Analysis of Fire Risks in Domestic Premises in England

XXXXXXXXXXXXXXXX

1. ***Abstract****: In order to produce useful information, data*

*mining methods are now greatly needed due to the availability of enormous volumes of data. An essential supervised learning strategy in this new age of digitisation is categorisation, which is covered in-depth information regarding data mining techniques. In order to execute classification analysis on various types of accessible data, WEKA software is also considered a tool of choice. With the help of WEKA, it is possible to extract useful information from data and choose an appropriate method for producing an accurate prediction model. Tasks involving categorisation and analysis comprise several machine learning techniques.*

*Additionally, it includes general-purpose environment tools for data preparation, regression, classification, association rules, clustering, feature selection, and visualisation. This paper, which offers several machine learning techniques for data mining jobs, will have categorisation as its primary emphasis, and the algorithms may be applied to the prepared dataset. The classification approach differentiates the seed types Canadian, Rose, and Kama according to their geometrical characteristics. We use intelligent system methods to distinguish between the wide varieties of wheat kernels. From the UCI machine learning database, we selected the kernel dataset. Specifically, there are seven geometrical characteristics (area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, length of kernel groove). In WEKA classification, the following algorithm will be considered, Decision Tree(DT), Neural Network(NN)/Multi-Layer Perception(MLP), Support Vector Machine(SVM), Naïve Bayes(NB), and K-Nearest Neighbors(KNN). After that, the training test option that considers predicting model accuracy is Use Training Set(UTS), Cross-Validation(Folds-10)(CV-10), and Percentage Split(66%)(PS-66). The accuracy percentage for all the algorithms are:*

*Decision Tree (DT): UTS 100%, CV-10 89.5238%, PS-66 91.5493%*

*Neural Network (NN): UTS 99.5238%, CV-10 95.2381%, PS-66 95.7746%*

*Support Vector Machine (SVM): UTS 94.2857%, CV-10 93.8095%, PS-66 91.5493%*

*Naïve Bayes (NB): UTS 90.9524%, CV-10 91.4286%, PS-66 91.5493%*

*K-Nearest Neighbors (KNN): UTS 100%, CV-10 94.2857%, PS-66 91.5493%*

*Finally, It was observed that the Training Set method gives higher accuracy than Cross Validation during the classification process.*

1. ***Introduction:*** *In the twenty-first century, digitalisation and*

*data mining have made managing the exponential rise of data much more accessible. Businesses built data warehouses with millions of entries and characteristics, but they still await a return on their investment. They cannot create enough because they need more employees, expertise, and equipment. The automated categorisation of instances using data patterns gleaned from a dataset is known as data mining. Various algorithms have been created and put into use to extract information and find knowledge patterns that can be helpful for decision assistance. Data preparation, pattern recognition, clustering, and classification are some of the standard data mining technologies, also known as KDD (knowledge discovery in databases) [1]. The classification components of data mining will be the main emphasis of this study. In classification, models are trained using a training data set to create classifications. Creating a data collection, dimension reduction, feature selection, model selection, training the model, and predicting unidentified input samples are all steps in the classification process. There is a wealth of literature for data mining classification methods in the biological sciences and agriculture. [2]. One of the most popular cereal grains eaten worldwide is wheat. Sorting wheat makes it more competitive. Manual sorting, in which farmers used to separate wheat using just their eyes, is exceedingly labour-intensive, necessitates a larger workforce, and produces inaccurate results [3]. It is a laborious, time-consuming, expensive, and sometimes inaccurate process to analyse and manually sort wheat using the human eye and to sort wheat using machines that consider just a few aspects. As a result, Sorting the wheat is crucial, and erroneous sorting may be prevented by not categorising the wheat. Machine learning has shown to be a valuable tool for analysis across a range of fields and applications. Since digitalisation is being utilised in the project to categorise wheat, it does away with antiquated, manual processes and human error.*

1. ***Dataset Description:*** *The UCI website, an excellent dataset*

*repository, is where the wheat seed data of size 13.2kilobytes was collected. Two hundred and ten(210) wheat seed samples from the three wheat classes, Kama, Rosa, and Canadian, are gathered for the categorization procedure. Seeds are divided into three types of wheat based on seven geometrical or morphological characteristics (area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient, and length of kernel groove).*

*Chart, bar chart, histogram

Description automatically generated*

*Fig 3.1. All attributes for the dataset.*

*The above fig 3.1, shows that the output data class ratio is complete and balanced.*

1. ***Methodology:*** *The wheat dataset was read through Jupiter*

*Notebook with the following pandas and numpy sets of python codes below and screenshots for executing the various codes.*

*>>>import pandas as pd*

*>>>import numpy as np*

*>>>wheatData = pd.read\_csv('/Users/apple/Desktop/Data Science Foundation/Course Work/WheatData(1).csv')*

*>>>wheatData*

*Table

Description automatically generated*

*Fig 4.1. 210-instances, 7-attributes and 1-class.*

*>>>wheatData.info()*

*>>>wheatData.shape*

*Table

Description automatically generated*

*Fig 4.2. The numbers of rows & columns, data-type float64-attributes and object-class.*

*>>>wheatData.columns*

*Graphical user interface, text, application

Description automatically generated*

*Fig 4.3. Columns name.*

*>>>wheatData.describe().transpose()*

*Table

Description automatically generated*

*Fig 4.4. The count, mean, stand deviation, minimum & maximum attributes value, and percentile attributes.*

*To analyse the link between structural activity and classification, Weka software is employed. After processing for categorisation, the data mentioned above was produced in CSV format. In our system, four classification techniques—Function, Bayes, Meta, and Lazy—are tested on seed datasets to assess their performance.*

1. ***Modelling:***

*Naïve Bayes (NB): Naive Bayes is a set of supervised learning algorithms, and the input values are traditionally assumed to be nominal, even if the distribution assumption supports numerical inputs. The posterior probability for each class is calculated, and the class with the most significant probability is predicted. As a result, it handles issues involving binary and multi-class classification.*

Text

Description automatically generated with low confidence

***Fig 5.1. bayes.NaiveBayes algorithm accuracy data***

*Decision Tree(DT): Decision trees/ Classification And Regression Trees (CART) support classification and regression problems. It analyse a data instance, and they build a tree, starting at the root and progressing to the leaves (roots) until a prediction can be formed. Building a decision tree involves repeatedly choosing the optimum split point to use as a prediction tool until the tree reaches a set depth.*

A picture containing text

Description automatically generated

***Fig 5.2. trees.RandomTree algorithm accuracy data***

*K-Nearest Neighbors (KNN): KNN is a simple algorithm that supports both classification and regression. When producing a prediction, it stores the whole training dataset and queries it to find the k most similar training patterns. As a result, the sole computation is querying the training dataset when a prediction is needed. There is no model other than the primary training dataset. It is a straightforward method, yet it makes just one significant assumption—that the distance between data instances matters for generating predictions. As a result, it often produces excellent performance. KNN will use the mode (most frequent class) of the k most comparable cases in the training dataset for generating predictions for classification tasks.*

Text

Description automatically generated with medium confidence

***Fig 5.3.*** ***lazy.IBk algorithm accuracy data***

*Neural Network (NN): Neural Network is a machine learning model that mimics the functions of the human brain by taking inspiration from biological neural networks. NN comprises three layers: the input layer, the hidden layer, and the output layer.*

*1. Input Layer indicates the input variables to be supplied into the network.*

*2. Hidden Layer are the training-relevant computation layers (or parameters).*

*3. Output Layer refers to the model's output, using the class label in a*

*classification task or the actual number in a regression task as examples.*

Diagram

Description automatically generated with low confidence

*Fig 5.4. Describes input, hidden, and output layer mesh.*

*Text

Description automatically generated with medium confidence*

***Fig 5.5.*** ***functions.SMO algorithm accuracy data***

*Support Vector Machine (SVM): Support Vector Machines were created for binary classification issues using numerical input variables, while the method has since been extended to cover multi-class classification and regression issues. Additionally, it automatically transforms nominal values into numerical values. Additionally, before being utilised, input data is standardized. SVM identifies the optimal line to divide the data into two groups. In order to do this, an optimisation method is used; however, it only takes into account data instances from the training dataset that is most closely related to the line that best demarcates the classes. The examples are referred to as support vectors, therefore the technique's name.*

*Text

Description automatically generated with medium confidence*

1. STORAGE OF THE DATA

The data is a recorded in Financial Years from 2010/2011 to 2016/2017. The data was downloaded as a CSV file (comma separated variable). It was then converted into a JSON file (JavaScript Object Notation) which is lightweight, text-based, and human readable [6], imported into MongoDB and stored in a single collection. An example of a single JSON document from the data is shown in [Figure 1.](#_bookmark0)



Figure 1: JSON document

Once the domestic fire data set was downloaded, a decision had to be made regarding which platform would be most suitable to store and analyse the data. Ultimately MongoDB, which is a Document Oriented Database, was chosen as the technology platform for storage and querying of the data.

MongoDB is an open source non-relational database, it stores data as documents in a binary representation called BSON (Binary JSON) [6]. Groups of documents are then stored as part of collections which have dynamic schemas – permitting documents within a single collection to have different “shapes”. MongoDB also groups collections into databases which has its own permissions and each database is stored in separate files on disk. The architecture layout is shown in [Figure 2.](#_bookmark1)

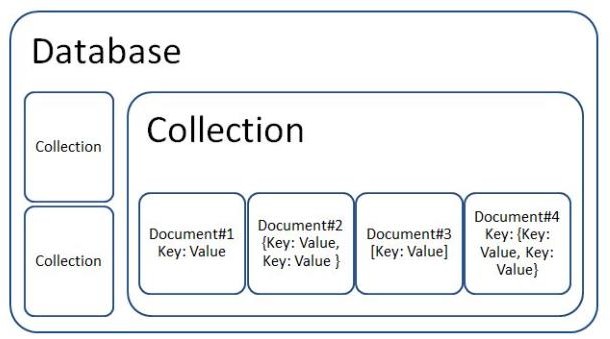


Figure 2: Document Oriented Database Architecture [7]

As part of this analysis, a single database has been crated, which contains a single BSON collection of over 230000 documents.

A Relational Database Management System (RDBMS) was also considered for storage and analysis of the data. MySQL was explored as a possible alternative; however it was

considered inappropriate for analysis of this data based on the following:

* + A comparison of MongoDB and MySQL was performed in [8] on two separate databases of similar structure and similar number of fields (MySQL) and documents (MongoDB). It was found that MongoDB had lower execution times for the four basic operators (Insert, Query, Update & Delete).
  + In future, domestic fire data from other regions such as Ireland, Scotland and even the USA will be added to the database. This will highly increase the amount of data and likely result in a varied structure, therefore the efficiency of MongoDB to handle large volumes of data will be required.
  + Given that only a single large collection is to be assessed, no complex Join operators on multiple tables would be required and all queries can be performed on the same collection. Therefore, while the data is suitably structured and would lend itself to a MySQL database, MongoDB is considered to perform more efficiently.
  + MongoDB does not involve the use of predefined schemas. This will be useful in future assessments where fire data from other regions will be added in to the database which will have different shapes to the data set in question in this report.

The data was queried and analyzed through the mongo shell, which is an interactive JavaScript interface to MongoDB [9]. In addition, use was made of Studio 3T which is a MongoDB Graphical User Interface (GUI) [10] and reduces the requirement for writing programming query’s and applies filters to the data through drag and drop functionality.

Smaller scale key queries were undertaken in Mongo through the shell and Studio 3T which could then be plotted in tables and graphs in Microsoft Excel. Additionally, where more complex data needed to be analyzed, subsections of the data were filtered in the shell and Studio 3T and exported as CSV files to be read in excel with key items visualized.

1. ANALYZING THE DATA

The data was analyzed using Microsoft Excel and Rapid Miner. Initially, the total number of domestic fires were plotted for each Financial Year, to determine if there were any trends in the number of fires occurring on a yearly basis. The total amount of fatalities/casualties was then determined for each dwelling type and a % of fatalities/casualties was then plotted per the number of fires in each type of dwelling.

To delve further into the data, additional items were assessed which could offer insight into the reason behind the number of fatalities occurring. These included:

* + Alarm system activation;
  + Cause of fire;
  + Location where the fire started;
  + Item ignited; and
  + Occupancy type

1. RESULTS

From Financial Year 2010/11 to 2016/17, there was a total percentage reduction of 17.1% in the number of fire incidents recorded. This is shown in [Figure 3.](#_bookmark2) By Financial Year 2016/17 there were 6273 fewer fires recorded than in 2010/11. The overall number of fatalities reduced but not by an appreciable amount.

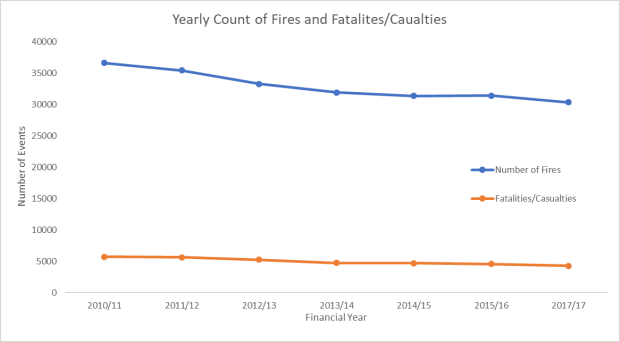


Figure 3: Number of Domestic Fires and Fatalities/Casualties Each Financial Year

The total amount of fatalities/casualties which occurred in each type of dwelling type was calculated ([Figure 4](#_bookmark3)) and an overall percentage of fatalities per number of fires was found in each case ([Figure 5](#_bookmark4)).

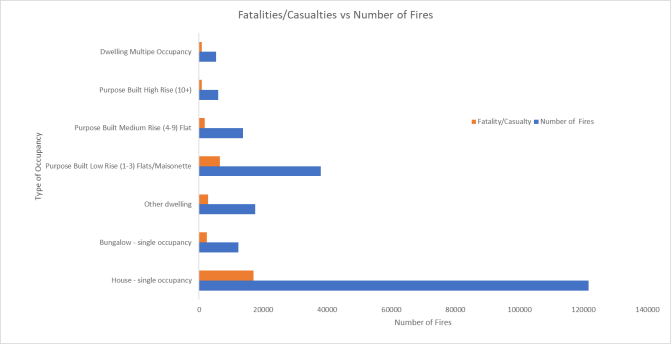


Figure 4: Number of Fatalities/Casualties recorded in total number of fire events

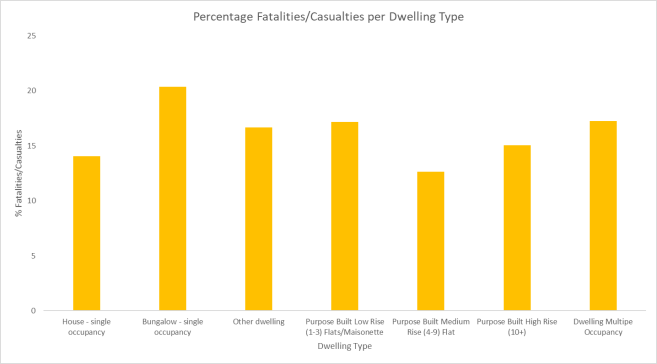


Figure 5: Fatalities/Casualties per Dwelling Type

It was then determined that four dwelling types should be analyzed in detail. While bungalow occupancies had the highest percentage of fatalities/casualties, this type of dwelling also had the third lowest number of fire incidents recorded. Therefore,

this was not selected for inclusion in the detailed analysis. It is considered that bungalows could fall under the Houses in Single Occupancy category also, for which the most number of fires were recorded. Therefore, analysis of the Houses data would likely yield insights into the Bungalow data also. The four types of dwelling chosen were:

* + Houses in single occupancy;
  + Low rise residential (flats/masionettes);
  + Medium rise residential (4-9 Storeys); and
  + High rise residential (10+ storeys)

The top 6 causes of fires in the above dwelling types are shown in [Figure 6.](#_bookmark5)

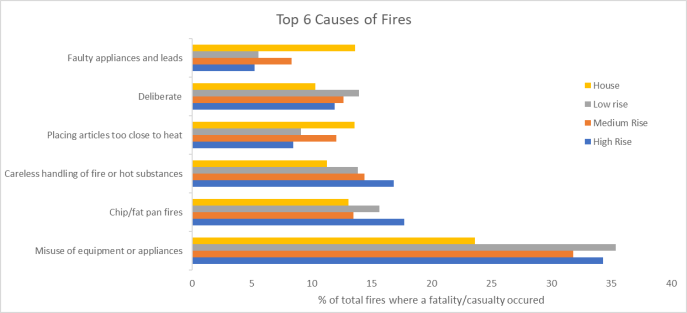


Figure 6: Top 6 causes of Fires where fatalities/casualties occurred

The top 6 fire locations were also recorded as percentages of the total number of fires where there was a fatality/casualty. This is highlighted in [Figure 7.](#_bookmark6)

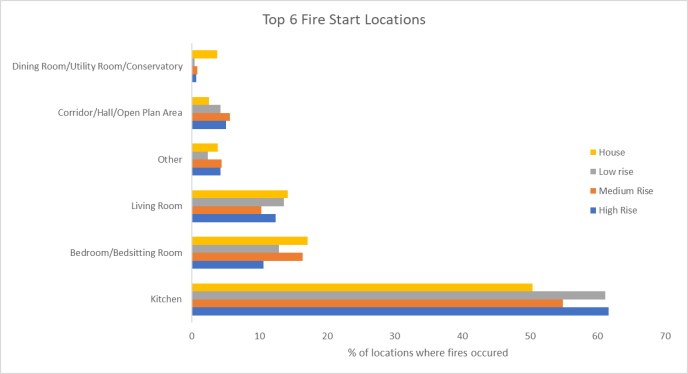


Figure 7: Top 6 Fire Start Locations

The various types of occupancy characteristics were also assessed with regard to determining trends within the data. The top 6 most impacted occupancy types where fatalities/casualties occurred were recorded, see [Figure 8.](#_bookmark7)

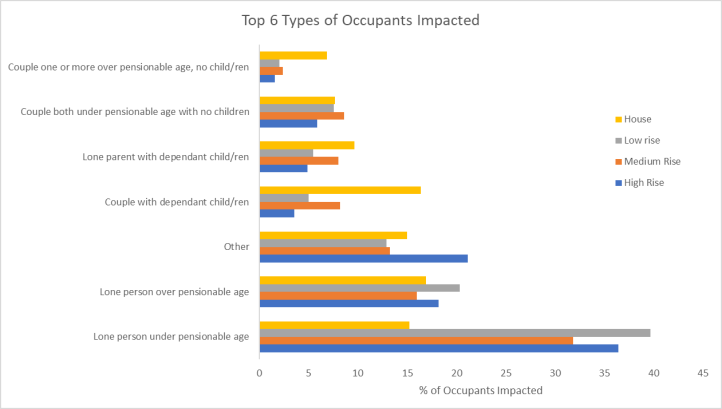


Figure 8: Top 6 Occupant Types Where Fatalities Occurred

Fire alarm data was also considered applicable in order to delve into the data and extract patterns. [Figure 9](#_bookmark8) outlines data relating to fire alarm systems in the selected residential dwellings and the overall percentage of instances where the scenario occurred.

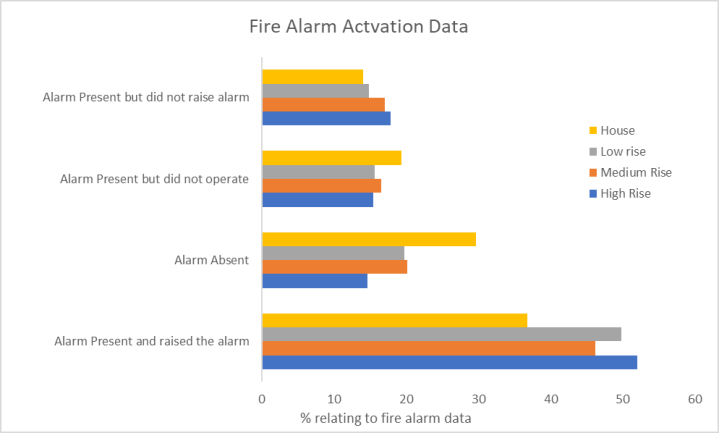


Figure 9: Fire Alarm Data

1. ANALYSIS OF RESULTS

Various key observations can be made from examining the data in detail.

The total number of fire events recorded each year appears to be reducing at a steady rate. This could be due to greater awareness being raised about fire safety in the home. Data is not available for Financial Year 2017/18, analysis would be required of the most recent data before a solid conclusion could be made regarding whether the number of fires were still being reduced.

The most number of fires occur in single occupancy dwellings. This may not necessarily be an indictment of houses as being more prone to fire, but rather that there are more types of these buildings when compared to the other three types which were selected. Therefore, it was necessary to plot all further data as a percentage of the total number of instances where they occurred. There does not appear to be an appreciable reduction in the amount of fatalities/casualties however.

The majority of fires across all dwelling types started in kitchen areas and occurred for the most part due to misuse of appliances or equipment and fires as a result of chip pans. For single occupancy houses, the causes of fires was more evenly distributed amongst the various possibilities. For example, the percentage of fires caused by faulty appliances and leads was highest in single occupancy houses when compared to the other dwelling types.

The occupant types most likely to be the subject of a casualty or fatality are those people who live on their own, particularly those under pensionable age.

Of interest also, is that in a large number of cases where a casualty/fatality was recorded, the fire alarm system was present, operational and raised the alarm. This is especially so in the case of high-rise residential blocks, given that when a fatality/casualty was recorded, in over 50% of these cases, the

fire alarm was raised. This could potentially be as a result of a fire in the apartment of fire origin growing to a large size and not setting off the fire alarm in adjacent apartment until it is too late for occupants to escape prior to injury.

Alarms appeared to be absent in single occupancy houses more than in any other dwelling type. This could potentially be down to inability to enforce house occupants to regularly test their smoke alarms or in some cases they may have been removed by the occupant. Whereas in managed residential buildings, the building owner/landlord has more of an input into ensuring adequate installation and operation of fire alarm systems.

1. CONCLUSION

From the data analyzed, a few points are abundantly clear in instances where fatalities/casualties occur:

* Most domestic fires happen in the kitchen;
* The fires are mainly cooking & appliance related;
* Lone occupants are most at risk of death or injury; and
* Even when a fire alarm activates, there is still a high likelihood of a fatality/casualty.

Based on the above, one can imagine many scenarios where all four of the above points are included. However, more research is required to delve into these items further e.g. Why when he fire alarm activates in so many cases, is there still a casualty? The most obvious answer is the sleeping risk associated with domestic premises i.e. the alarm is simply not heard given that occupants are not awake. However further research needs to be done into methods to increase alarm effectiveness.

1. FURTHER STUDY

It is recommended that in further studies the following key items should be taken into consideration:

* The impact time of year has on fire casualty rates. e.g. does the casualty rate increase at Christmas time?
* City vs Smaller settlements: Delve further into the impact living in a city has vs a town or village on dwelling fire casualties.
* Fire alarm system: review the type of system installed in premises and comment on suitability for the premises i.e. the system may have operated however may not have operated a quickly as necessary.
* What type of appliances cause kitchen fires and is this a fault with the appliance or the occupant?
* Investigate similarities/differences in domestic fire fatality/casualty rates in other regions e.g. in the USA to determine if any lessons can be learnt and applied to the UK Building Regulations.
* Given the amount of fatalities to lone occupants, development of a community alarm/mobile application which would serve occupants in blocks of flats could be looked in to. Such an app might alert neighbors of lone vulnerable occupants that a fire has occurred in their apartment and so they can aid in an escape.

REFERENCES

1. S. Lundy, “Hackitt Review - Building Regulations and Fire Safety.” [Online]. Available: https://4siteconsulting.co.uk/blog/hackitt- review/. [Accessed: 01-Dec-2018].
2. J. Hackitt, *Building a safer future - Independent Review of Building Regulations and Fire Safety: Interim Report*, no. December. 2017.
3. G. Genco, “Lacrosse Building Fire Report,” vol. 2015, no. April, 2015.
4. C. G. Barney Henderson, “Dubai Skyscraper Fire,” *Telegraph.co.uk*. [Online]. Available: https://[www.telegraph.co.uk/news/2017/08/03/dubai-skyscraper-](http://www.telegraph.co.uk/news/2017/08/03/dubai-skyscraper-) fire-1100ft-torch-tower-engulfed-flames/. [Accessed: 02-Nov- 2018].
5. H. Office, “Fire Statistics: Incident Level Datasets.” [Online]. Available: https://data.gov.uk/dataset/527847a2-83ab-4f01-81c5- a0b58cc50bc8/fire-statistics-incident-level-datasets. [Accessed: 11- Nov-2018].
6. “MongoDB and MySQL Compared.” [Online]. Available: https://[www.mongodb.com/compare/mongodb-mysql.](http://www.mongodb.com/compare/mongodb-mysql) [Accessed: 09-Dec-2018].
7. S. Subramanian, “A Primer on Open-Source NoSQL Databases.” [Online]. Available: https://dzone.com/articles/a-primer-on-open- source-nosql-databases. [Accessed: 07-Dec-2018].
8. A. O. Cornelia GYORODI, Robert GYORODI, George

PECHERLE, “A Comparative Study\_ MongoDB vs. MySQL.pdf,” no. JUNE, 2015.

1. “Mongo Shell.” [Online]. Available: https://docs.mongodb.com/manual/mongo/. [Accessed: 07-Nov- 2018].
2. “Studio 3T.” [Online]. Available: https://studio3t.com/. [Accessed: 12-Dec-2018].

*Appendix A – MongoDB Process*

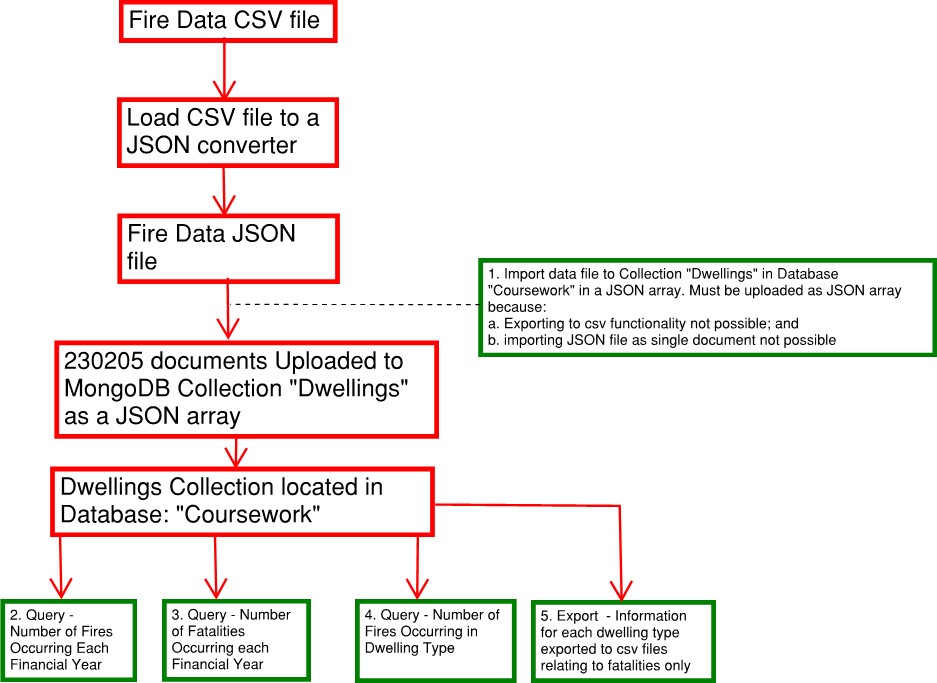


Figure 10: MongoDB Process

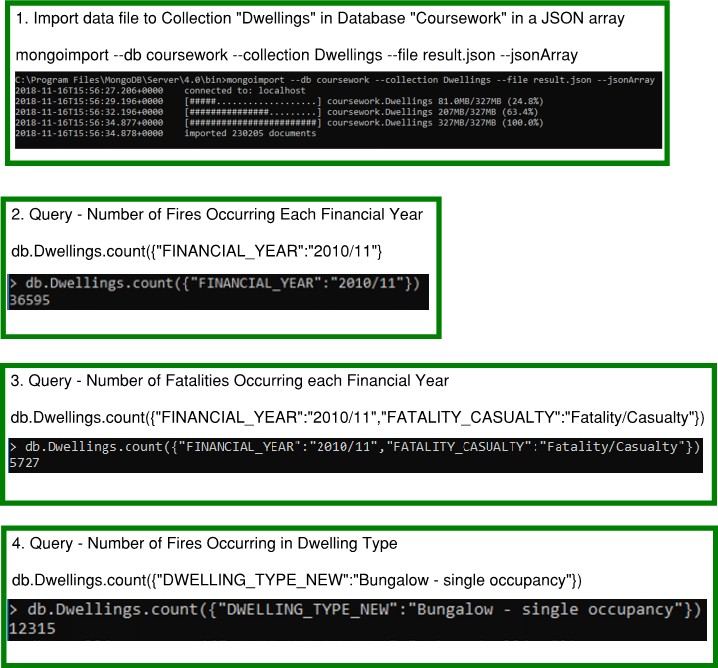


Figure 11: Import of Data and Filter Queries carried out in Mongo Shell

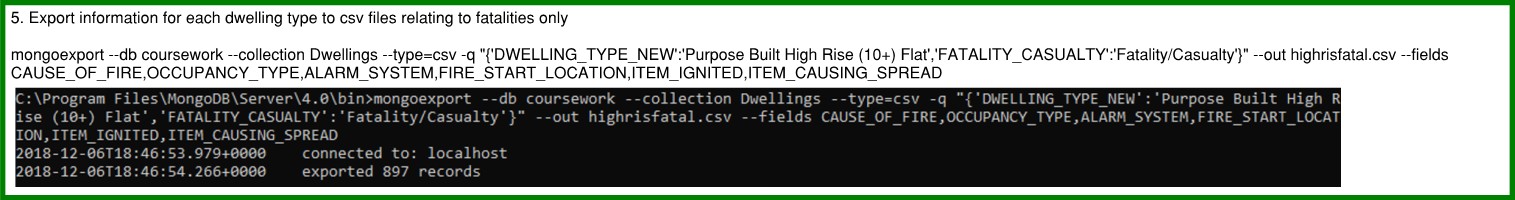


Figure 12:Export Filtered Data to CSV File

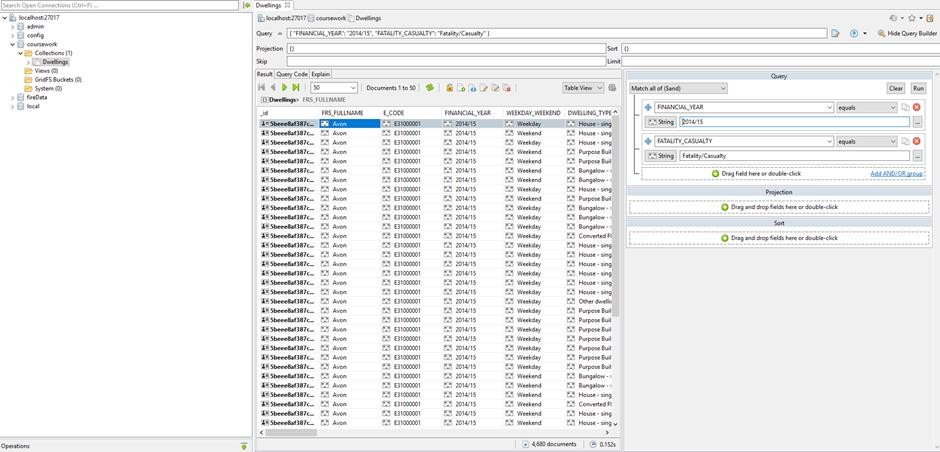


Figure 13: Filtered Data using Studio 3T – Finding Count of Fatalities Each Year

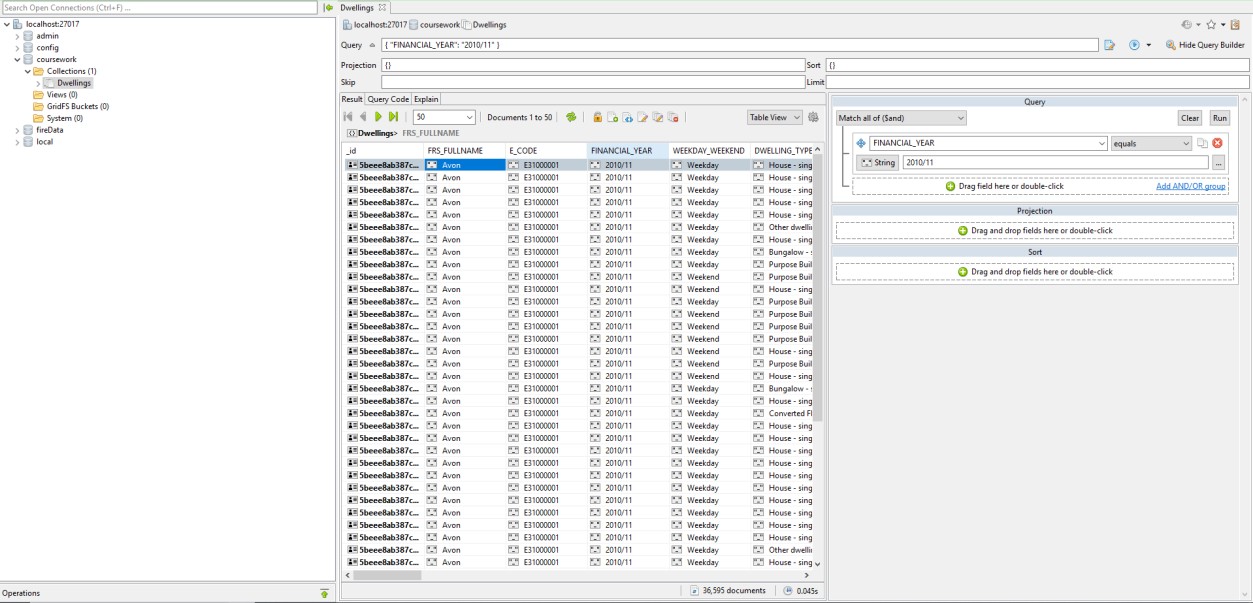


Figure 14: Filtered Data using Studio 3T –Counting Number of Fires Each Year

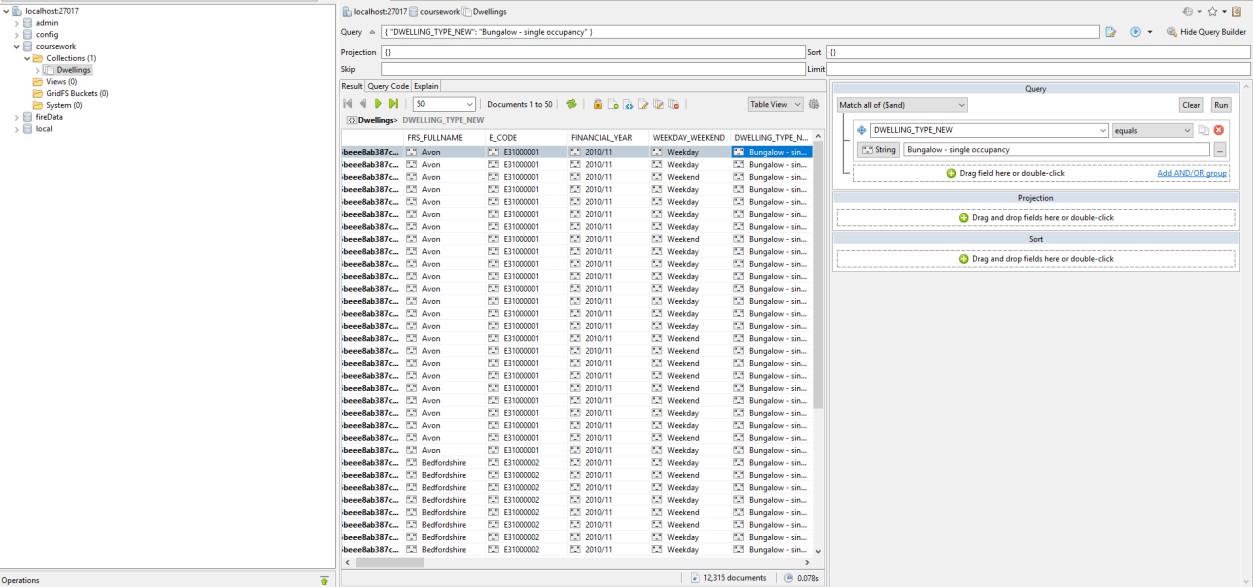


Figure 15: Filtered Data using Studio 3T –Counting Number of Fires in Each Dwelling Type