Understanding and Personalising Smart City Services Using Machine Learning, the Internet-of-Things and Big Data

Jeannette Chin
Department of Computing and Technology
Anglia Ruskin University
Cambridge, United Kingdom

Vic Callaghan
School of Computer Science and Electronic Engineering
Essex University
Colchester, United Kingdom

Ivan Lam
University College London
London, United Kingdom

***Abstract*—This paper explores the potential of Machine Learning (ML) and Artificial Intelligence (AI) to lever Internet of Things (IoT) and Big Data in the development of personalised services in Smart Cities. We do this by studying the performance of four well-known ML classification algorithms (Bayes Network (BN), Naïve Bayesian (NB), J48, and Nearest Neighbour (NN)) in correlating the effects of weather data (especially rainfall and temperature) on short journeys made by cyclists in London. The performance of the algorithms was assessed in terms of accuracy, trustworthy and speed. The data sets were provided by Transport for London (TfL) and the UK MetOffice. We employed a random sample of some 1,800,000 instances, comprising six individual datasets, which we analysed on the WEKA platform. The results revealed that there were a high degree of correlations between weather-based attributes and the Big Data being analysed. Notable observations were that, on average, the decision tree J48 algorithm performed best in terms of accuracy while the kNN IBK algorithm was the fastest to build models. Finally we suggest IoT Smart City applications that may benefit from our work**

Keywords—classification; personalisation; machine learning; artificial intelligence; profiling; data mining, recommendation systems; algorithms; Internet-of-Things; Smart Cities; Big Data, Data Analytics

I. INTRODUCTION

Technologies such as the Internet-of-Things, Big Data, and Data Analytics are changing the way we live, work, and play, creating many new opportunities. One such opportunity is the so-called Smart City which proponents claim may contain thousands of sensors generating massive amounts of data, Big Data, which through analysis, could enable city services to be more responsive to the needs of the inhabitants. In this way businesses and other organisations would be able to offer their clients more personalized services, which better fit their needs. To achieve a successful personalised service, two fundamental requirements are needed. The first is the ability to understand the behaviour of the users and the second is the ability to adapt efficiently, to the user's changing behaviour over time. Artificial Intelligence (AI) in various guises is commonly used in applications to understand and adapt user behaviours, but while these systems tend to work well in specific domains, the search for General Artificial Intelligence (GAI) is proving illusive [3]. For Big-Data, where information about individuals can be aggregated and reasoned about, profiling through machine learning plays a crucial part in the provision of personalized services. There are three methods in machine learning for profiling, namely content-based [1] [18], collaborative methods [19] [20], and hybrid between these two [21], [22],[23]

By way of an exemplary study, this paper investigated the relationships between weather and short cycling journeys. The motivation behind the investigation was to gain a better understanding of the potential for big-data, of a type that in future might be gathered from city based IoT systems, and could be utilised to produce meaningful information for Smart City applications. The study was based on London Santander Bikes[1] hire usage from datasets provided by the Department of Transport UK, together with two different weather related datasets provided by the British Atmospheric Data Centre (BADC)[2] from which the rainfall and temperature records employed in this study were extracted. Licenses for the bike datasets were obtained under an open government license[3] while those for the weather datasets were provided for research purposes only. There were no privacy issues as the data did not include personal information. The datasets were tested against four Machine Learning Classification algorithms (see section III) and results are shown in section V. The study was conducted on WEKA[4] platform.

The paper is structured as follows: Section II, background work on profiling techniques and personalisation; Section III, discussions on classification techniques and algorithms; Section IV, detailed descriptions on the datasets including pre-processing strategies; Section V, data analysis and evaluation; and, finally, Section VI, the summary and conclusions.

II. RELATED WORK

Profiling can be regarded as, the construction and application of descriptions that characterise users (a user profile). Many of the parameters involved are in continual temporal flux (eg changes in environment, age, health etc) posing challenges to its computerisation, in terms of

[1] Santander Cycles - https://tfl.gov.uk/modes/cycling/santander-cycles
[2] BADC - http://badc.nerc.ac.uk/home/index.html
[3] Open license: http://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/
[4] WEKA - http://www.cs.waikato.ac.nz/ml/weka/

978-1-5090-1412-5/17/$31.00 ©2017 IEEE

maintaining such information (eg classifying, learning behaviours, updating information etc). Various classification algorithms have been explored, the majority of which has centred on harvesting, processing and visualizing personal information, either via explicit (user input) or implicit (device tracking). Applications that harness personal information to provide useful recommendations enable organisations and clients to enjoy more personalised services, improving efficiency which saves money and improves relationships. In 2004, a comprehensive review of various algorithms for such systems was conducted by Schafer et al who proposed a system based on usage patterns and feature weighting [3]. Kim et-al [4] explored collaborative filtering algorithms for provision of personalized TV programmes. Cufoglu et-al [5] conducted user profiling studies by comparing various classification algorithms, reporting that, although it takes longer, the Naïve Bayesian Tree (NBTree) should be considered for applications where accuracy is important. Web page filtering has also been an important area for profiling research [6] [7] [8]. Finally, the WEKA platform, developed by the University of Waikato in New Zealand, is powerful Machine learning (ML) tool for processing large collections of datasets. Various studies on different datasets [9] [10] [11] have been conducted using this tool which is provided as open source software, under a General Public License.

III. METHODOLOGY – ALGORITHMS AND CLASSIFIERS

As mentioned earlier, rainfall and temperature are the only two weather attributes considered in this study. Classification can be regarded as a process of learning a certain model (ie relationships) from a given dataset such that the model can be used to predict the classification of a novel instance whose classification is unknown [13]. This technique is commonly used in ML and we use this to investigate the relationships between weather and cycling through possible predictions. For class performance and accuracy examinations, we randomly sampled 300,000 instances for each month studied, out of the total number of usage instances of that month (Table II). The total number of usage instances for quarter 1 and 3 are shown in Table I. Each instance is represented by 12 attributes; and each attribute has its own set of associated values (Table III). The datasets were then tested and trained on four different classifier algorithms, namely, Bayes Network (BN), Naïve Bayesian (NB), J48, and Nearest Neighbour (NN) algorithms on WEKA ML platform.

A. Naïve Bayesian Classifier

Bayesian classifiers utilize a statistical learning algorithm to compute the probability of an event occurring, given particular attributes. They are based on the Bayes conditional probability rule, which states how the probability of an event occurring may be calculated given the probabilities of related events or attributes.

The Naïve Bayes classifier is a simplified form of a Bayesian classifier, where each attribute is considered independently of each other. Based on Bayes [14] theorem, which describes a way of calculating the posterior probability:

Assume;

- Ll_a as class label where $Ll = \{Ll_1, Ll_2, \dots, Ll_a\}$ and $a = 1,2, \dots, A$.
- Y_j as unclassified test instance where $Y_j = \{y_j(1), y_j(2), \dots, y_j(B)\}$ for $b = 1,2, \dots, B$.

Then, Y_j will be classified into class Ll_a with the maximum posterior class probability,

$$P(Ll_a|Y_j) = \arg\max_{Ll_a} P(Ll_a)P(Y_j|Ll_a) \quad (1)$$

While Naive Bayes [15], a linear classifier[5], assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors, which alters (1) to:

$$\mathrm{P}(\mathrm{Ll_a}|\mathrm{Y_j}) = \arg\max_{\mathrm{Ll_a}} \mathrm{P}(\mathrm{Ll_a})\mathrm{P}(\mathrm{y_j}(1), \mathrm{y_j}(2), \dots, \mathrm{y_j}(\mathrm{B})|\mathrm{Ll_a}) \quad (2)$$

$$\mathrm{P}(\mathrm{Ll_a}|\mathrm{Y_j}) = \arg\max_{\mathrm{Ll_a}} \mathrm{P}(\mathrm{Ll_a}) \prod_{\mathrm{t=1}}^{\mathrm{B}} \mathrm{P}(\mathrm{y_j}(\mathrm{t})|\mathrm{Ll_a}) \quad (3)$$

B. J48 Tree Classifier

J48 is a decision tree classifier implemented in the WEKA data mining software, based on the C4.5 algorithm developed by Ross Quinlan, which itself is an extension of ID3, an earlier algorithm also developed by Quinlan [17][16].

The J48 classifier creates a decision tree based on attributes from a training data set. The attribute that creates the split with the lowest entropy (and therefore highest information gain) is used as the root node. This procedure is applied recursively to create further nodes. Leaf nodes represent the final classification target of the data. Decision trees created by J48 may, optionally, be pruned using subtree replacement or subtree raising methods.

The entropy, *E*, of a node with two branches is given by

$$E = -pP \log_2(pP) - pN \log_2(pN) \quad (4)$$

where pP is the proportion of positive training examples and pN is the proportion of negative training examples [17]

For a node with *n* branches, the entropy is given by

$$E = \sum_{i=1}^{n} -p_i \log_2(p_i) \quad (5)$$

C. Nearest Neighbour Classifier

As the name suggests, the Nearest Neighbour (NN) classifiers use neighbour-based learning methods for supervised and unsupervised learning. The technique, sometime referred as instance-based learning, involves normalized Euclidean distance to compare each test instance with the training instances. The closest training instance predicted to have the same class label with the test instance [8]. In the case of more than one training instance qualifing as the closest, the class label of the first one is assigned to be the class label of the test instance [2]. The k-Nearest Neighbor (kNN) Classifier provides more users' controls in its metric measure; generally it is known as non-generalising ML methods useful

[5] Linear Classifiers: http://nlp.stanford.edu/IR-book/html/htmledition/linear-versus-nonlinear-classifiers-1.html

for classification problems with certain conditions. The Instance-Based Learning with parameter k (IBK) classifier is a comprehensive form of kNN, which has been used in this study to compare with other algorithms. In IBK the comparison between the test instance and the training instance is done as follows:

Assume training instance $X_i = \{x_i(1), x_i(2), \dots, x_i(B)\}$ and test instance $Y_j = \{y_j(1), y_j(2), \dots, y_j(B)\}$. Here comparison between training instance and test instance is done feature by feature as:

- If the feature is numeric,

$$g\left(y_j(k), x_i(k)\right) = (y_j(k) - x_i(k))^2 \quad (6)$$

- If the feature is symbolic,

$$g(y_j(k) - x_i(k)) = \begin{cases} 0, if\ y_j(k) = x_i(k) \\ 1, if\ y_j(k) \neq x_i(k) \end{cases} \quad (7)$$

Where the $g\left(y_j(k), x_i(k)\right)$ function shows the similarity between k values of the training and test instances and the distance is calculated as;

$$dist(Y_j, X_i) = \sqrt{\sum_{k=1}^{B} g(y_j(k), x_i(k))} \quad (8)$$

IV. DATASETS

The study is based on combining two different datasets. The first is derived from an easy-access self-service model known as the Santander Cycles scheme (formerly called Barclays Cycle Hire), owned by Transport for London (TfL) provides over 10,000 bikes and 700 docking stations (Fig 1), covering every 300 to 500 metres in the London city centre[6]. The study was conducted using TfL Bikes hire usage dataset in 2012 (Table 1). No TfL bike hire data before 4th January 2012 was publically available. This dataset was obtained under open government license. We combine this dataset with two weather related datasets by the Met Office CEDA[7], more specifically, the rainfall and temperature weather datasets provided by The British Atmospheric Data Centre (BADC)[8].

Fig. 1. The London Santander Bike hire docking stations.

The weather related datasets are obtained under the license for research purposes only.

[6] Santander Cycles: https://tfl.gov.uk/modes/cycling/santander-cycles/find-a-docking-station
[7] CEDA - http://www.ceda.ac.uk/blog/category/met-office-data/
[8] BADC - http://badc.nerc.ac.uk

A. London Santander Bikes hire usage datasets

Due to the size of the datasets and computation limitations, for this study we decided to sample the datasets according to the classification of quarters of the year: each quarter is a three-month period, similar to those on a financial calendar. Given weather is the focus of the study, we chose winter and summer quarters as they reflected the most extreme (opposite) climates. The datasets have: Q1 – January, February and March and Q3 – July, August and September, hence six datasets in total. Note, for 2012, Q1 begins with 04 January to 31 March and 29 Feb is excluded. Table II shows the total monthly usage of the datasets sampled. Data validation included the following rule: all data must meet the following two criteria: (1) start date and time must not fall outside the quarters being sampled (2) each entry must has a match with both the weather datasets. Entries with missing data or data with NULL values are excluded in the analysis.

TABLE I. TFL BIKES HIRE STATS

Year	Number of Bikes Hire	
	Q1	*Q3*
2012	1,794,360	3,192,490

TABLE II. SAMPLES TFL BIKES HIRE MONTHLY USAGE

Year	Break-down number of Bikes Hire					
	Jan	Feb	Mar	Jul	Aug	Sep
2012	453,297[a]	451776[b]	767,700	881,899	1,040,989	958,143

[a] From 04-31. [b] 29 Feb excluded

B. Met Office CEDA rainfall datasets

The British Atmospheric Data Centre (BADC) provides various weather related datasets in different regions globally. Datasets related to daily rainfall (the UK Daily Rainfall data) and temperature readings (the UK Hourly Weather Observation data) in London are used in the study. The rainfall dataset has the following classifications: 0mm as 'DRY', less than 0.4mm as "Drizzle", between 0.4mm to 4mm as "Moderate rain", above 4mm as "Heavy rain". The classification coding used is according to the ones defined by the MetOffice[9]. It has observed that there are: one reading made per day in daily rainfall dataset and two readings made per day in temperature dataset, with, first reading was made at 09:00 and the second at 21:00. The readings were made at different weather stations across the city of London. The temperature dataset has the following classifications: 0C or less as "<0C", less than 5C as "0-5C", between 5C and less than10C as ">=5-<10C", between 10C and less than 15C as ">=10-<15C", between 15C and less than 20C as ">=15C-<20C", between 20C and less than 25C as ">=20C-<25C", above 25C as ">=25C".

C. Datasets pre-processing and Attributes

The pre-processing procedures were based on the assumption that the weather related readings were applicable to all TfL bike routes in the city used in the study. For rainfall,

[9] Rainfall Classifications - http://www.metoffice.gov.uk/media/pdf/f/c/Fact_sheet_No._3.pdf

given there is only one reading per day, the pre-processing procedure included finding a match to dates between the two datasets.

TABLE III. ATTRIBUTES USED IN THE STUDY

Attributes	Distinct	Unique
journey_name	97862	43160 (14%)
start	37397	4802 (2%)
day	31	0
wday	7	0
hour	24	0
duration	10144	3976 (1%)
sstation_id	569	0
sstation	569	0
temp_max	6	0
temp_max_raw	107	0
rain	4	0
rain_raw	25	0

For temperature however, it is worth noting that the readings made at 21:00 were mostly higher than those made at 09:00, thus the pro-processing procedures included splitting a day into two 12-hour periods and associating the day-journeys to 21:00 reading and evening-journeys to 09:00 reading. We created journeys information that contain details of start-station-ids and end-station-ids as an independent attribute. The study considered 12 attributes in total, each attribute has its unique set of values, which shows in Table III.

V. COMPARISON ANALYSIS AND RESULTS

A random sample of 300,000 instances from each dataset (Table II), i.e. a total of 1,800,000, was selected for the analysis. The datasets were tested and analyzed on the WEKA platform. A 10-fold cross-validation was used for each test and this was used as the basis of the full training set. The tests were conducted with a focus on comparisons between four classification algorithms, specifically for four main weather related attributes: rain_raw (rainfall raw reading), rain (rainfall classification), temp_max_raw (maximum forecast raw temperature reading), and temp_max (maximum forecast temperature reading), with the following objectives:

- If there is any correlation between these attributes, and with the other attributes within the same dataset
- The performance of four popular classification algorithms with respect to the classification of cycling journeys into the right classes.

The distributions of rainfall (fig 2 and fig 3) for Q1 and Q3 appeared to suggest Q1 had dry months compared to Q3 with July being the wettest month sampled. On average, the total number of hires appeared to be lower for wet days. The total number of hires in Q3 was about 75% more than in Q1 suggests a strong correlation between warmer weather and short-cycling journeys.

A. Rainfall Comparison Analysis

The analysis was conducted for both winter and summer quarters (Q1 & Q2) over six months period. The first comparison used rain_raw and rain attributes as the class attribute. The rain_raw attribute has 25 distinct values while the rain attribute has 4 distinct values, and both attributes have zero unique case. The results show in Table IV and Table VI review that all four algorithms performed remarkably well for correctly classified instances into the appropriate classes. All but IBK correctly classified all instances within the datasets, which is 100%. This indicates that there are high correlations between this attribute and the instances within the datasets. For IBK, based on k=1, it was observed that IBK performed better with the RAIN attribute.

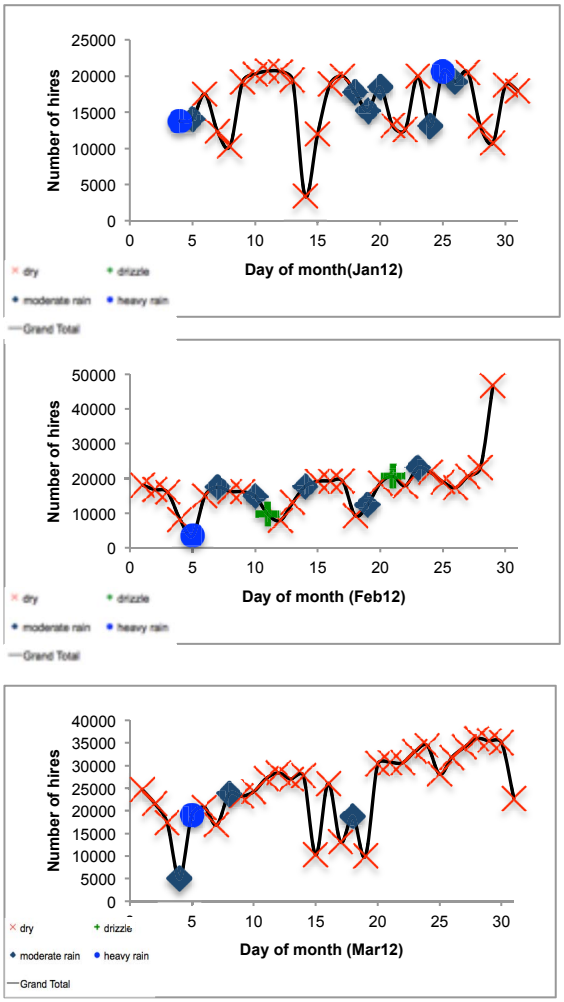

Fig 2 Q1 rainfall distributions

The results indicated that IBK's performance was linked to the number of distinct cases within the datasets. In this study, the RAIN_RAW attribute has 4 times the number of such cases compared to RAIN attribute. In terms of speed performance, such as building the models and training the datasets, each algorithm performed differently (Table V and Table VII). Here, NB and IBK out-performed the other two algorithms for building the models. Interestingly, the results also revealed that the performance for both BN and J48 were linked to the number of distinct cases within the datasets. However, from observation, IBK performed worst in terms of the time taken to train the datasets compared to BN, NB and J48, and followed by NB.

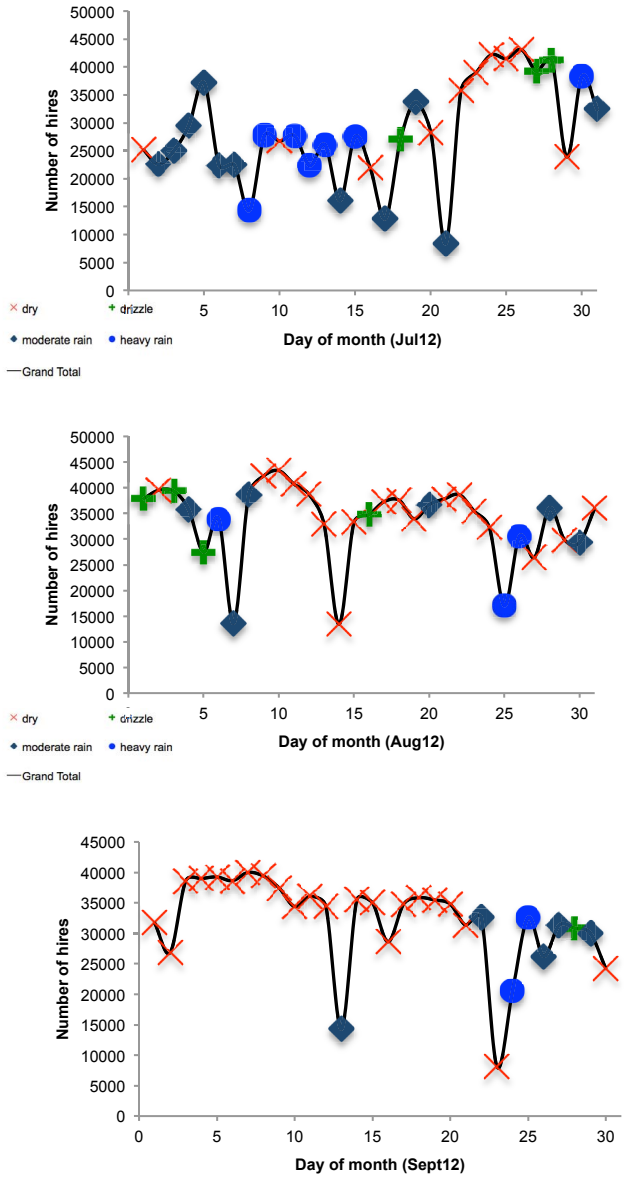

Fig 3. Q3 rainfall distributions

B. *Temperature Comparison Analysis*

Same datasets were subsequently used to evaluate temperature attributes.

TABLE IV. RAIN_RAW - CORRECTLY CLASSIFIED INSTANCES IN PERCENTAGE

datasets	BN	NB	J48	IBK
Jan-12	100	100	100	99.999
Feb-12	100	100	100	99.997
Mar-12	100	100	100	99.999
Jul-12	100	100	100	99.997
Aug-12	100	100	100	100
Sep-12	100	100	100	100

TABLE V. RAIN_RAW - TIME TAKEN TO BUILD MODEL IN SECONDS

datasets	BN	NB	J48	IBK
Jan12	0.45	0.12	0.68	0.03
Feb12	0.25	0.08	0.34	0.04
Mar12	0.28	0.07	0.31	0.04
Jul12	0.69	0.09	0.71	0.09
Aug12	0.32	0.09	0.39	0.12
Sep12	0.29	0.06	0.34	0.09

TABLE VI. RAIN - CORRECTLY CLASSIFIED INSTANCES IN PERCENTAGE

datasets	BN	NB	J48	IBK
Jan-12	100	100	100	100
Feb-12	100	100	100	99.9997
Mar-12	100	100	100	100
Jul-12	100	100	100	100
Aug-12	100	100	100	100
Sep-12	100	100	100	100

TABLE VII. RAIN - TIME TAKEN TO BUILD MODEL IN SECONDS

datasets	BN	NB	J48	IBK
Jan-12	0.18	0.06	0.19	0.03
Feb-12	0.22	0.06	0.27	0.04
Mar-12	0.23	0.06	0.28	0.04
Jul-12	0.24	0.07	0.28	0.11
Aug-12	0.32	0.07	0.24	0.11
Sep-12	0.22	0.08	0.24	0.09

Here the TEMP-MAX attribute represented the temperature classification described in section IV. The raw reading is represented by attribute TEMP-MAX-RAW. The results shown in Table VIII and IX revealed that, although there high accuracy was achieved for all four algorithms, J48 was the winner achieving 100% for evaluating all six datasets. However, it had a lowest performance in terms of the time taken to build the models, followed by BN,

TABLE VIII. TEMP-MAX_RAW - CORRECTLY CLASSIFIED INSTANCES IN PERCENTAGE

datasets	BN	NB	J48	IBK
Jan-12	98.9723	98.7653	100	95.462
Feb-12	99.9273	99.9123	100	96.727
Mar-12	99.939	99.923	100	98.088
Jul-12	99.307	99.187	100	96.5433
Aug-12	99.9177	99.8567	100	97.4433
Sep-12	99.813	99.7623	100	97.5537

TABLE IX. TEMP-MAX_RAW - TIME TAKEN TO BUILD MODEL IN SECONDS

datasets	BN	NB	J48	IBK
Jan-12	0.47	1.16	1.2	0.05
Feb-12	0.69	0.13	1.23	0.04
Mar-12	0.61	0.12	1.22	0.04
Jul-12	0.76	0.11	1.44	0.11
Aug-12	0.71	0.13	1.32	0.09
Sep-12	0.7	0.12	1.53	0.1

indicating the association of number of distinct cases within the datasets. Again, it was observed that IBK required longest time for data training, followed by NB.

VI. Conclusion

This paper explored the potential for AI to lever IoT and Big Data to support the realisation of personalised services in Smart Cities through the use of ML techniques to correlate weather conditions with short-cycling journeys made in London. The rationale was that such sensor generated data might provide valuable insights in to algorithms that Smart City ML and IoT system may use. The ability to understand the behaviour of the users, for given circumstances, and to be able to adapt services to better fit their needs efficiently is fundamental to a successful personalised service. In this work the study sought to understand the correlation between weather-based conditions and short-cycling behaviour, using using four well-known ML classification algorithms operating on data taken from six datasets (of 1,800,000 instances).All four classification algorithms were consistent and produced high accuracy results in classifying instances into the right class with J48 achieving 100% for all cases. A one (or close) Kappa statistic was obtained for all tests, indicating there was a high level of confidence in the results obtained. With respect to the speed of building models, IBK outperformed the others, followed by NB. However, J48 took the longest time to build the tree models and IBK required the most timefor data training (more than 6 hours in some cases -attributed to the "cold start" problem), followed by NB, BN and J48. Based on the results it was concluded that the kNN algorithms was not suitable for small computations, such as IoT applications, due to the time required to train the algorithm and attributed to "cold start" constraint whereas decision tree algorithms were well suited for applications where accuracy was important. Concerning the data, the results revealed that there was a strong correlation between weather attributes (rainfall and temperature) within each dataset. Overall, the results indicate that a combination of ML, IoT and Big Data offer great potential to developers of smart city technologies and services.

References

[1] G. Araniti, Pasquale De Meo, A. Iera and D. Ursino, "Adaptively controlling the QoS of multimedia wireless applications through "user profiling" techniques," in IEEE Journal on Selected Areas in Communications, vol. 21, no. 10, pp. 1546-1556, Dec. 2003. doi: 10.1109/JSAC.2003.815226

[2] M. J. Martin-Bautista, D. H. Kraft, M. A. Vila, J. Chen and J. Cruz, User profiles and fuzzy logic for web retrieval issues, Soft Computing (Focus), 15(3-4), pp. 365-372. 2002

[3] V. Callaghan, J. Miller, R. Yampolskiy, S. Armstrong (eds) "The Technological Singularity: Managing the Journey", Springer-Verlag GmbH Germany 2017, ISBN 978-3-662-54031-2

[4] E. Kim, S. Pyo, E. Park and M. Kim, "An Automatic Recommendation Scheme of TV Program Contents for (IP)TV Personalization," in IEEE Transactions on Broadcasting, vol. 57, no. 3, pp. 674-684, Sept. 2011. doi: 10.1109/TBC.2011.2161409

[5] A. Cufoglu, M. Lohi and K. Madani, "A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling," Computer Science and Information Engineering, 2009 WRI World Congress on, Los Angeles, CA, 2009, pp. 708-712. doi: 10.1109/CSIE.2009.954

[6] D. Godoy and A. Amandi, "User profiling for Web page filtering," in IEEE Internet Computing, vol. 9, no. 4, pp. 56-64, July-Aug. 2005. doi: 10.1109/MIC.2005.90

[7] J. Zhang and M. Shukla, "Rule-Based Platform for Web User Profiling," Sixth International Conference on Data Mining (ICDM'06), Hong Kong, 2006, pp. 1183-1187. doi: 10.1109/ICDM.2006.137

[8] J. A. Iglesias, P. Angelov, A. Ledezma and A. Sanchis, "Creating Evolving User Behavior Profiles Automatically," in IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 854-867, May 2012. doi: 10.1109/TKDE.2011.17

[9] R. Duriqi, V. Raca and B. Cico, "Comparative analysis of classification algorithms on three different datasets using WEKA," 2016 5th Mediterranean Conference on Embedded Computing (MECO), Bar, 2016, pp. 335-338. doi: 10.1109/MECO.2016.7525775

[10] D. Heredia, Y. Amaya and E. Barrientos, "Student Dropout Predictive Model Using Data Mining Techniques," in IEEE Latin America Transactions, vol. 13, no. 9, pp. 3127-3134, Sept. 2015. doi: 10.1109/TLA.2015.7350068

[11] L. Jiang, H. Zhang and Z. Cai, "A Novel Bayes Model: Hidden Naive Bayes," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 10, pp. 1361-1371, Oct. 2009. doi: 10.1109/TKDE.2008.234

[12] Shaheen, S., Guzman, S., Zhang, H., 2010. Bikesharing in Europe, the Americas, and Asia: past, present, and future. In: 2010 Transportation Research Board Annual Meeting. Washington, DC, USA.

[13] Senthil K. Palanisamy, Association Rule Based Classification, Thesis of Degree of Master of Science, Worcester Polytechnic Institute, 2006

[14] Ian H. Witten, Eibe Frank, Data mining : practical machine learning tools and techniques, ISBN: 0-12-088407-0, 2000

[15] S. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Informatica*, vol. 31, pp. 249-268, 2007.

[16] R. Arora and S. Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA", *International Journal of Computer Applications*, vol. 54, no. 13, pp. 21-25, 2012.

[17] A. Yadav and S. Chandel, "Solar energy potential assessment of western Himalayan Indian state of Himachal Pradesh using J48 algorithm of WEKA in ANN based prediction model", Renewable Energy, vol. 75, pp. 675-693, 2015.

[18] Balabanović, Marko, and Y.S., 1997. Fab: content-based, collaborative recommendation. Communications of the ACM, 40(3), pp.66–72.

[19] Bobadilla, J. et al., 2011. A framework for collaborative filtering recommender systems. Expert Systems with Applications, 38(12), pp.14609–14623.

[20] Karydi, E. & Margaritis, K., 2014. Parallel and Distributed Collaborative Filtering: A Survey. arXiv preprint arXiv:1409.2762. Available at: http://arxiv.org/abs/1409.2762.

[21] Chen, B. et al., 2005. Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis. 2005 IEEE Computational Systems Bioinformatics Conference, Workshops and Poster Abstracts, pp.105–108.

[22] Barragáns-Martínez, A.B. et al., 2010. A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. Information Sciences, 180(22), pp.4290–4311.

[23] Albadvi, A. & Shahbazi, M., 2009. A hybrid recommendation technique based on product category attributes. Expert Systems with Applications, 36(9), pp.11480–11488. Available at: http://dx.doi.org/10.1016/j.eswa.2009.03.046.