Classification of Imbalanced Travel Mode Choice to Work Data Using Adjustable SVM Model

Yufeng Qian 1,*, Mahdi Aghaabbasi 2,*, Mujahid Ali 3,*, Muwaffaq Alqurashi 4, Bashir Salah 5, Rosilawati Zainol 2, Mehdi Moeinaddini 6 and Enas E. Hussein 7

Citation: Qian, Y.; Aghaabbasi, M.; Ali, M.; Alqurashi, M.; Salah, B.; Zainol, R.; Moeinaddini, M.; Hussein, E.E. Classification of Imbalanced Travel Mode Choice to Work Data Using Adjustable SVM Model. Appl. Sci. 2021, 11, 11916. https://doi.org/10.3390/app112411916

Abstract: The investigation of travel mode choice is an essential task in transport planning and policymaking for predicting travel demands. Typically, mode choice datasets are imbalanced and learning from such datasets is challenging. This study deals with imbalanced mode choice data by developing an algorithm (SVMAK) based on a support vector machine model and the theory of adjusting kernel scaling. The kernel function’s choice was evaluated by applying the likelihood-ratio chi-square and weighting measures. The empirical assessment was performed on the 2017 National Household Travel Survey–California dataset. The performance of the SVMAK model was compared with several other models, including neural networks, XGBoost, Bayesian Network, standard support vector machine model, and some SVM-based models that were previously developed to handle the imbalanced datasets. The SVMAK model outperformed these models, and in some cases improved the accuracy of the minority class classification. For the majority class, the accuracy improvement was substantial. This algorithm can be applied to other tasks in the transport planning domain that deal with uneven data distribution.

Keywords: imbalanced data; travel mode choice data; hybrid support vector machine-based model

1. Introduction

A considerable amount of people’s daily trips is associated with their work. Transport planners and engineers attempt to discover work-related travel behaviors and establish strategies for reducing the adverse impacts of motorized transport on traffic, health, and the environment. One of these important behaviors is work mode choice which refers to the process where an individual chooses a certain mode for his/her trip to work. According to the literature, a variety of factors influence the work mode choice. Socioeconomic factors [1–3], household attributes e.g., [4], trip characteristics e.g., [5], job e.g., [6–9], and built environment [2,10,11] are some of these factors [3,11].

Mode choice data include a wide range of variables and samples. Typically, these data are complex and incomplete [12]. Furthermore, since motorized transport is dominant in most parts of the world, the travel surveys yield unbalanced mode choice classes; that is, there are more people who use cars than people who use other commute modes.
To date, many studies investigated the choice of travel mode to work (Table 1). These methods employed both traditional statistical methods and machine learning (ML) techniques. However, the former is criticized because of its linearity assumptions concerning mode choice data [13–15]. Thus, the employment of ML techniques has received more attention recently [16–23]. The classification of the new cases established concerning the existing samples is an essential task in ML models. If at least one of the categories comprises a smaller number of samples than other categories, the classification process becomes complex [24]. The class imbalance issue is simply an uneven data distribution amongst the different categories of the target. The precision of the classification algorithms will be unreliable when they are influenced by the majority class. In this case, the new samples are distributed to the majority category since the classification model tends to predict the minority category with less accuracy, which is an undesirable consequence [25].

Support vector machine (SVM) is a renowned ML technique for classification [26]. This algorithm also was used as a base to cope with imbalanced data. Batuwita and Palade [27] developed the fuzzy SVM model and dealt with imbalanced data in the presence of noises and outliers. Wang and Japkowicz [28] suggested boosting-SVMs with asymmetric cost. Their model runs by adjusting the classifier utilizing cost assignment, though it compensates the bias presented with adjustment through utilizing a combination system that is comparable, in effect, to adjust the distribution of data. Wu and Chang [29] suggested the class-boundary alignment algorithm to augment the SVM model to deal with imbalanced data. They modified the class boundary by converting the kernel function if data is represented in a vector space. This modification can also be performed by adjusting the kernel matrix if the data do not possess a vector-space representation. To enhance forecast performance, Liu, et al. [30] suggested consolidating an integrated sampling system, which mixes both over-sampling and under-sampling, with an ensemble of SVM. These studies investigated the binary classification problem based on the SVM model; however, less examinations have been done concerning multiclass imbalanced classification based on this model.

Many studies in other domains, including medicine, economy, crash severity, and so on, tried to reduce the issues of imbalanced data, e.g., [31–33]. However, to the best of the authors’ knowledge, a very small number of studies have investigated the issues of imbalanced travel mode choice data and proposed a solution for it [33].

So far, many scholars have provided useful strategies to manage the issue of class imbalance. These strategies have been helpful and competent in explaining the issue partially through enhancing classifiers’ performance. The majority of models developed for binary category imbalance issues are improper for the multiclass imbalanced datasets like work mode choice. In addition, rare studies have provided a solution for the imbalanced mode choice datasets. The shortcomings mentioned above prompted the authors to cope with the multiclass imbalance mode choice data issue and contribute to the body of research on this topic. Thus, this study developed the adjustable kernel-based SVM classification algorithm (SVM\textsubscript{AK}) that is suitable for handling multiclass imbalanced data. Initially, the estimated hyperplane is obtained employing the regular SVM model. Subsequently, the parameter function and the weighting factor concerning each support vector in every iteration is determined. The likelihood-ratio chi-square test is utilized to estimate these parameters. Following this, the kernel transformation or the new kernel functions are determined. The unequal boundaries of class are enlarged, and data skewness is adjusted, thanks to this function of kernel conversion. Consequently, the estimated hyperplane is remedied through the developed model, and it also solves the problem of performance degradation.

The rest of this paper is structured as follows. In Section 2, the source of data, dataset characteristics, and methodology used for improving the performance of the SVM model for classification of imbalanced data are presented, and evaluation metrics are provided. Section 3 presents the results obtained with the model as well as a series of comparisons against other ML models and SVM-based models for classifying imbalanced datasets.
Section 4 describes the sensitivity analysis method and its outcomes. Finally, a conclusion of the paper is presented in Section 5.

Table 1. Some investigation on work mode choice.

| Author                    | Main Factors Used                                      | Modelling Method                        |
|---------------------------|--------------------------------------------------------|----------------------------------------|
| Lu and Kawamura [34]      | Mode preferences and responsiveness to level-of-service| Multinomial logit                       |
| Badoe [4]                 | Households (two-workers)                               | Multinomial logit                       |
| Xie et al. [35]           | Sociodemographic and Level-of-service attributes       | Multinomial logit, Decision trees (DT), and Neural Networks (NN) |
| Patterson et al. [36]     | Gender                                                 | Multinomial logit                       |
| Al-Ahmadi [37]            | Cultural, socioeconomic, safety, and religious parameters | Disaggregate models and utility maximization |
| Gang [1]                  | Socioeconomic                                          | Multinomial logit                       |
| Vega and Reynolds-Feighan [5] | Travel time, travel cost, and employment destinations | Binary logit and GIS                   |
| Vega and Reynolds-Feighan [38] | Central and non-central and suburban employment patterns. | GIS and Cross-Nested Logit (CNL)       |
| Day, Habib and Miller [6] | Commuter trip timing, occupation groups, labor rates, work hour rules, free parking availability, and the spatial distribution of work locations | Multinomial logit                       |
| Habib [7]                 | Work start time and work duration                      | Multinomial logit                       |
| Heinen, Maat and van Wee [8] | Office culture and colleagues’ and employers’ attitudes | Binary logit                            |
| Hamre and Buehler [9]     | free car parking, public transportation benefits, showers/lockers, and bike parking at work | Multinomial logit                       |
| Heinen and Bohte [39]     | Attitudes Toward Mode Choice                           | Multinomial logit                       |
| Tran, Zhang, Chikaraishi and Fujiwara [10] | Neighborhood and travel preferences, land use policy, land use diversity and population density | Multinomial logit                       |
| Kunhikrishnan and Srinivasan [40] | Contextual heterogeneity             | Binary logit                            |
| Franco [41]               | Downtown parking supply                                | Spatial general equilibrium model        |
| Simons, De Bourdeaudhuij, Clarys, De Geus, Vandelanotte, Van Cauwenberg and Deforche [2] | Gender, socio-economic-status (SES) and living environment (urban vs. rural) | Zero-inflated negative binomial (ZINB) regression |
| Indriany et al. [42]      | Risk and uncertainty                                  | Binomial logit                          |
| Irfan et al. [43]         | Econometric Modeling                                   | Multinomial logit                       |
| Hatamzadeh et al. [44]    | Gender                                                 | Binary logit                            |
2. Methods and Data

2.1. Data

This study employed the 2017 National Household Travel Survey (NHTS)–California dataset. These data are provided by the US Federal Highway Administration and the California Department of Transportation and are freely available to all researchers and practitioners [45]. The NHTS is the definitive source on public travel behavior in the United States. It is the only national source of data that lets researchers and practitioners look at patterns in personal and household travel. This data comprises non-commercial travel information by all modes on a daily basis and the characteristics of the people who travel, their households, and their transport means. It appeared that 26,095 household samples of California were involved in this dataset. This research eliminated records that contained incomplete or inaccurate data. Additionally, the dataset included 458 variables. Thus, based on the literature, the authors selected those variables that linked to work mode choice. Finally, the dataset included 151,597 samples (based on the individuals’ records), 26 inputs, and one target variable (mode choice to work). However, at the same time, it was found that the target variable included uneven distribution of classes, which is called imbalanced data. Table 2 shows the composition of the dataset used in this study. The work mode choice had nine classes, and as expected, “car” is the majority class. The imbalance ratio is large (777.5). A list of variables used in this study is provided in Table 3.

### Table 2. Description data.

| 2017 National Household Travel Survey–California Dataset |
|----------------------------------------------------------|
| Total sample size | 151,597 |
| Number of inputs | 26 |
| Number of classes | 9 |
| Number of samples in each class | |
| Walk | 345 |
| Bicycle | 2091 |
| Car | 108,885 |
| SUV | 12,582 |
| Van | 2998 |
| Pickup truck | 22,188 |
| Motorcycle/moped | 14 |
| Public or commuter bus | 2454 |
| Private/charter/tour/shuttle bus | 40 |
| Imbalance ratio | 777.5 |

### Table 3. Variables included in the final dataset.

| Variable       | Description                  | Type   | Mean/Mode |
|----------------|------------------------------|--------|-----------|
| age            | Age of respondent            | Continuous | 36.947   |
| educ           | Highest education level      | Categorial | 3.00     |
| gtjblwk        | More than one job            | Flag   | 2.00      |
| flextime       | Flexibility of work start time | Flag   | 2.00      |
| race           | Race                         | Nominal | 1.00     |
| sex            | Gender                       | Flag   | 1.00     |
| wkftpt         | Full-time or part-time worker | Flag   | 1.00     |
Table 3. Cont.

| Variable         | Description                                                                 | Type     | Mean/Mode |
|------------------|-----------------------------------------------------------------------------|----------|-----------|
| Household information |                                                                             |          |           |
| drvrcnt          | Number of drivers in household                                              | Continuous | 3.003     |
| hh_ontd          | Number of household members on trip including respondent                    | Continuous | 1.884     |
| hhfaminc         | Household income                                                            | Categorial | 7.00      |
| hhsize           | Count of household members                                                  | Continuous | 4.576     |
| hhvehcnt         | Count of household vehicles                                                 | Continuous | 3.647     |
| numadult         | Count of adult household members at least 18 years old                      | Continuous | 3.194     |
| vehowned         | Owned vehicle longer than a year                                            | Flag     | 1.00      |
| wrkcount         | Number of workers in household                                              | Continuous | 0.25      |
| Trip characteristics |                                                                             |          |           |
| bikemore         | Reason for not biking                                                       | Nominal  | 7.00      |
| timetowk         | Trip time to work in                                                        | Continuous | 22.97    |
| walkmore         | Reason for not walking                                                      | Nominal  | 5.00      |
| Health           |                                                                             |          |           |
| health           | Opinion of health                                                           | Categorial | 2.00     |
| medcond          | Any condition or handicap that makes it difficult to travel outside of the home? | Flag     | 2.00      |
| medcond6         | Medical condition, how long?                                                | Categorial | 2.00     |
| Living environment |                                                                             |          |           |
| urbansize        | Urban area size where home address is located                               | Categorial | 6.00     |
| urbrur           | Household in urban/rural area                                                | Categorial | 1.00     |
| homeloc          | Home location/reason to choose your current home location?                  | Nominal  | 1.00      |
| wrktrans*        | Mode to work                                                                | Nominal  | 3.00      |

* Target variable.

2.2. Proposed Approach

The algorithm proposed in this study aims at dealing with the imbalanced mode choice data effectively. The theory of adjusting the kernel scaling method [46] is behind the model developed in this research to manage the multi-category imbalanced data. This study combines the SVM classification algorithm with the adjusting kernel scaling technique, which is named SVM\_AK.

2.2.1. Standard Support Vector Machine Model

Support Vector Machine (SVM) is a broadly employed and praised ML technique for classifying data [26]. The principal purpose of creating this model was to draw the input data into high dimensional space with the aid of the kernel function in such a way that the categories can be linearly divisible [47–49]. For the binary class issue, the greatest boundary that can divide the hyperplanes is as follows:

\[ w \cdot a + l = 0 \]  

(1)

The decision function for SVM based on the optimal pair \((w_0, l_0)\) is expressed by:

\[ f(a) = \sum_{i \in SV} \lambda_i y_i \langle a, a_i \rangle + l \]  

(2)

where, \(\lambda_i\) stands for support vector, \(a_i\) denotes data sample and \(i = 1, 2, \ldots, K\). Concerning greater dimensional feature space, the value of \(\langle a, a_i \rangle\) is substituted by the kernel function \(Q \langle a, a_i \rangle\), that is:

\[ Q \langle a, a_i \rangle = \langle a, a_i \rangle \]  

(3)
From the regular SVM, the kernel function was selected for estimating the boundaries’ space. In the beginning, the dataset $S$ is divided into different samples, which are $S^1$, $S^2$, $S^3$, ..., $S^i$, and subsequently, the kernel transformation function is implemented (Equation (4)).

$$f(a) = \begin{cases} 
  e^{-z_1 b(a)^2}, & \text{if } a \in S^1 \\
  e^{-z_2 b(a)^2}, & \text{if } a \in S^2 \\
  \vdots \\
  e^{-z_K b(a)^2}, & \text{if } a \in S^K 
\end{cases}$$

where, $h_a = \sum_{i \in SV} \lambda_i y_i (a, a_i) + l$ (where, $\lambda_i$ signifies support vector), $S^i$ refers to the $i$th sample of the training dataset, $z_i$ is calculated using likelihood-ratio chi-square, which is described in the following sections.

2.2.2. Likelihood-Ratio Chi-Square

Likelihood-ratio chi-square ($G^2$) is a renowned non-parametric test which assesses the target-input independence and is suitable for categorial attributes. $G^2$ ascertains a frequency distribution-based relationship among the categorial attribute assortments. To put it another way, it can be said that this technique should be employed to assess the association between the groups. The importance of determining the $G^2$ is to ascertain the connection amongst the samples of every class and parameter $z_i$. Equation (5) presents the analytical formulation for estimating $G^2$.

$$G^2 = 2 \sum D_o \log \left( \frac{D_o}{D_e} \right)$$

where, $D_o$ and $D_e$ signify observed and expected frequencies, respectively.

2.2.3. Calculating the Factor of Weighting

Ascertaining the factor of weighting is a challenging and vital task while handling an imbalanced category because finding a suitable weight is comparatively complicated. A practical technique to manage such issues is to assign smaller weight to the mainstream category and larger weight to the minority category through fulfilling the weight condition $z_i \in (0, 1)$. For dealing with the multi-category imbalance issue in the SVM$_{AK}$ algorithm, this study employed Equation (6).

$$w_i = \frac{N}{n_i \sum_{i=1}^{K} n_i}$$

where, $C$ and $N$ express class and training sample sizes, respectively. $n_i$ symbolizes the size of each class when $i = 1, 2, ..., K$.

For computing the parameter $z_i$, let $S$ denotes the dataset that comprises the $N$ number of samples and $K$ classes. The $z_i$ parameter is estimated employing Equations (2) and (3). The $G^2$ value in optimal distribution can be calculated as follows:

$$G^2 = 2 \sum_{i=1}^{K} n_i \log \left( \frac{n_i}{N/K} \right)$$

Let $X_i = n_i \log \left( \frac{n_i}{N/K} \right)$.

Then,

$$G^2 = 2 \sum_{i=1}^{K} A_i$$
Hence, the parameter \( Z_i \) can be characterized as
\[
Z_i = w_i \times \frac{A_i}{G^2}
\]  
(9)

In Equation (8), place the \( G^2 \) value
\[
Z_i = w_i \times \frac{A_i}{2 \sum_{i=1}^{K} A_i}
\]  
(10)

where, \( n_i \) is the size of the sample in the \( i \)th class and \( I = 1, 2, \ldots, K \).

2.3. Model Development Steps

In the beginning, the NHTS data was prepared and cleansed. Later, these data were used for achieving the primary partition. Then, the authors determined the weighting factor \((w_i)\) value as well as \( Z_i \) parameters for all support vectors in every iteration. The \( Z_i \) value was estimated using the likelihood-ratio chi-square test. The kernel conversion function was estimated in the next step. Eventually, utilizing the newly estimated kernel matrix \( K_{mt} \), the model was retrained. Figure 1 indicates the flowchart of the suggested algorithm.

![Algorithm Flowchart](image)

**Figure 1.** Developed algorithm flowchart.

2.4. Evaluation Metrics

This study employed four evaluation criteria to evaluate the performance of the models developed in this study. These criteria included accuracy, precision, recall, and F1 score. The formulas for calculating these criteria are shown in Equations (11)–(14). Accuracy refers to the ratio of the precisely forecasted class across the whole experiment class. Precision indicates the proportion of the true positive class over the whole number of an actual positive and false-positive category. Recall refers to the quantity of forecasted
positive categories that fall out of whole positive cases in the data. F1 score shows the equilibrium between recall and precision.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
\text{F1 score} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}
\]

where, \(TP\), \(TN\), \(FP\), and \(FN\) denote the number of true positives, true negatives, false positives, and false negatives, respectively.

3. Models’ Development and Evaluation

The authors of this study present the results of the SVM\(_{AK}\) model as well as other classification models, including the standard SVM model, BN, ANN model, and some SVM-based models in literature proposed for handling the imbalance data.

It is a challenging task to determine the most suitable classification model for handling imbalance data issues. The travel mode choice dataset was taken for the empirical investigation. Figure 2 shows the nine classes plotted on the \(x\)-axis, and the size of samples in each class plotted on the \(y\)-axis. As can be seen in this figure, it is obvious that the NHTS dataset includes uneven category distribution; practically, it is called imbalance. Hence, it grows to be more complicated to manage such a circumstance through regular classification techniques. The category-wise distributions of the dataset based on the sample size are: The car category contains 108,885 samples, the pickup truck category consists of 22,188 samples, the SUV class contains 12,582 samples, the van category consists of 2998 samples, the public or commuter bus class includes 2454 samples, the bicycle class category comprises 2091 samples, walk contains 345 samples, the private/charter/tour/shuttle bus category consists of 40 samples, and finally the motorcycle/moped category holds 14 samples. The class imbalance ratio of the dataset is 777.5.

![Figure 2. Category wise distribution of NHTS dataset.](image)
The principal intention of this study is to discover the most competent classification technique that can examine the imbalanced class. Periodically, various scholars had proposed useful techniques to handle class imbalance issue. The majority of the proposed methods were for binary category imbalance issues, which were not fit for the multi-category imbalance issue. These shortcomings urged authors to adjust the algorithm which can effectively handle binary category and multi-category imbalance issues without jeopardizing the performance of algorithms. The classification mentioned above will further assist in attaining the desirable answer toward urban transport planning and prediction of travel mode choice. This study employed four renowned conventional classification techniques and SVM-based models proposed in the literature along with the SVM$_{AK}$ model suggested in this study for the empirical assessment. The model applied in this study was compared with other models to ascertain effectiveness, fitness, and precision. This study evaluated the performance of the models developed using six criteria. Moreover, the authors employed a 10-fold cross-validation procedure as the validation scheme.

Four criteria, including accuracy, F1 score, precision, and recall were used to evaluate the outcomes of the classification algorithms applied in this study. The authors validated the classification techniques using the accuracy of classification. As is known, the NHTS dataset includes imbalanced category distribution, which may influence the performance of classification techniques. The overall performance of the models developed in this part of the study is shown in Table 4. All models achieved an overall accuracy above 80%. However, the SVM$_{AK}$ outperformed other models. The worst model was BN. Regarding other evaluation criteria, SVM$_{AK}$ again had the best values. It is worth mentioning that the SVM$_{AK}$ improved the performance of the SVM$_S$ model, which shows the capability of the proposed model of this study to handle the imbalanced mode choice data and enhance the performance of the typical SVM model for dealing with such data.

Table 4. Models’ overall performance.

| Model  | Accuracy (%) | Precision | Recall | F1 Score |
|--------|--------------|-----------|--------|----------|
| SVM$_{AK}$ | 99.81 | 0.99 | 1.00 | 0.99 |
| SVM$_S$ | 93.18 | 0.87 | 0.89 | 0.88 |
| XGBoost | 85.4 | 0.20 | 0.28 | 0.21 |
| NN | 83.06 | 0.37 | 0.43 | 0.39 |
| BN | 80.54 | 0.71 | 0.86 | 0.77 |

An evaluation of the models developed by each class also is provided in Figure 3. For the class of car, which had the largest sample size, the SVM$_{AK}$ improved the prediction accuracy from 82.33% (BN model) to 99.81%. Concerning the category of motorcycle/moped, which had the smallest sample size, the models developed yielded almost a similar accuracy. For other classes, the SVM$_{AK}$ model almost achieved better accuracy.

As previously mentioned, the performance of the SVM$_{AK}$ model was compared with some existing SVM models, which tried to alleviate the severe effects of using imbalanced data. In these methods the SVM model was hybridized with some techniques, including boosting [28], fuzzy [27], and class-boundary alignment [29], and ensemble [30]. The outcomes of the mentioned comparison are presented in Table 5. As can be seen, the SVM$_{AK}$ obtained the highest overall accuracy among all models developed.
Appl. Sci. 2021, 11, 11916

Figure 3. Models’ performance by each class: (a) accuracy (%); (b) precision; (c) recall; (d) F1 score; CLASS 1 = Walk; CLASS 2 = Bicycle; CLASS 3 = Car; CLASS 4 = SUV; CLASS 5 = Van; CLASS 6 = Pickup truck; CLASS 7 = Motorcycle/moped; CLASS 8 = Public or commuter bus; CLASS 9 = Private/charter/tour/shuttle bus.

Table 5. The overall accuracy of SVM-based models and the SVM$_{AK}$ model.

| Model                        | Accuracy (%) |
|------------------------------|--------------|
| Wang and Japkowicz [28]      | 83.33        |
| Batuwita and Palade [27]     | 92.99        |
| Wu and Chang [29]            | 93.89        |
| Liu, Yu, Huang and An [30]   | 90.16        |
| SVM$_{AK}$                   | 99.81        |

4. Sensitivity Analysis

Many factors impact the travel mode choice; however, their effects are not the same. Thus, it is necessary to ascertain the magnitude of these impacts and identify the most influential factors on travel mode choice. For this purpose, the authors employed the mutual information (MI) test method [50], which computes the importance of the inputs. MI means a filtering system that captures the random association between inputs and the target. MI examines the dependence among variables and confirms the strength of the connection among them. The MI size among inputs is measured employing the information gain:

$$\text{Gain}(C,D) = \text{Ent}(C) - \sum_{h=1}^{H} \frac{|C^h|}{|C|} \text{Ent}(C^h)$$

(15)

where, $h$ denotes the number of all probable values of $D$, $C^h$ is the set of $C$ when $D$ takes the value $D_h$, and $\text{Ent}(C)$ signifies the information entropy. The larger the value of $\text{Gain}(C,D)$, the better the relationship between $D$ and $C$.

Ultimately, the importance magnitude of each attribute for predicting travel mode choice was achieved based on the scores obtained in the MI test. The outcomes of this analysis are shown in Figure 4. The most important attributes were reason for not walking (walkmore), number of drivers in household (drvcnet), and count of adult household...
members at least 18 years old (numadlt). On the other hands, the lowest scores belonged to flexibility of work start time (flextime), owned vehicle longer than a year (vehowned), and gender (r_sex).

![Weighted importance of travel mode choice factors](image)

**Figure 4.** Outcomes of sensitivity analysis.

Reasons for not walking among respondents included unsafe street crossings, heavy traffic, and insufficient night lighting. It is clear that any improvement in these street conditions can encourage people to shift from motorized transport to walking. Thus, it makes sense that this factor is among the most influential travel mode choice factors to work [51–53]. The significance of the number of drivers in households can be attributed to its influence on the usage of vehicles and the generation of more trips. In practice, the likelihood of choosing active transportation options reduces as the number of drivers in a family grows [54]. As the number of adults in a family increases, the need for independent trips rises. Because of the different responsibilities that each adult in the family has, it is not easy to consolidate trips into one trip. This can be easily one of the principal sources of more trip generation and use of motorized transportation.

The flexibility of work start time was among the least important factors. This could be attributable to work culture of the respondents in the US. However, several previous studies showed that the flexibility of work start time influences the mode choice e.g., [54,55]. The possession of a vehicle for longer than a year was also an unimportant factor for predicting the choice of travel mode to work. A possible reason for this is that people usually look for flexible and convenient travel options to work. Usually, they are reluctant to replace their private cars with healthy travel modes unless they face new challenges. These challenges can be health problems, heavy traffic, so on. Thus, it is sensible that this factor does not influence the mode choice substantially.

### 5. Conclusions

In this research, the authors offered a novel method for learning from imbalanced mode choice data by the adjustable kernel based SVM classification model (SVM_{AK}). The likelihood-ratio chi-square test and weighting measures were used in this suggested method for selecting the kernel function. The aforementioned kernel transformation function makes it possible to increase the class limits and offset the irregular class limits. The authors also performed a sensitivity analysis which showed that the reason for not walking (walkmore), the number of drivers in the household (drvcnet), and the count of adult household members no less than 18 years old (numadlt) were the most influential...
factors. On the other hand, the lowest scores found for flexibility of work start time (flextime), owning a vehicle longer than a year (vehowned), and gender (r_sex) were the most influential factors on travel mode choice.

The outcomes of this model were compared with those of various SVM-based models and ML models. The authors employed four criteria, including accuracy, F1 score, precision, and recall to perform this comparison. The results of this study showed that the SVM\textsubscript{AK} model achieved the best results and outperformed other models. The results also showed that this model improved the classification accuracy of most categories, especially the car class that had the largest samples.

Prediction of travel mode choice is an essential component of transport planning and traffic engineering. An accurate prediction is viable only if the data are precisely classified. Therefore, precise mode choice classification utilizing the algorithm suggested in this study would be efficacious for enhancing the current transport systems and further boosting the capacities for an efficient response to the worst traffic and transport scenarios.

The classes of choice of travel mode to work in the NHTS dataset are distinct from other mode choice datasets since this dataset considered the “SUV” category as different from the “car” category. However, future studies could combine these two classes and create a new class of travel mode to work. In the US, private motorized transport is dominant. Thus, the NHTS and, in turn, the result of this study is affected by this issue. Future studies can employ the method developed in this study to predict choice of travel mode to work in different environments, such as those in which walking, cycling, and public transport are dominant.

**Author Contributions:** Conceptualization, M.A. (Mahdi Aghaabbasi); methodology, M.A. (Mahdi Aghaabbasi); software, M.A. (Mahdi Aghaabbasi); formal analysis, M.A. (Mahdi Aghaabbasi); investigation, M.A. (Mahdi Aghaabbasi); resources, Y.Q., M.A. (Muwaffaq Alqurashi), M.A. (Mujahid Ali), B.S., E.E.H.; data curation, M.A. (Mahdi Aghaabbasi); writing—original draft preparation, M.A. (Mahdi Aghaabbasi) and M.M.; writing—review and editing, M.A. (Mahdi Aghaabbasi) and M.A. (Mujahid Ali); supervision, R.Z.; funding acquisition, Y.Q., M.A. (Muwaffaq Alqurashi), M.A. (Mujahid Ali), B.S., E.E.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was funded by the King Saud University, Saudi Arabia, through researchers supporting project number (RSP-2021/145), and the Taif University, Saudi Arabia, through Taif University Researchers Supporting Project, grant number [TURSP-2020/324].

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Acknowledgments:** The authors of this paper thank King Saud University, Saudi Arabia for funding this work through researchers supporting project numbers (RSP-2021/145). The authors also would like to acknowledge Taif University Researches Supporting Project number (TURSP-2020/324), Taif University, Taif, Saudi Arabia for supporting this work.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Gang, L. A behavioral model of work-trip mode choice in Shanghai. *China Econ. Rev.* **2007**, *18*, 456–476.
2. Simons, D.; De Bourdeaudhuij, I.; Clarys, P.; De Geus, B.; Vandelanotte, C.; Van Cauwenberg, J.; Deforce, B. Choice of transport mode in emerging adulthood: Differences between secondary school students, studying young adults and working young adults and relations with gender, SES and living environment. *Transp. Res. Part A Policy Pract.* **2017**, *103*, 172–184. [CrossRef]
3. Ali, M.; de Azevedo, A.R.G.; Marvila, M.T.; Khan, M.I.; Memon, A.M.; Masood, F.; Almahbashi, N.M.Y.; Shad, M.K.; Khan, M.A.; Fediuk, R.; et al. The Influence of COVID-19-Induced Daily Activities on Health Parameters—A Case Study in Malaysia. *Sustainability* **2021**, *13*, 7465. [CrossRef]
4. Badoe, D. Modelling work-trip mode choice decisions in two-worker households. *Transp. Plan. Technol.* **2002**, *25*, 49–73. [CrossRef]
5. Vega, A.; Reynolds-Feighan, A. Employment sub-centres and travel-to-work mode choice in the Dublin region. *Urban Stud.* **2008**, *45*, 1747–1768. [CrossRef]
6. Day, N.; Habib, K.N.; Miller, E.J. Analysis of work trip timing and mode choice in the Greater Toronto Area. *Can. J. Civ. Eng.* **2010**, *37*, 695–705. [CrossRef]
7. Habib, K.B.M.N. Modeling commuting mode choice jointly with work start time and work duration. *Transp. Res. Part A Policy Pract.* **2012**, *46*, 33–47. [CrossRef]
8. Heinen, E.; Maat, K.; van Wee, B. The effect of work-related factors on the bicycle commute mode choice in the Netherlands. *Transportation* **2013**, *40*, 23–43. [CrossRef]
9. Hamre, A.; Buehler, R. Commuter mode choice and free car parking, public transportation benefits, showers/lockers, and bike parking at work: Evidence from the Washington, DC region. *J. Public Transp.* **2014**, *17*, 4. [CrossRef]
10. Tran, M.T.; Zhang, J.; Chikaraishi, M.; Fujiwara, A. A joint analysis of residential location, work location and commuting mode choices in Hanoi, Vietnam. *J. Transp. Geogr.* **2016**, *54*, 181–193. [CrossRef]
11. Ali, M.; Dharmowijoyo, D.B.; Harahap, I.S.; Puri, A.; Tanjung, L.E. Travel behaviour and health: Interaction of Activity-Travel Pattern, Travel Parameter and Physical Intensity. *Solid State Technol.* **2020**, *63*, 4026–4039.
12. Aghaabbasi, M.; Shekari, Z.A.; Shah, M.Z.; Olakunle, O.; Armaghani, D.J.; Moeinaddini, M. Predicting the use frequency of ride-sourcing by off-campus university students through random forest and Bayesian network techniques. *Transp. Res. Part A Policy Pract.* **2020**, *136*, 262–281. [CrossRef]
13. Stylianou, K.; Dimitriou, L.; Abdel-Aty, M. Big data and road safety: A comprehensive review. *Mobil. Patterns Big Data Transp. Anal.* **2019**, *297–343*. [CrossRef]
14. Rashidi, S.; Ranjitkar, P.; Hadas, Y. Modeling bus dwell time with decision tree-based methods. *Transp. Res. Rec.* **2014**, *2418*, 74–83. [CrossRef]
15. Ali, M.; Dharmowijoyo, D.B.E.; de Azvedo, A.R.G.; Fediuk, R.; Ahmad, H.; Salah, B. Time-Use and Spatio-Temporal Variables Influence on Physical Activity Intensity, Physical and Social Health of Travelers. *Sustainability* **2021**, *13*, 12226. [CrossRef]
16. Parsajoo, M.; Armaghani, D.J.; Mohammed, A.S.; Khari, M.; Jahandari, S. Tensile strength prediction of rock material using non-destructive tests: A comparative intelligent study. *Transp. Geotech.* **2021**, *31*, 100652. [CrossRef]
17. Harandizadeh, H.; Armaghani, D.J.; Asteris, P.G.; Gandomi, A.H. TBM performance prediction developing a hybrid ANFIS-PNN predictive model optimized by imperialism competitive algorithm. *Neural Comput. Appl.* **2021**, *33*, 16149–16179. [CrossRef]
18. Li, E.; Zhou, J.; Shi, X.; Armaghani, D.J.; Yu, Z.; Chen, X.; Huang, P. Developing a hybrid model of salp swarm algorithm-based support vector machine to predict the strength of fiber-reinforced cemented paste backfill. *Eng. Comput.* **2020**, *37*, 1–22. [CrossRef]
19. Jahed Armaghani, D.; Kumar, D.; Samui, P.; Hasanipanah, M.; Roy, B. A novel approach for forecasting of ground vibrations resulting from blasting: Modified particle swarm optimization coupled extreme learning machine. *Eng. Comput.* **2021**, *37*, 3221–3235. [CrossRef]
20. Armaghani, D.J.; Harandizadeh, H.; Momeni, E.; Maizir, H.; Zhou, J. An optimized system of GMDH-ANFIS predictive model by ICA for estimating pile bearing capacity. *Artif. Intell. Rev.* **2021**, *54*, 1–38. [CrossRef]
21. Li, Z.; Yazdani Bejarbaneh, B.; Asteris, P.G.; Koopialipoor, M.; Armaghani, D.J.; Tahir, M. A hybrid GEP and WOA approach to estimate the optimal penetration rate of TBM in granitic rock mass. *Soft Comput.* **2021**, *25*, 11877–11895. [CrossRef]
22. Yu, C.; Koopialipoor, M.; Murlidhar, B.R.; Mohammed, A.S.; Armaghani, D.J.; Mohamad, E.T.; Wang, Z. Optimal ELM–Harris Hawks optimization and ELM-Grasshopper optimization models to forecast peak particle velocity resulting from mine blasting. *Nat. Resour. Res.* **2021**, *30*, 2647–2662. [CrossRef]
23. Menardi, G.; Torelli, N. Training and assessing classification rules with imbalanced data. *Data Min. Knowl. Discov.* **2014**, *28*, 92–122. [CrossRef]
24. Daskalaki, S.; Kopanas, I.; Avouris, N. Evaluation of classifiers for an uneven class distribution problem. *Appl. Artif. Intell.* **2006**, *20*, 381–417. [CrossRef]
25. Vapnik, V. *The Nature of Support Vector Machine*; Springer: Berlin/Heidelberg, Germany, 1999.
26. Batuwita, R.; Palade, V. FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning. *IEEE Trans. Fuzzy Syst.* **2010**, *18*, 558–571. [CrossRef]
27. Wang, B.X.; Japkowicz, N. Boosting support vector machines for imbalanced data sets. *Knowl. Inf. Syst.* **2009**, *25*, 1–20. [CrossRef]
28. Wu, G.; Chang, E.Y. Class-Boundary Alignment for Imbalanced Dataset Learning. In *Proceedings of the Workshop Learning from Imbalanced Datasets II, Washington, DC, USA, 21 August 2003*; pp. 49–56.
29. Liu, Y.; Yu, X.; Huang, J.X.; An, A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Inf. Process. Manag.* **2011**, *47*, 617–631. [CrossRef]
30. Liu, Z.; Tang, D.; Cai, Y.; Wang, R.; Chen, F. A hybrid method based on ensemble WELM for handling multi class imbalance in cancer microarray data. *Neurocomputing* **2017**, *266*, 641–650. [CrossRef]
31. Hordri, N.F.; Yuhaniz, S.S.; Azmi, N.F.M.; Shamsuddin, S.M. Handling class imbalance in credit card fraud using resampling methods. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 390–396. [CrossRef]
32. Kim, S.; Lym, Y.; Kim, K.-J. Developing crash severity model handling class imbalance and implementing ordered nature: Focusing on elderly drivers. *Int. J. Environ. Res. Public Health* **2021**, *18*, 1966. [CrossRef]
33. Rezaei, S.; Khajandizadeh, H.; Hamed, A.M.; Brakewood, C.; Jin, M.; Cherry, C. Performance evaluation of mode choice models under balanced and imbalanced data assumptions. *Transp. Lett.* **2021**, *13*, 1–13. [CrossRef]
34. Lu, Y.; Kawamura, K. Data-mining approach to work trip mode choice analysis in Chicago, Illinois, area. *Transp. Res. Rec.* **2010**, *2156*, 73–80. [CrossRef]
35. Xie, C.; Lu, J.; Parkany, E. Work travel mode choice modeling with data mining: Decision trees and neural networks. *Transp. Res. Rec.* **2003**, *1854*, 50–61. [CrossRef]

36. Patterson, Z.; Ewing, G.; Haider, M. Gender-based analysis of work trip mode choice of commuters in suburban Montreal, Canada, with stated preference data. *Transp. Res. Rec.* **2005**, *1924*, 85–93. [CrossRef]

37. Al-Ahmadi, H. Development of intercity work mode choice model for Saudi Arabia. In *WIT Transactions on The Built Environment*; WITT Press: Southampton, UK, 2007; Volume 96.

38. Vega, A.; Reynolds-Feighan, A. A methodological framework for the study of residential location and travel-to-work mode choice under central and suburban employment destination patterns. *Transp. Res. Part A Policy Pract.* **2009**, *43*, 401–419. [CrossRef]

39. Heinen, E.; Bohte, W. Multimodal commuting to work by public transport and bicycle: Attitudes toward mode choice. *Transp. Res. Rec.* **2014**, *2468*, 111–122. [CrossRef]

40. Kunhikrishnan, P.; Srinivasan, K.K. Choice set variability and contextual heterogeneity in work trip mode choice in Chennai city. *Transp. Lett.* **2019**, *11*, 174–189. [CrossRef]

41. Franco, S.F. Downtown parking supply, work-trip mode choice and urban spatial structure. *Transp. Res. Part B Methodol.* **2017**, *101*, 107–122. [CrossRef]

42. Indriany, S.; Sjafruddin, A.; Kusumawati, A.; Weningtyas, W. Mode choice model for working trip under risk and uncertainty. In Proceedings of the AIP Conference Proceedings, Maharashtra, India, 5–6 July 2018; p. 020041.

43. Irfan, M.; Khurshid, A.N.; Khurshid, M.B.; Ali, Y.; Khattak, A. Policy implications of work-trip mode choice using econometric modeling. *J. Transp. Eng. Part A Syst.* **2018**, *144*, 04018035. [CrossRef]

44. Hatamzadeh, Y.; Habibian, M.; Khodaii, A. Walking mode choice across genders for purposes of work and shopping: A case study of an Iranian city. *Int. J. Sustain. Transp.* **2020**, *14*, 389–402. [CrossRef]

45. Transportation Secure Data Center. 2017 National Household Travel Survey—California; Transportation Secure Data Center: Golden, CO, USA, 2019.

46. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.

47. Wang, L. *Support. Vector Machines: Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005; Volume 177.

48. Foody, G.M.; Mathur, A. Toward intelligent training of supervised image classifications: Directing training data acquisition for SVM classification. *Remote Sens. Environ.* **2004**, *93*, 107–117. [CrossRef]

49. Verron, S.; Tiplica, T.; Kobi, A. Fault detection and identification with a new feature selection based on mutual information. *J. Process. Control.* **2008**, *18*, 479–490. [CrossRef]

50. Aghaabbasi, M.; Shah, M.Z.; Zainol, R. Investigating the Use of Active Transportation Modes Among University Employees Through an Advanced Decision Tree Algorithm. *Civil. Sustain. Urban. Eng.* **2021**, *1*, 26–49. [CrossRef]

51. Aghaabbasi, M.; Moeinaddini, M.; Shah, M.Z.; Asadi-Shekari, Z. A new assessment model to evaluate the microscale sidewalk design factors at the neighbourhood level. *J. Transp. Health* **2017**, *5*, 97–112. [CrossRef]

52. Tabatabaei, S.; Aghaabbasi, M.; Mahdiyar, A.; Zainol, R.; Ismail, S. Measurement Quality Appraisal Instrument for Evaluation of Walkability Assessment Tools Based on Walking Needs. *Sustainability* **2021**, *13*, 11342. [CrossRef]

53. Sultana, S. Factors Affecting Parents’ Choice of Active Transport Modes for Children’s Commute to School: Evidence from 2017 NHTS Data; The University of Toledo: Toledo, OH, USA, 2019.

54. Thorhauge, M.; Cherchi, E.; Rich, J. How flexible is flexible? Accounting for the effect of rescheduling possibilities in choice of departure time for work trips. *Transp. Res. Part A Policy Pract.* **2016**, *86*, 177–193. [CrossRef]