market report: What's happening to rents?* Accessed 2024-23-01. [Online]. Available: https://www.zoopla.co.uk/discover/property-news/rental-market-report-march-2023/.

[7] D. Diez, M. Cetinkaya-Rundel and C. D. Barr, *OpenIntro Statistics, Fourth Edition*. OpenIntro.org, 2019.

[8] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, *Data Mining Practical Machine Learning Tools and Techniques Fourth Edition*. Morgan Kaufmann - an imprint of Elsevier, 2017.

[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.

[10] Tableau, *Guide to data cleaning: Definition, benefits, components, and how to clean your data*, Accessed 2024-09-12. [Online]. Available: https://www.tableau.com/learn/articles/what-is-data-cleaning.

[11] C. Klein, *Bathrooms defined*, Accessed 2024-09-02. [Online]. Available: https://realestatechuck.com/bathrooms-defined/.

[12] U. O. of Administrative Rules, *Rule 840: Lead-based paint program purpose, applicability, and definitions*, Accessed 2024-09-02. [Online]. Available: https://adminrules.utah.gov/public/rule/R307-840/Current%20Rules?.

[13] M. H. et al., 'Data cleansing mechanisms and approaches for big data analytics: A systematic study,' *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 1, pp. 99–111, 2023. DOI: 10.1007/s12652-021-03590-2.

[14] K. Amadeo, *How much rent can americans afford on minimum wage?* Accessed 2024-09-02. [Online]. Available: https://www.thebalancemoney.com/how-much-rent-can-i-afford-on-minimum-wage-4175100.

[15] R. ADLER, *What exactly is a small apartment?* Accessed 2024-09-02. [Online]. Available: https://www.housedigest.com/993166/what-exactly-is-a-small-apartment/.

[16] E. Reynolds, *Home of the week: This $22 million estate is the priciest listing in san francisco's glen park neighborhood*, Accessed 2024-16-02. [Online]. Available: https://robbreport.com/shelter/homes-for-sale/22-million-estate-most-expensive-san-francisco-glen-park-1235366522/.

[17] D. Lelazier, *A highlight of the best san francisco neighborhoods to live: Glen park*, Accessed 2024-16-02. [Online]. Available: https://daniellelazier.com/a-highlight-of-the-best-san-francisco-neighborhoods-to-live-glen-park/.

[18] FIFA, *Stadium guidelines - 5.3 pitch dimensions and surrounding areas*, Accessed 2024-09-02. [Online]. Available: https://publications.fifa.com/en/football-stadiums-guidelines/technical-guideline/stadium-guidelines/pitch-dimensions-and-surrounding-areas/.

[19] D. Cielen, A. D. B. Meysman and M. Ali, *Introducing Data Science Big Data Machine Learning and More Using Python Tools*. Mannung, 2016.

[20] J. Brownlee, *What is the difference between test and validation datasets?* Accessed 2024-15-02. [Online]. Available: https://machinelearningmastery.com/difference-test-validation-datasets/.

[21] D. Shulga, *5 reasons why you should use cross-validation in your data science projects*, Accessed 2024-12-02. [Online]. Available: https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79.

[22] J. Brownlee, *How to compare the performance of machine learning algorithms in weka*, Accessed 2024-13-02. [Online]. Available: https://machinelearningmastery.com/compare-performance-machine-learning-algorithms-weka/.

[23] J. Brownlee, *How to transform your machine learning data in weka*, Accessed 2024-16-02. [Online]. Available: https://machinelearningmastery.com/transform-machine-learning-data-weka/.

[24] A. Raj, *Unlocking the true power of support vector regression - using support vector machine for regression problems*, Accessed 2024-16-02. [Online]. Available: https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0.

[25] L. Breiman, 'Random forests,' *Machine Learning*, vol. 45, pp. 5–32, 2001. [Online]. Available: https://api.semanticscholar.org/CorpusID:89141.

[26] J. Platt, 'Sequential minimal optimization: A fast algorithm for training support vector machines,' *Advances in Kernel Methods-Support Vector Learning*, vol. 208, Jul. 1998.

[27] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kauffmann Publishers, 1993.

[28] K. Deeba and B. Amutha, 'Classification algorithms of data mining,' *Indian Journal of Science and Technology*, vol. 9, no. 39, 2016.

[29] Weka Sourceforge, *Class nominaltobinary*, Accessed 2024-17-02. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/filters/supervised/attribute/NominalToBinary.html.

[30] S. Shevade, S. Keerthi, C. Bhattacharyya and K. Murthy, 'Improvements to the smo algorithm for svm regression,' *IEEE Transactions on Neural Networks*, vol. 11, no. 5, pp. 1188–1193, 2000. DOI: 10.1109/72.870050.

[31] L. Chen-Xu and Y. Gui-Lan, 'Chapter 3 - neural networks in phononics,' in *Intelligent Nanotechnology*, ser. Materials Today, Y. Zheng and Z. Wu, Eds., Elsevier, 2023, pp. 47–70, ISBN: 978-0-323-85796-3. DOI: https://doi.org/10.1016/B978-0-323-85796-3.00003-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323857963000032.

[32] W. Aha, D. Kibler and M. Albert, 'Instance-based learning algorithms,' *Machine Learning*, vol. 6, pp. 37–66, Jan. 1991. DOI: 10.1023/A:1022689900470.

[33] Weka Sourceforge, *Class reptree*, Accessed 2024-16-02. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/REPTree.html.

[34] Weka Sourceforge, *Class removewithvalues*, Accessed 2024-16-02. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/filters/unsupervised/instance/RemoveWithValues.html.

[35] Weka Sourceforge, *Class em*, Accessed 2024-16-02. [Online]. Available: https://weka.sourceforge.io/doc.dev/weka/clusterers/EM.html.

[36] S. Kotsiantis, D. Kanellopoulos, P. Pintelas *et al.*, 'Handling imbalanced datasets: A review,' *GESTS international transactions on computer science and engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[37] S. Galli, *Exploring oversampling techniques for imbalanced datasets*, Accessed 2024-01-03. [Online]. Available: https://www.blog.trainindata.com/oversampling-techniques-for-imbalanced-data/.

[38] IBM, *What is overfitting?* Accessed 2024-16-02. [Online]. Available: https://www.ibm.com/topics/overfitting.

[39] B. Ratner, 'The correlation coefficient: Its values range between $+1/-1$, or do they?' *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 17, no. 2, 2009.

[40] S. Allwright, *What is a good mae score? (simply explained)*, Accessed 2024-14-02. [Online]. Available: https://stephenallwright.com/good-mae-score/.

[41] M. Taylor, *Mean absolute percentage error*, Accessed 2024-17-02. [Online]. Available: https://www.vexpower.com/brief/mean-absolute-percentage-error.

[42] W. Lemahieu, S. vanden Broucke and B. Baesens, *Principles of Database Management: The Practical Guide to Storing, Managing and Analyzing Big and Small Data*. Cambridge University Press, 2018.

[43] OpenClassrooms.com, *Model a database with uml*, Accessed 2024-01-03. [Online]. Available: https://openclassrooms.com/en/courses/7569661-model-a-database-with-uml/7785424-link-tables-using-foreign-keys.

[44] A.-A. Corodescu, N. Nikolov, A. Q. Khan, A. Soylu, M. Matskin, A. H. Payberah and D. Roman, 'Big data workflows: Locality-aware orchestration using software containers,' *Sensors*, vol. 21, no. 24, 2021, ISSN: 1424-8220. DOI: 10.3390/s21248212. [Online]. Available: https://www.mdpi.com/1424-8220/21/24/8212.

[45] H. Lee and T. Park, 'Allocating data and workload among multiple servers in a local area network,' *Information Systems*, vol. 20, no. 3, pp. 261–269, 1995, ISSN: 0306-4379. DOI: https://doi.org/10.1016/0306-4379(95)00012-S. [Online]. Available: https://www.sciencedirect.com/science/article/pii/030643799500012S.

[46] A. S. Gillis, *Message passing interface (mpi)*, Accessed 2024-26-02. [Online]. Available: https://www.techtarget.com/searchenterprisedesktop/definition/message-passing-interface-MPI.

[47] Broadcom, *Vmware esxi*, Accessed 2024-25-02. [Online]. Available: https://www.vmware.com/products/esxi-and-esx.html.

[48] N. Sultan, 'Making use of cloud computing for healthcare provision: Opportunities and challenges,' *International Journal of Information Management*, vol. 34, no. 2, pp. 177–184, 2014, ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2013.12.011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0268401213001680.

[49] V. Korkhov, I. Gankevich, A. Gavrikov, M. Mingazova, I. Petriakov, D. Tereshchenko, A. Shatalin and V. Slobodskoy, 'Finding bottlenecks in message passing interface pro-

grams by scalable critical path analysis,' *Algorithms*, vol. 16, no. 11, 2023, ISSN: 1999-4893. DOI: 10.3390/a16110505. [Online]. Available: https://www.mdpi.com/1999-4893/16/11/505.

[50] S. Thota, 'Vmware virtualization-physical to virtual migration,' *Int. J. Comput. Trends Technol*, vol. 58, pp. 65–75, 2018.

[51] Microsoft, *What is azure?* Accessed 2024-25-02. [Online]. Available: https://azure.microsoft.com/en-gb/resources/cloud-computing-dictionary/what-is-azure/.

[52] Microsoft, *Get started for azure it operators*, Accessed 2024-25-02. [Online]. Available: https://learn.microsoft.com/en-us/azure/guides/operations/azure-operations-guide.

[53] A. Khan, X. Yan, S. Tao and N. Anerousis, 'Workload characterization and prediction in the cloud: A multiple time series approach,' in *2012 IEEE Network Operations and Management Symposium*, 2012, pp. 1287–1294. DOI: 10.1109/NOMS.2012.6212065.

[54] Microsoft, *Azure virtual machine scale sets*, Accessed 2024-26-02. [Online]. Available: https://azure.microsoft.com/en-gb/products/virtual-machine-scale-sets/.

[55] C. Yang, S. Lan, L. Wang, W. Shen and G. G. Huang, 'Big data driven edge-cloud collaboration architecture for cloud manufacturing: A software defined perspective,' *IEEE access*, vol. 8, pp. 45 938–45 950, 2020.

[56] Microsoft, *Azure stack edge*, Accessed 2024-26-02. [Online]. Available: https://azure.microsoft.com/en-gb/products/azure-stack/edge/.

[57] H. Zhonglin and H. Yuhua, 'A study on cloud backup technology and its development,' in *Innovative Computing and Information*, M. Dai, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–7, ISBN: 978-3-642-23993-9.

[58] A. Cavoukian, 'Privacy by design: The definitive workshop. a foreword by ann cavoukian, ph.d,' *Identity in the Information Society*, vol. 3, no. 2, 2010.

[59] I. Rubinstein and N. Good, 'Privacy by design: A counterfactual analysis of google and facebook privacy incidents,' *Berkeley Technology Law Journal*, vol. 28, no. 12-43, 2013.

[60] I. Kroener and D. Wright, 'A strategy for operationalizing privacy by design,' *The Information Society*, vol. 30, no. 5, pp. 355–365, 2014. DOI: 10.1080/01972243.2014.944730. eprint: https://doi.org/10.1080/01972243.2014.944730. [Online]. Available: https://doi.org/10.1080/01972243.2014.944730.

[61] C. Kurtz, M. Semmann and T. Böhmann, 'Privacy by design to comply with gdpr: A review on third-party data processors,' 2018.

[62] T. Linden, H. Harkous and K. Fawaz, 'The privacy policy landscape after the gdpr,' *Proceedings on Privacy Enhancing Technologies*, vol. 2020, pp. 47–64, 2018. DOI: 10.2478/popets-2020-0004.

[63] Information Commissioner's Office, *A guide to the data protection principles*, Accessed 2024-27-02. [Online]. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/.

[64] R. Page, *The state of privacy regulations across asia*, Accessed 2024-27-02. [Online]. Available: https://www.csoonline.com/article/572461/the-state-of-privacy-regulations-across-asia.html.

[65] L. de la Torre, *Gdpr matchup: The california consumer privacy act 2018*, Accessed 2024-27-02. [Online]. Available: https://iapp.org/news/a/gdpr-matchup-california-consumer-privacy-act/.

[66] State of California Department of Justice, *California consumer privacy act (ccpa)*, Accessed 2024-27-02. [Online]. Available: https://oag.ca.gov/privacy/ccpa.

[67] California Legislative Information, *1.81.5. california consumer privacy act of 2018*, Accessed 2024-27-02. [Online]. Available: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5.

[68] K. B. Park and M. Kang, *South korea - data protection overview*, Accessed 2024-27-02. [Online]. Available: https://www.dataguidance.com/notes/south-korea-data-protection-overview.

[69] H. Ko, J. Leitner, E. Kim and J. Jeong, 'Structure and enforcement of data privacy law in South Korea,' *International Data Privacy Law*, vol. 7, no. 2, pp. 100–114, Apr. 2017, ISSN: 2044-3994. DOI: 10.1093/idpl/ipx004. eprint: https://academic.oup.com/idpl/article-pdf/7/2/100/17932194/ipx004.pdf. [Online]. Available: https://doi.org/10.1093/idpl/ipx004.

[70] International Organization for Standardization, 'Iso/iec 27021:2017 information technology security techniques competence requirements for information security management systems professionals,' Standard, 2017.

[71] J. Williams, *Owasp risk rating methodology*, Accessed 2024-27-02. [Online]. Available: https://owasp.org/www-community/OWASP_Risk_Rating_Methodology.

[72] National Cyber Security Centre, *Authentication methods: Choosing the right type*, Accessed 2024-27-02. [Online]. Available: https://www.ncsc.gov.uk/guidance/authentication-methods-choosing-the-right-type.

[73] J. A. Obar and A. Oeldorf-Hirsch, 'The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services,' *Information, Communication & Society*, vol. 23, no. 1, pp. 128–147, 2020. DOI: 10.1080/1369118X.2018.1486870. eprint: https://doi.org/10.1080/1369118X.2018.148687. [Online]. Available: https://doi.org/10.1080/1369118X.2018.1486870.

[74] N. Steinfeld, '"i agree to the terms and conditions": (how) do users read privacy policies online? an eye-tracking experiment,' *Computers in Human Behavior*, vol. 55, pp. 992–1000, 2016, ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2015.09.038. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563215301692.

[75] Information Commissioner's Office, *Consent*, Accessed 2024-27-02. [Online]. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/a-guide-to-lawful-basis/lawful-basis-for-processing/consent/.