Mobility Behavior of Mobile and Immobile People

G. Arnarson, F. Bartke, M. Van Laere, S. Nguyen

2025-01-20

1 Introduction

Mobility plays a very important role in society and can be a huge inconvenience. People can be immobile for a lot of reasons. In this project, it will be researched what characterizes immobile people in and around Grenoble and these people will be predicted based on certain variables. Several different models will be used to see which model is most suited to predict. This research can be useful for trying to decrease immobility. The following research questions are posed: - How are immobile people characterized? - Which model can best predict their immobility?

2 Literature Review

Some literature was reviewed to determine the relevant variables and get some more insight in mobility data.

Paper 1: The spatial dimensions of immobility in France In this paper, immobility is defined as not leaving your house. Structural Equation Modelling was used to declare immobility. The used variables were either social like retirement, income and car ownership, spacial like population density and region and individual like age, physical

limitations, and education. It was concluded that social, spacial and individual variables all have a significant impact on the mobility of people in France.

Paper 2: Determinants of car ownership among young households in the Netherlands Households from 18-29 years are researched to see what influences whether they have a car or not. Logistic regression was used to declare this variable using the grade of urbanization, age, ethnicity, family structure and income. The paper concludes that the decreasing possession of a car can be partially declared by the growing urbanization.

Paper 3: Modelling car ownership in urban areas: a case study of Hamilton, Canada An online survey and a GIS was used to collect data. Then, multinomial logit modelling was used to predict the likelihood of car ownership. Household structure, income, income, working status, population density and proximity to and accessibility for facilities turned out to be the most relevant variables.

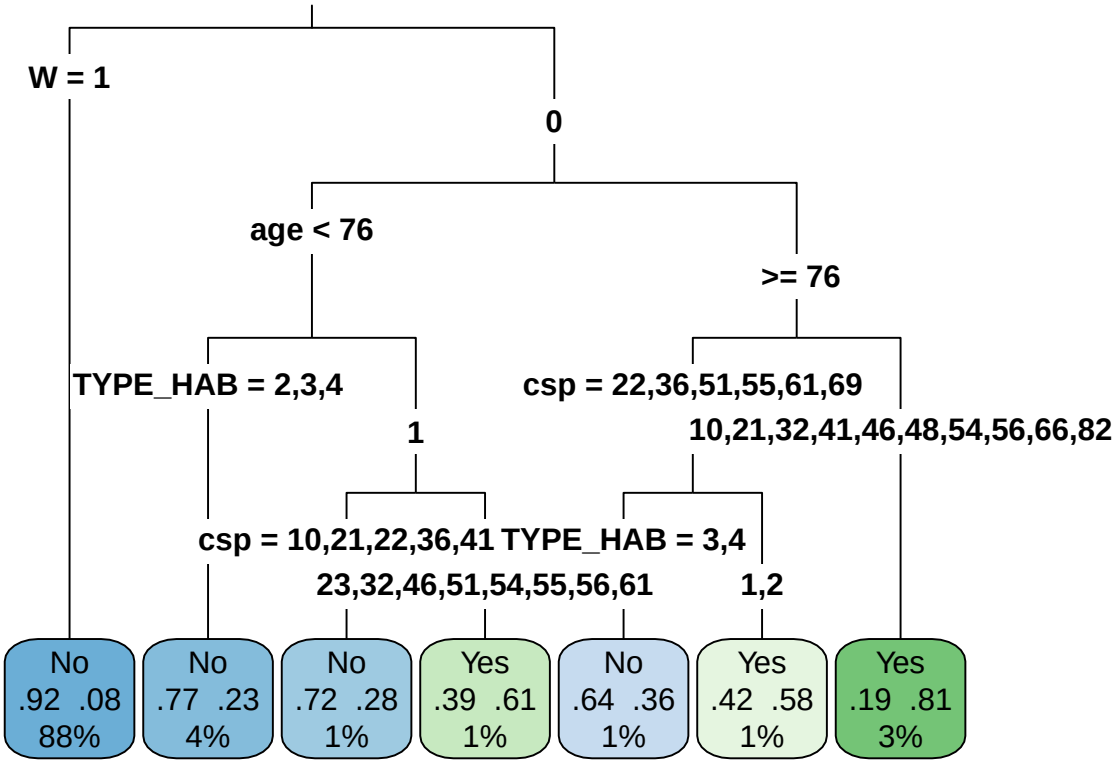
It can be seen that in all papers social, spacial and individual variables have a significant impact. Logit modelling seems to be a good way of investigating these variables and their interactions among themselves.

2.1 Variables Choice

Based on the reviewed literature, the following variables are chosen to be further researched: age, gender, type of living space, socio professional category, occupation, retirement, work from home, car ownership, number of cars, parking difficulty, living region and finally, a variable W is created. This variable is defined as whether a person has a direct mean of transport.

3 Dataset Preparation

Since variables from all datasets are being used, these all needed to be merged. Then, these variables were selected, categorized, renamed and if necessary redefined. The W variable was also defined: if Nb_2Rm, Nb_velo, dispovp, ABO_TC, abonpeage and LIEU_STAT are all 0, then W is 1. Some variables have a lot of categories so these needed to be grouped. To decide on how to group these, a decision tree was used. This is how the variables age_grouped, OCCU1_grouped, TYPE_HAB and csp_grouped were defined. The following tree was used. Then, variables were iteratively removed from the tree to group variables that are not shown on the tree yet. Age is split in 0-5, 5-76 and 76+. Csp is (mostly) split in active working people and inactive people, scholars or interns. OCCU1 is split up in 3 groups: 1. working and stay-at-home, 2. students and interns, 3. inactive and retired. Homes are split between individual homes and collectives.

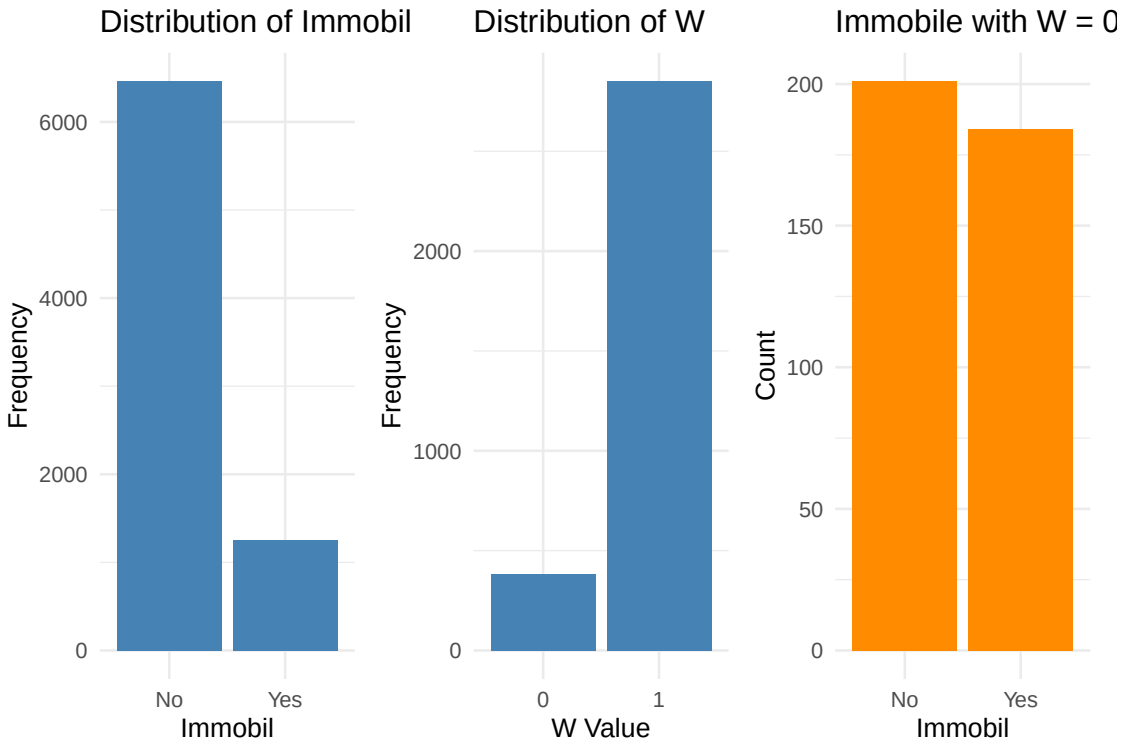


This results in the following dataset. This is the dataset that will be used for all further analysis. (except for the GIS) Most variable names speak for themselves. Parking_diff stands for parking difficulty. Fullygrouped is the sort of region a person lives in, divided in 3 levels: city, mountains and countryside. It is clear that plenty social, individual and spatial variables are in this dataset.

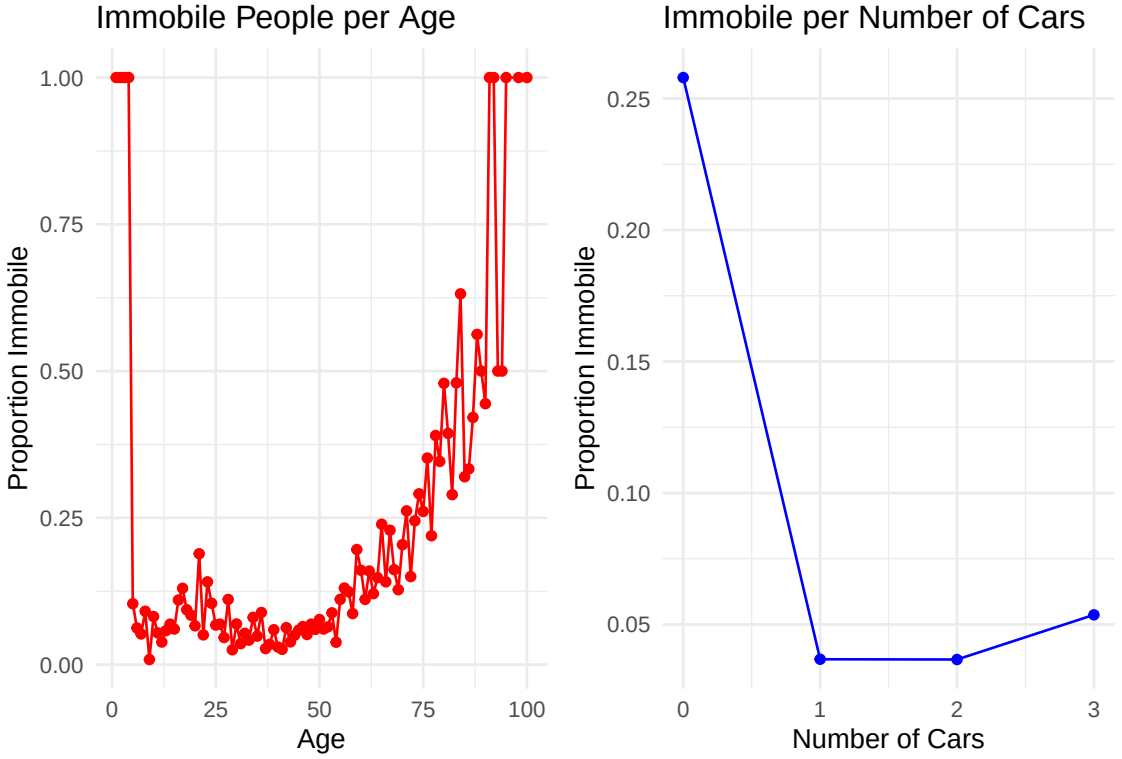
```
## 'data.frame':    7720 obs. of  16 variables:
##  $ id_pers      : num  1.01e+08 1.01e+08 1.01e+08 1.01e+08 1.01e+08 ...
##  $ immobil      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ sexe         : Factor w/ 2 levels "Male","Female": 1 1 2 1 2 1 2 1 2 1 ...
##  $ dispovp      : num  0 1 2 3 3 3 2 2 2 0 ...
##  $ age          : int  24 42 28 33 30 30 27 57 52 17 ...
##  $ has_car      : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 2 2 2 2 1 ...
##  $ W            : Factor w/ 2 levels "No","Yes": 2 2 NA 2 NA NA NA 2 NA NA ...
##  $ TYPE_HAB     : Factor w/ 2 levels "Groep 1","Groep 2": 2 2 NA 2 NA NA NA 2 NA NA ...
##  $ zoneres      : Factor w/ 32 levels "1","2","3","4",..: 3 3 NA 1 NA NA NA 2 NA NA ...
##  $ travdom      : Factor w/ 2 levels "1","2": NA 2 NA 2 NA NA NA 2 NA NA ...
##  $ parking_diff : Factor w/ 2 levels "No","Yes": 1 1 NA 2 NA NA NA 1 NA NA ...
##  $ retrait      : Factor w/ 2 levels "No","Yes": 1 1 NA 1 NA NA NA 1 NA NA ...
##  $ fullygrouped : Factor w/ 3 levels "City","Montagnes",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ age_grouped  : Factor w/ 2 levels "5-76","76+": 1 1 NA 1 NA NA NA 1 NA NA ...
##  $ csp_grouped  : Factor w/ 2 levels "Groep 1","Groep 2": 1 1 NA 1 NA NA NA 1 NA NA ...
##  $ OCCU1_grouped: Factor w/ 3 levels "Groep 1","Groep 2",..: 1 2 NA 3 NA NA NA 2 NA NA ...
```

4 Data exploration

In this section, some further exploring of the data will be done to identify patterns, with a focus on the variables mentioned in the assignment. Below are the distribution of W and immobil. It turns out that around 85% of people are mobile. Noticeably, the distribution of W is similar. In the third plot we can see that around half of the people that have no mean of travel are immobile. So, whether or not a person has a direct mean of transport has a big influence on a persons mobility.



The following graphs show that people under the age of 5 and over the age of 75 are more often immobile than people in between these ages. Thus, the grouping decided by the regression tree makes sense. As can be expected, people without a car tend to less mobile.



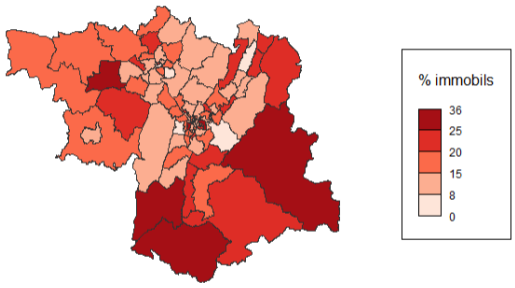
The following table shows the correlation between the chosen variables and the immobil variable. All variables, except for the sort of living area variable are significant. As expected from the literature, possesion of a car or a mean of travel is the most significant. The age is also very significant.

```
##         Variabel  Testtype       P_value Significant
## 1    csp_grouped Chisquare  9.218181e-10         Yes
## 2           sexe Chisquare  2.319995e-04         Yes
## 3        has_car Chisquare 1.168671e-145         Yes
## 4    age_grouped Chisquare  8.693929e-69         Yes
## 5       TYPE_HAB Chisquare  1.094061e-02         Yes
## 6   parking_diff Chisquare  1.334971e-06         Yes
## 7              W Chisquare 1.430520e-100         Yes
## 8  OCCU1_grouped Chisquare  7.925460e-30         Yes
## 9        travdom Chisquare  8.572656e-06         Yes
## 10  parking_diff Chisquare  1.334971e-06         Yes
## 11       retrait Chisquare  5.260818e-27         Yes
## 12  fullygrouped Chisquare  7.614163e-02          No
## 13       id_pers    T-test  3.925016e-01          No
## 14       dispovp    T-test 8.149656e-158         Yes
## 15           age    T-test  1.452498e-06         Yes
```

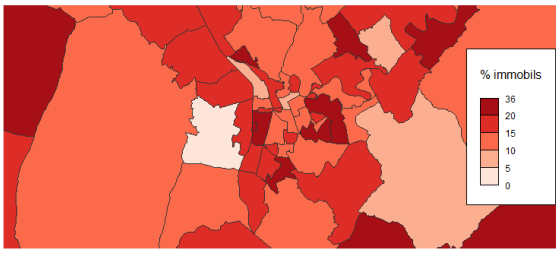
4.1 Mapping

The following map shows the percentage immobile people per zone in and around Grenoble. This confirms the earlier conclusion that the sort of region a person lives in does not necessarily affect their mobility. Since, people living in the Chartreuse mountain region seem to be rather mobile, while some people in Grenoble are immobile. It is also noticable that people living in the Belledonne region are less mobile than people in the Chartreuse region, so the mobility does not necessarily depend on the sort of geologic region, but more so on different factors.

Percentage immobile people per zone



Zoomed: Percentage immobile people per zone



5 Modeling

In this section, 4 different models will be used to do further research in mobility and will be compared to one another and a benchmark model to see which model is most suited for predicting in this context. The same 3 declaring variables will be used to declare for each model: age_grouped, dispovp and W. The models that will be used are: logistic regression, decision trees, random forest and neural networks.

5.1 Logit model

As seen in the literature, the logit model is a good model to do some analysis on the different variables. After testing each selected variable in a separate model, it is concluded that every one of them is significant in this separate model. However, when putting them all together in a model, they are no longer significant. They least significant variable gets iteratively removed in the following order to reduce multicollinearity and achieve a model with only significant variables left: dispovp, sexe, fullygrouped, parking_diff, retrait. This leaves only significant variables in the model. This gives us an idea of which variables have the biggest influence. In a multivariate model like this, it is possible that a very significant variable becomes redundant, because another variable has a similar effect. It makes sense that dispovp was the least significant, since W is still in the model. Apparently, the influence of the gender of a person is not that big of an influence. As concluded earlier with the map and chisquared, the region a person lives in does also not necessarily influence their mobility. Parking difficulty probably has some overlap with W and retrait has some overlap with OCCU1 and csp.

The resulting model is shown below. The model has a McFadden's R-squared of 0.71. This is on the higher side. With an AIC of 2036.9, the model seems to have a good balance between fit and complexity. The difference between residual and null deviance shows that the model improves significantly with the chosen variables. The baseline odds (intercept) show that the chance for immobility is lower than mobility. People over the age of 76 have an about 250% higher chance of being immobile. People that are non working have 146% higher odds of immobility than working people. People having a car have an 82% lower chance of being immobile. People living in a collective building have a 29% lower chance of immobility. Students and interns have a 327% higher chance of being immobile than the reference group, active working people. This is likely partly due to children being included in this group. Finally, having a direct mean of transport decreases chances of immobility by 78%. It should be noted that these numbers should be taken with a grain of salt, since only 3236 observations are taken into account. So, these odds ratios are in a pretty wide confidence interval. But, they give a good idea of the influence of these variables.

```
##
## Call:
## glm(formula = immobil ~ age_grouped + csp_grouped + has_car +
##     TYPE_HAB + OCCU1_grouped + W, family = binomial, data = allgreI_filtered)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -0.5462     0.1452  -3.762 0.000169 ***
## age_grouped76+         0.9196     0.1451   6.336 2.36e-10 ***
## csp_groupedGroep 2     0.3804     0.1377   2.763 0.005733 **
## has_carYes            -1.7182     0.2158  -7.961 1.70e-15 ***
## TYPE_HABGroep 2       -0.3401     0.1209  -2.814 0.004898 **
## OCCU1_groupedGroep 2   0.6730     0.1895   3.552 0.000382 ***
## OCCU1_groupedGroep 3   1.1861     0.2928   4.051 5.10e-05 ***
## WYes                  -1.5012     0.1433 -10.477  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 2524.6  on 3235  degrees of freedom
## Residual deviance: 2020.9  on 3228  degrees of freedom
##   (4484 Beobachtungen als fehlend gelöscht)
## AIC: 2036.9
##
## Number of Fisher Scoring iterations: 5
```

```
## McFadden's R-squared: 0.705569
```

```
## Waiting for profiling to be done...
```

```
## [1] "Odds Ratios met 95% betrouwbaarheidsintervallen:"
```

```
##                                 Variable Odds_Ratio  CI_Lower  CI_Upper
## (Intercept)                  (Intercept)  0.5791179 0.4349059 0.7687800
## age_grouped76+             age_grouped76+  2.5082315 1.8857942 3.3322973
## csp_groupedGroep 2     csp_groupedGroep 2  1.4628079 1.1145202 1.9126209
## has_carYes                   has_carYes  0.1793932 0.1176730 0.2745586
## TYPE_HABGroep 2           TYPE_HABGroep 2  0.7116773 0.5607176 0.9009063
## OCCU1_groupedGroep 2 OCCU1_groupedGroep 2  1.9600320 1.3478945 2.8349881
## OCCU1_groupedGroep 3 OCCU1_groupedGroep 3  3.2741628 1.8282013 5.7728864
## WYes                                 WYes  0.2228683 0.1680940 0.2948657
```

Now the predictive power of the logit model will be investigated. First, a benchmark model is defined to see how much the model improve upon this.

```
## Benchmarkmodel results:
```

```
## Accuracy on testdata: 0.8458549
```

```
## Recall: 0
```

```
## Precision: NaN
```

```
## F1-Score: NaN
```

For the logit model, a cut-off value of 0.35 is chosen, since the immobil variable is unbalanced. While this might decrease the accuracy, this will improve the recall. Even with this low cut-off value, the recall is still very low at 0.44. If the cut-off value is lowered further, the accuracy would decrease even further. If a cut-off value of 0.5 would be chosen, the accuracy would increase up to 0.89, but the recall would decrease to 0.24 with a precision of 0.61. This means that the model would be correct more often if it predicts immobile, but it barely does so. Since it is more important to correctly determine true immobile people, the cut-off value with higher recall is chosen. From the dispovp odd ratio we can see that the odds of someone being immobile decreases with around 53% with each extra car they have.

```
##
## Call:
## glm(formula = immobil ~ age_grouped + dispovp + W, family = binomial,
##     data = train_data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.5271     0.1494  -3.529 0.000416 ***
## age_grouped76+   1.0388     0.1644   6.318 2.64e-10 ***
## dispovp         -0.6362     0.1232  -5.165 2.40e-07 ***
## WYes            -1.4936     0.1630  -9.165  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1833.5  on 2274  degrees of freedom
## Residual deviance: 1498.1  on 2271  degrees of freedom
##   (3129 Beobachtungen als fehlend gelöscht)
## AIC: 1506.1
##
## Number of Fisher Scoring iterations: 6


## [1] "Odds Ratios:"


##                   Coefficient Odds_Ratio
## (Intercept)       (Intercept)  0.5902892
## age_grouped76+ age_grouped76+  2.8258734
## dispovp               dispovp  0.5292940
## WYes                     WYes  0.2245544


##           Actual
## Predicted  No Yes
##       No  786  62
##       Yes  64  49


## Accuracy op testdata (Logistisch model): 0.8688866


## Recall: 0.4414414


## Precision: 0.4336283


## F1-Score: 0.4375
```

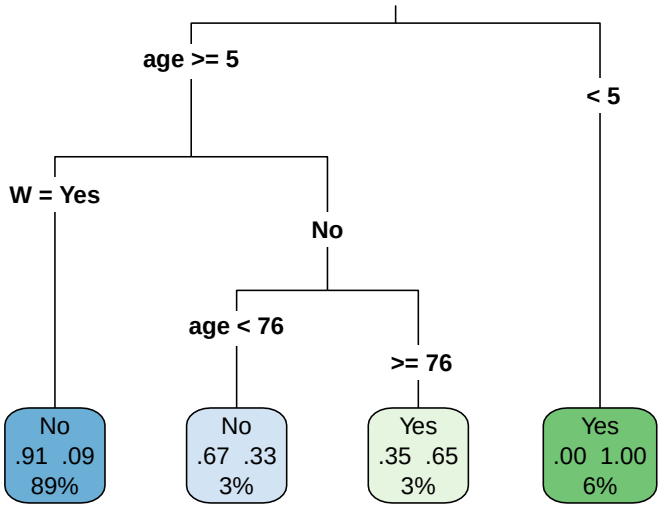
5.2 Decision tree

We trained Decision Trees using the explanatory variables age, dispovp, and the focus variable W to predict the target variable "immobil." Two trees were built: one using the numerical variable age and the other using the categorized variable age_group, as both showed promising results. These trees produced nearly identical structures because age_group and W were the most relevant variables in earlier models as well.

A third Decision Tree was constructed for comparison, incorporating the variables has_car, OCCU1, and W, which are significant predictors of immobility according to the literature. These variables also performed well in regression models.

When comparing the evaluation metrics of the three trees, the third model lacks defined Recall and F1-Score values, as it classified all observations as "mobile." This outcome results from the imbalanced dataset, where False Negatives are absent. Accuracy, as expected, is high in such cases but misleading. To preserve the dataset's balance, children under 5 were retained despite their marginal contribution to classification.

For the first two models, both Accuracy and F1-Score are consistent, with an F1-Score of 0.59. This consistency arises because the split thresholds for age align closely with the boundaries of age_group. The categorization proved optimal, as indicated by initial tests. Compared to earlier models, these trees offer only marginal improvements but excel in simplicity. On the test data, they perform equally well while being easier to interpret.




```
##            used.variables Accuracy Precision Recall F1.Score
## 1 age_group + dispovp + W     0.91      0.44   0.90     0.59
## 2       age + dispovp + W     0.90      0.45   0.87     0.59
## 3     has_car + OCCU1 + W     0.85      0.00    NaN      NaN
```

5.3 Random forest

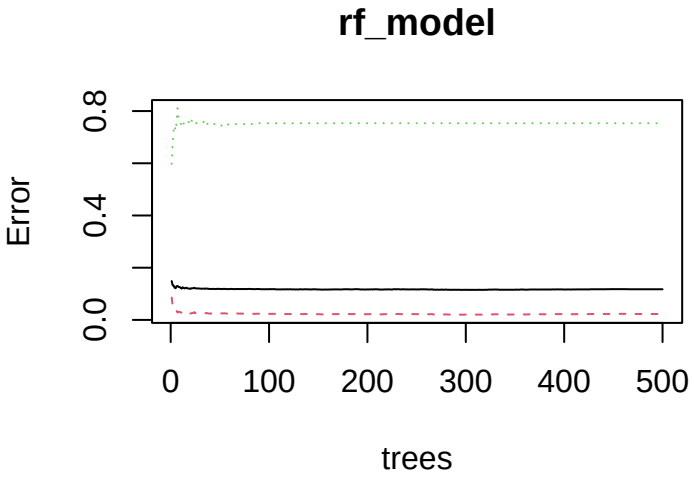
A Random Forest was trained, and its performance is summarized in two diagrams.

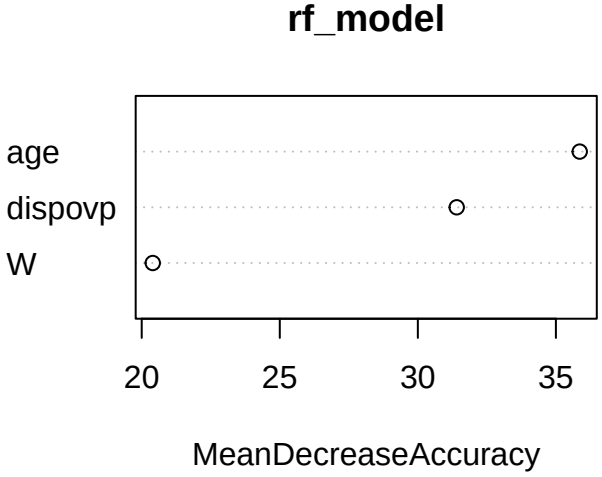
1. Error Curve The first diagram illustrates how the model's error changes with the number of trees.

The overall error (black line) decreases quickly as the first 50–100 trees are added and then stabilizes, indicating that the model combines enough trees to make robust predictions. Adding more trees provides little benefit and does not lead to overfitting,

which is a key strength of Random Forests. Class-specific errors (red and green lines) reveal that the model performs well for one class (e.g., “mobile”) but struggles with the other (e.g., “immobile”). This discrepancy is likely due to class imbalance in the dataset, where the underrepresented class is harder to predict accurately. 2. Variable Importance The second diagram shows the importance of each variable, measured by MeanDecreaseAccuracy, which indicates how much accuracy is lost when a variable is excluded.

Age (age) is the most influential predictor, aligning with its dominance in Decision Trees. dispovp and W have smaller impacts, with dispovp gaining importance due to the Random Forest’s ability to use diverse subsets of variables. Key Insights Decision Trees prioritize the most dominant variables (age), often ignoring others like dispovp. Random Forests, however, distribute importance across variables because of random feature selection during splits. Random Forests are more robust against overfitting and better leverage variable interactions. However, in imbalanced datasets, their performance on minority classes can suffer compared to Decision Trees, which may focus excessively on the majority class.





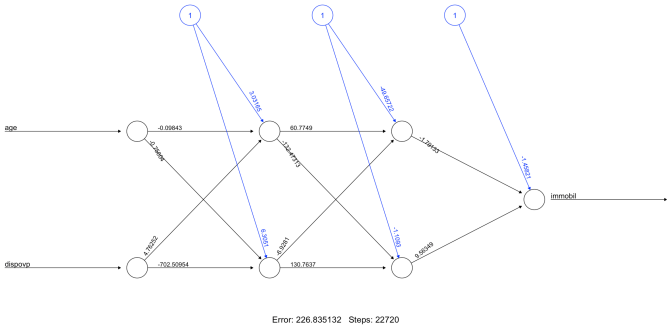
5.4 Clustering

Experiments were made with K-means clustering to see if it could provide us with meaningful groups for individuals based on the variables of interest. Two seperate set of clusters were made with different sets of variables. Clustering 1 used age_group, cspgroup, W, OCCU1, travdom and dispovp. Clustering 2 used age_group, dispovp and W. Categorical variables are converted to numeric since K-means clustering only works for numerical variables. We chose to give missing values the mean of the column and

lastly we normalize the numerical variables for better results. Elbow plots and experimentation with different numbers of clusters was used and 3 clusters were chosen as a best fit for this data based on the elbow plots and exploration of cluster evaluations.

5.5 Neural network

A neural network (NN) was utilized to see how well it could predict the mobility status of individuals compared to other models. Three seperate NN´s were made with different numbers of variables utilized. In NN1 only age and dispovp are used as numeric variables. NN2 uses age, dispovp, W, cspgroup, OCCU1 and travdom. NN3 uses age_group, dispovp and W. All of them were trying to predict the value of immobil as a number between 0-1 as immobil was changed to a numeric variable in the training. When predictions are made a threshold of 0,5 is used to determine if the prediction falls into 0 or 1 for immobility. All networks used a structure with 2 nodes and 2 hidden layers, linear.output is set to false as we are dealing with a classification problem, training and test data was split 70/30, a stepmax of 100.000 was used and all networkes were able to converge. NA values were omitted were necessary and one-hot encoding was used for the categorical variables. Further experiments were made by trying to omitt less of NA values by either creating new categories for such variables or setting them to the average value but these experiments did not improve our model. Oversampling was also tested to try to balance the data but it did not seem to improve the model either. The model that uses only age and dispovp had the best performance in our experiments.



6 Results

	Model	Accuracy	Precision	Recall	F_score
1	Logit Model	0.884	0.525	0.432	0.474
2	Decision tree	0.900	0.450	0.870	0.590
3	Random Forest	0.881	0.725	0.221	0.339
4	Neural Network1	0.900	0.926	0.384	0.543
5	Neural Network2	0.954	0.700	0.233	0.350