HEC MONTRÉAL

Perception of Canadian Parents about Their Children's Physical Health During COVID-19: a Machine Learning Approach

par

Issam Abdo-Ahmad

Decio Coviello
HEC Montréal
Codirecteur de recherche

Benoit Dostie
HEC Montréal
Codirecteur de recherche

Sciences de la gestion
(Spécialisation Économie Financière Appliquée)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Mars, 2024

Résumé

Cette étude vise à examiner les facteurs affectant la perception des parents canadiens sur la santé physique de leurs enfants pendant la pandémie de COVID-19 en utilisant un ensemble de données transversales de Statistique Canada intitulé "Impacts du COVID-19 sur le rôle parental des Canadiens pendant la pandémie". L'étude est remarquable pour son vaste ensemble de données, qui comprend plus de 24 000 observations provenant de tout le pays, ainsi que pour sa méthodologie innovante qui intègre des méthodes d'apprentissage automatique non paramétriques (arbre de décision et Gini Impurity-Based Feature Importance) avec une modélisation économétrique paramétrique (Generalized Ordered Logit with Partial Proportional Odds). Nos résultats montrent que la consommation de malbouffe par les enfants, les difficultés à maintenir des liens sociaux avec les amis et la famille, le fait de jongler entre la garde des enfants et les tâches professionnelles, l'appartenance à une minorité visible et l'intention d'utiliser des services de garde d'enfants après la pandémie sont les cinq principaux facteurs qui influencent les préoccupations des parents au sujet de la santé physique de leurs enfants au cours de l'enquête COVID-19. Étant la première à aborder cette question de recherche, notre étude fournit des informations précieuses aux décideurs politiques qui souhaitent développer des interventions ciblées pour atténuer les effets d'une éventuelle pandémie à l'avenir.

Mots clés : COVID-19 ; Santé physique des enfants ; Perception des parents ; Analyse économétrique ; Modèle logit ordonné généralisé ; Apprentissage automatique ; Arbre de décision

Méthodes de recherche : Régression logistique ordonnée généralisée avec cotes proportionnelles partielles ; arbre de décision ; importance des caractéristiques basée sur l'impureté de Gini.

Abstract

This study aims to examine the factors affecting the perception of Canadian parents about their children's physical health during the COVID-19 pandemic utilizing a cross-sectional dataset from Statistics Canada titled "Impacts of COVID-19 on Canadians' Parenting During the Pandemic". The study is noteworthy for its extensive dataset, which includes over 24,000 observations from all over the country, as well as its innovative methodology that integrates non-parametric machine learning methods (Decision Tree and Gini Impurity-Based Feature Importance) with parametric econometric modeling (Generalized Ordered Logit with Partial Proportional Odds). Our findings show that children's junk food consumption, challenges in maintaining social connections with friends and family, juggling childcare and work-related tasks, belonging to a visible minority group, and the intention to use childcare services after the pandemic, are the five main factors influencing parental concerns about their children's physical health during COVID-19. Being the first to tackle this research question, our study provides valuable insights for policymakers aiming at developing targeted interventions to lessen the effect of a possible pandemic in the future.

Keywords: COVID-19; Children's Physical Health; Parental Perception; Econometric Analysis; Generalized Ordered Logit Model; Machine Learning; Decision Tree.

Research methods: Generalized Ordered Logistic Regression with Partial Proportional Odds; Decision Tree; Gini Impurity-Based Feature Importance.

Table of contents

List of Tables

List of Figures

List of Equations

Preface

One of the three main contributions of this paper, which is the integration of Machine Learning Technique with Econometric Analysis, was originally inspired by the "Applied Econometrics and Machine Learning in Economics" course, taught by my co-supervisor, Professor Decio Coviello. The thesis topic was also discussed by Professor Benoit Dostie, an expert on Applied Econometrics and Labor Economics, and then later developed to incorporate machine learning in collaboration with my second co-supervisor Professor Decio Coviello. This work came to life because of their continuous support and guidance.

Acknowledgments

Foremost, I would like to thank my co-supervisors, Professor Decio Coviello and Professor Benoit Dostie, for their great support and patience throughout the process of writing this thesis. I would like to also thank my parents, family, and friends in Canada, especially Nathir Haimoun, Ismail Bourgi, Ghina Abdul-Baki, and Firas Alameddine whose unwavering support and presence have been a constant source of motivation. While I was writing alone, their presence proved that I wasn't on my own.

Chapter 1: Introduction

In December 2019, China reported the first case of COVID-19 in the city of Wuhan. Shortly after that, the (World Health Organization, 2020) declared the emergence of COVID-19 as a pandemic on the 11th of March 2020 whereby then all provinces and territories in Canada declared a state of public health emergency eventually leading to lockdowns (Dawson, 2020). This has disrupted the daily life of Canadians regardless of their age. For adults, many were now asked to work either partially or fully from home to limit human contact and thus control the spread of the virus. As for children, they were required to switch to online learning following the closure of schools and the termination of many recreational activities that kept children active. This has thus led to a significant drop in the physical activity of Canadian children triggering considerable increases in sedentary behavior along with disruptions to children's mental health and sleep behavior.

In a study that aimed to explore the effect of the COVID-19 pandemic on the physical activity and screen time of Canadian children residing in London, Ontario, Ostermeier et al. (2021) argue that there was a significant drop in children's physical activity. Specifically, they note that children's physical activity drops from 4.39 days per week to 3.78 days per week whereas recreational time in front of a screen went up by an hour per day. As such, many children during COVID-19 did not meet the recommended Canadian 24-hour movement guidelines[1], as also highlighted by Moore et al. (2021). They compare the movement behavior of children and youth in two different periods during COVID-19 by repeating two cross-sectional surveys in October 2020 (first wave) and in April 2021 (second wave), and they report that less than 5% of children and less than 2% of youth met the Canadian 24-hour movement guidelines during the first and second waves. Unfortunately, pre-COVID-19 statistics are not any better whereby Chaput et al. (2017) argue that only 13% of children aged 3 to 4 years old met the Canadian 24-Hour Movement Guidelines. They also note that only 17% of those aged between 5 and 17 years old met the latter guidelines. Considering the several benefits of healthy movement during childhood, such as

[1] According to the Canadian Society for Exercise Physiology (2016), the Canadian 24-hour movement guidelines for children and youth suggest at least an hour per day of physical activity of moderate to vigorous intensity. Moreover, the guidelines also suggest that physical activity that aims to strengthen bones and muscles be practiced at least 3 days per week.

improved physical activity (Janssen and Allana, 2010; Hayes et al., 2019) and improved sleeping behavior (Williamson et al., 2019), such findings about the Canadian movement guidelines are concerning.

Within this context, Szpunar et al. (2022) argue that understanding parents' perspectives about their children's physical health is important as research shows that children's activities, both structured and unstructured, are highly influenced by their parents (Trost and Loprinzi, 2011). Hence, the author hypothesizes that parents' perceptions about their children's physical activity may affect the likelihood that children improve their physical activity though improvement movement that complies with the Canadian guideline. In fact, research has found that certain parental activities have been positively linked with those of their children in some investigations. For example, Sigmundová et al. (2020) show that an increase in steps for fathers corresponds to an increase for sons, and a similar increase for mothers corresponds to an increase for daughters. This might be attributed to the finding of Moore et al. (2020) who argue that parental support and co-participation significantly improved the physical activity of children during COVID-19.

In this context, we aim to study the factors that affect parents' perception of their children's physical health during COVID-19 in Canada. To do so, we rely on a publicly available cross-sectional dataset titled "*Impacts of COVID-19 on Canadians' Parenting During the Pandemic*" published by Statistics Canada. The data was collected between June 9, 2022, and June 22, 2022, and aimed at surveying Canadian parents about changes to them and their families in the context of COVID-19 starting from March 15, 2020, to June 9, 2022.

As such, the main goal of this survey data is to examine the concerns of Canadian parents about their children's mental and physical health during COVID-19. Consequently, the data covers the different impacts of COVID-19 on children's social lives, childcare, and schooling activities. The data also shows the labor impacts caused by COVID-19 as well as a set of different parental attributes of the surveyed parents in the context of COVID-19. Hence our research question reads as follows:

What Factors Predict Parents' Perception about Their Children's Physical Health During the COVID-19 Pandemic in Canada?

Our research question is important because it recognizes the key role of parents in influencing their children's physical activity by examining the factors that affect parents' perceptions of their children's physical health. In their paper, Zecevic et al. (2010) discuss how parents shape the way children move and play showing that they have a big say in how active their children are. If parents love to be active, their kids often do too. Moreover, the authors demonstrate through their findings that children who received greater encouragement for physical activity from their parents were considerably more likely to engage in active behaviors themselves compared to children lacking such guidance and backing. Similar findings were shown by the research of Moore et al. (1991) who examined the effect of parents' physical activity levels on that of young children. The authors argue that children with moms who moved a lot – as depicted by their hourly step count being above the median - had twice the chance to be active as those with moms less keen on action. They also show a similar effect to fathers being active whereby increasing kids' chances to be active by 3.5 times.

Understanding parents' perceptions of their children's physical health is crucial for informing public health policies and interventions, particularly during health crises like the COVID-19 pandemic. By analyzing the factors that affect parents' perceptions about their children's physical health factors, we are better able to know what makes parents more concerned about their children's physical health during a pandemic which is key for similar future ones. At the time of writing this paper, concerns are mainly about two emerging Omicron subvariants known as *EG.5* and *BA.2.86* and a bacterial infection referred to as *Mycoplasma Pneumoniae.* Regarding the former Omicron subvariants, expert thoughts remain mixed and uncertain as with any new variant. This uncertainty about the new variants was evident in the talk of Health Canada's Chief Medical Advisor Dr. Supriya Sharma who said that although current clinical data are not alarming, waiting remains the only way to be sure how these variants will behave[2].

Similarly, there is a current uncertainty about whether the new bacterial pneumonia, referred to as *Mycoplasma Pneumoniae,* will cause an epidemic in Canada. The latter bacterial infection was first identified in May 2023 in China leading to outbreaks in

[2] See: https://www.ctvnews.ca/health/coronavirus/what-you-should-know-about-omicron-subvariants-eg-5-and-ba-2-86-1.6559825

countries around the world. For example, the spread of *Mycoplasma Pneumoniae* in France led to the classification of this bacterial infection as an epidemic[3]. In Quebec, the diagnosis of Dr. Donald Vinh - an expert on infectious diseases at McGill University Health Centre – reveals how uncertain we are about the possible spread of pneumonia. In a recent talk to the Montreal Gazette, Dr. Vinh argued that although there have not been many reported cases of the bacteria, we cannot assure that we are safe from a possible outbreak[4]. Moreover, Dr. Vinh argues that testing for the bacterial spread in Quebec is still limited which might be the reason why we are not observing increased cases of *Mycoplasma Pneumoniae* yet.

As such, having information on factors that influence parents' concern about their children's physical health during a pandemic is key to guiding policy makers for specific policies. For example, we show later in our paper that certain parental attributes such as being old, immigrant, and belonging to a minority or indigenous group all contribute to higher concern about children's physical health during COVID-19. Accordingly, policy makers are encouraged to develop targeted support programs for these groups of parents who were shown to be more adversely affected by the COVID-19 pandemic than their respective counterparts. These policies may include special financial aid programs and counseling services for these groups of parents.

Another set of features affecting parents' concerns are changes in children's activity and eating behavior. In this regard, we show later in the paper that children spending a lot of time in front of a screen and eating junk food both adversely affect parents' concerns about their physical health during COVID-19. About this, policy makers are encouraged to have campaigns that alert parents and children to the health risks of frequent screentime and junk food consumption. Moreover, policies that keep children active and entertained can be effective in helping children reduce screen time and improve eating habits such as having online home sports and arts classes. Additionally, having stricter regulation on advertisements of unhealthy food targeted to children can play a key role in shifting their consumption towards healthy alternatives.

[3] See: https://www.euronews.com/next/2023/12/05/walking-pneumonia-epidemics-have-been-reported-in-parts-of-europe-heres-what-you-need-to-k#:~:text=In%20France%2C%20%22unusual%20increases%22,%22reflecting%20an%20epidemic%20situation%22

[4] See: https://montrealgazette.com/news/local-news/quebec-authorities-monitoring-for-possible-mystery-illness-in-kids

A further set of variables affecting parents' concern are the labor market-related impacts of COVID-19. Concerning this, our paper shows that parents facing the challenge of balancing childcare tasks and work and those who lost their jobs or experienced a drop in working hours are more concerned about their children's physical health during COVID-19. As such, policies to alleviate these effects may include support for flexible working hours to help parents better coordinate between tasks related to childcare and work. Moreover, policy makers may want to offer workshops to parents about parenting during a pandemic and ways for better stress management during such trying times.

The last variable shown to increase parents' concern about their children's physical health is children's challenging to remain connected with family and friends during COVID-19. For that, polices that ensure children remain connected with their community during a pandemic are crucial such as facilitating virtual meetups with friends and family. Another policy to keep children connected and active would encourage socially distanced outdoor activities such as sports that are naturally socially distanced such as tennis, frisbee, kite flying, group bicycle rides, and fitness classes in the open air. Other socially distanced outdoor activities might include cleaning a public beach or gardening in a public park.

Another reason for the importance of our research question is the limited research on the subject matter, whereby up to our knowledge, no paper has yet examined our research question. The only few papers that are in the scope of our work are those by McCormack et al., (2020), Ostermeier et al. (2022), and Szpunar et al. (2022) with each having several limitations that our research tackles. For example, common limitations for the three studies include being limited by their data's sample size and geographic location whereby McCormack et al., (2020) rely on data that was collected from 345 parents of children aged between 5 and 17 years old in Calgary, Alberta, intending to examine the relationship between parents' anxiety from COVID-19 and children's physical activity and sedentary practices. Similarly, Ostermeier et al. (2022) relies on data from 27 parents of children enrolled in the Grade 5 ACT-i-Pass Program in London, Ontario, who were interviewed to study the effect of COVID-19 on their children's engagement in physical activity. Analogously, the paper Szpunar et al. (2022) relies on data from 382 parents in Ontario, Canada to explain the perspectives of parents with children aged between 0 and 12 years old regarding their physical

activity during the COVID-19 pandemic. As such, we notice that none of the latter papers rely on data with a significant number of observations spanned across the different Canadian provinces as opposed to our paper which utilizes a wider dataset with 24,956 observations spanned across the different Canadian provinces and territories. Hence, by using such data in our paper, we acknowledge the regional differences thus providing insights that are more representative of the Canadian parents in general.

The third reason for the importance of our paper is the employed methodology used to answer the research question that comes at the intersection of econometric and machine learning techniques. Specifically, our paper employs a *Generalized Ordered Logit Regression with Partial Proportional Odds* (Williams, 2016) which is a parametric econometric technique along with *Decision Tree* (Breiman, 2017), which is a non-parametric machine learning technique. As such, our approach presents a case where researchers can benefit from combining non-parametric machine learning techniques with parametric econometric modeling in several ways.

First, it gives researchers different options to investigate the relationship between the response variables and the set of predictor variables. When determining potential effects without requiring prior assumptions to be tested and validated, non-parametric models perform well. The patterns found from the exploratory analysis carried out in the non-parametric models can then be quantified and tested using a parametric model.

The second advantage of combining parametric and non-parametric approaches is that one can be the robustness check for the other. Specifically, the non-parametric approach can act as a robustness check for the parametric one as the former doesn't force any assumptions unlike the latter which does. For example, our parametric approach (i.e., the *Generalized Ordered Logit Model*) identifies certain predictor variables with significantly high effect on the response variable which is then confirmed by our non-parametric model (i.e., the Decision Tree) which uses the same variables for the first few splitting nodes such acknowledging their importance in affecting the dependent variable[5].

[5] The terms *Dependent Variable* and *Response Variable* are used interchangeably throughout the paper.

Lastly, combining parametric and non-parametric approaches widens the audience of the paper. Technical audiences with a background in econometric analysis can delve deeper into the analysis of the effects of the predictor variables[6] on the response variable. As for the non-technical audience, they can benefit from the ease of visualization and interpretability of non-parametric models like *Decision Trees*.

Yang et al. (2022) argue that such integration between Econometric and Machine Learning techniques has attracted significant interest in academic research. Specifically, the authors argue that an empirical approach may involve the utilization of machine learning techniques to mine the data for key predictor features such as utilizing what is called Feature Importance Based on Gini Impurity which ranks a given set of predicator variables based on their ability to reduce prediction error (i.e. Gini Impurity) across the entire Decision Tree.

Findings from the non-parametric approach (i.e., Decision Tree) show five key features affecting parents' concern about their children's physical health during COVID-19 after the first two splits, namely: (1) Parents' Concern About Their Children's Consumption of Junk Food, (2) Parents Concern about Limitation for their Childrens to remain connected with Friends and Family, (3) Parents Concern for Balancing between Childcare and Work Tasks at Home, (4) Parents identifying with a visible Minority group in Canada, and lastly (5) Parents willing to make use of childcare services when they open after the pandemic[7].

As for findings from the parametric model (i.e., Generalized Ordered Logistic Regression with Proportional Odds) quantifying the effect of the latter variables shown by the first two splits of the Decision Tree, we note that Parents Concerned About Their Children Consuming Junk Food during COVID-19 and those Concerned For Their Children Remaining Connected With Family & Friends were both shown to be twice as likely to be more concerned about their children's physical health. Moreover, findings show that Parents Belonging to a Visible Minority and those Concerned About Balancing Between Childcare and Work tasks are 67% and 36.6%

[6] The terms *Feature(s)*, *Explanatory Variable(s)*, and *Predictor Variable(s)* are used interchangeably throughout the paper.

[7] The criteria involve the examination of the impact of including certain features in the model and the extent to which they contribute to the reduction of Gini Impurity, with a specified threshold of 0.05 (or 5%). Hence, only factors lowering Gini Impurity by at least 0.05 where chosen. Detailed discussion on this available in the Decision Tree section.

more concerned about their children's physical health, respectively. Lastly, parents willing to make use of childcare services when they open after the pandemic were shown to be 12% less concerned about their children's physical health.

Upon thorough examination and comparison with the existing body of literature, our findings were found to align with prior research both in the presence and absence of pandemics. Consequently, we anticipate these established relationships to persist even in non-pandemic scenarios. Hence the findings of this paper are not necessarily specific to a health crisis scenario rendering our research question relevant beyond pandemic situations.

One potential problem in our model is the possibility of having Endogeneity generated from two sources namely Reverse Causality and Omitted Variable Bias (OVB) which limits the ability to derive causal inferences. A possible solution to the Reverse Causality problem would be employing a Two-Stage Model which is used to address the endogeneity problem using Instrumental Variables. However, implementing a Two-Stage model poses some challenges such as identifying valid instruments and ensuring that the exogeneity condition applies to them.

As for the second source of Endogeneity, one could include as many explanatory variables as possible in the model to lower the bias from omitted variables. However, this approach suffers the *Curse of Dimensionality* limitation whereby the more variables we add, the more the model will use these features to fit the noise in the training dataset potentially leading to overfitting. This in turn causes the model to perform very well on the training data but poorly on the out-of-sample data, thus lowering the predicative ability of the model. As such, we opt to add multiple explanatory variables while avoiding having too many of them so that we don't weaken the predictive power of our model.

Given that this study aims to find key predictors associated with parents' concern about their children's physical health, we are only interested in the predictive ability of the model instead of deriving causal relationships. In other words, our predictive analysis aims to forecast variations in the labels of the dependent variable and hence the focus is more on the accurate predictions of the model as opposed to the unbiased estimation of causal effects. Since the accuracy score of our Generalized Ordered Logistic Model is around 50%, the model demonstrates a relatively strong predictive power on out-of-

sample data, especially considering the ordinal and categorical nature of the response variable with 4 labels whereby the random guess probability is 25%.

The subsequent sections of the paper are structured as follows. Chapter 2 presents Related Literature and develops the hypotheses to be tested. Chapter 3 presents the data, the independent variable, and the selection of predictor variables to be used for the parametric and non-parametric models. We also present the summary statistics and conduct two checks for multicollinearity. Chapter 4 presents a detailed discussion of the model selection using the machine learning technique referred to as cross-validation and discusses thoroughly the used parametric and non-parametric models. Chapter 5 presents the results of the models and discusses the findings in the context of current literature on the subject matter. Chapter 6 presents the limitations of the study. Lastly, Chapter 7 concludes the paper and proposes a set of policy recommendations.

Chapter 2: Literature Review

In this section, we present the related literature on factors either directly affecting children's physical health or affecting the concern of parents about their children's physical health. By doing so, we extract several features that are then grouped under five different categories namely (1) Parental Attributes, (2) Child Activities and Eating Impacts, (3) Labor Market Impacts, (4) Social Impacts, and (5) Childcare Impacts.

The aim of establishing these categories is to use them as references to extract similar or related variables from our data to incorporate as predicator variables. The contribution of these variables is then analyzed by two models one parametric referred to as a Generalized Ordered Logistic Regression with Partial Proportional Odds and another nonparametric referred to as a Decision Tree or Classification and Regression Tree (CART)[8].

Category #1: Parental Attributes

The literature on the factors affecting the parent's concern about their children's physical health points out a set of variables such as parents' gender, age, educational attainment, Immigration Status, Belonging to a Minority, and Belonging to an Indigenous Group.

Parents' Gender

The existing body of work highlights the heterogeneous responses exhibited by fathers and mothers regarding various family-related issues. In a study by Van der Vegt & Kleinberg (2020), the authors aim to examine the emotional responses of parents during COVID-19 regarding the challenges of the COVID-19 pandemic. To do so, the authors employ a Bayesian hypothesis testing for self-reported emotions using text data from the Real-World Worry Dataset, collected from 2,500 people in the United Kingdom. The authors argue that during a pandemic, women are most worried about the health of the family whereas fathers are shown to be more concerned about the economic impact of the pandemic.

[8] The terms *Decision Tree* and *Classification and Regression Tree (CART)* are used interchangeably throughout the paper.

Similar findings were reported by the study of Waters et al. (2000) who aimed to investigate connections between parents' reports of their health and their perception of their kids' health. For this, the authors utilize a logistic regression model that relies on survey data from a sample of children aged 5 to 18 years in Australia. They demonstrate a significant correlation between a mother's self-reported health and her perception regarding the health of her child. On the contrary, this was not observed for the surveyed fathers.

Parents' Age

Another factor that plays a role in affecting parents' concern about their children's physical health is the parents' age, although there has been no consensus yet on whether the effect is positive or negative. In this regard, de Buhr, E., & Tannen, A. (2020) examined the effect of parental health literacy, health knowledge, and parental age on the physical activity of children using bivariate and multivariate analyses and relying on a cross-sectional data with 4217 surveyed parents of children in German schools. The authors highlight a strong correlation between parental health literacy and parental age, which in turn was linked to children's healthier behaviors. As such, the authors argue that initiatives to raise parental health literacy may have a positive spillover effect on children's health.

Similar results were reached by Simpson (2022), Petersen et al. (2020), and Rhodes et al. (2020) who all conducted a systematic review of the literature examining the effect of parental support on child physical activity (PA) whereby confirming the role of parental age in affecting children's physical activity. On the other hand, another systematic study paper by Davids & Roman (2014) examines the relationship between various parenting characteristics such as parental age and kids' levels of physical activity shows that parental age does not significantly affect a child's physical health. However, one should note that all papers using systematic reviews share the limitation of publication bias whereby the review covers only papers that made it to publication which may lead to lost insights from papers that didn't reach the publication stage. Also, the studies included in a systematic review may vary in several aspects such as design, sample size, and methodology, which makes it challenging to safely aggregate results across all.

Parents' Educational Attainment

Parents' educational attainment was another variable shown to affect children's physical health. In this regard, de Buhr & Tannen (2020) argue that an increase in parental health literacy of German parents improves children's healthy behaviors such as increased consumption of fruits and vegetables, regular tooth brushing, and increased physical activity. In the Canadian context, we observe similar findings in the qualitative study of Zecevic et al. (2010) who investigated the effect of multiple parental characteristics on the physical activity and sedentary habits of preschool children in Canada using data collected from 102 questionnaires answered by parents and show that parents' educational level is positively related to children's physical activity. However, limitations of the study include being focused on one location, namely Sudbury in Ontario, and participants being self-selected which introduces bias.

Parents' Immigration Status

Another individual parental characteristic that was shown to have a relationship with the physical activity of children is whether parents are immigrants or not. For that, Lacoste et al. (2020) conducted a systematic review of the literature examining 11 studies on the patterns of physical activity of immigrants versus nonimmigrant children in Canada. The authors argue that immigrant children engaged in less physical activity and that Canadian-born parents were less worried about their children's physical activity. On a similar note, Clark (2008) presents a descriptive statistical analysis using data from General Social Surveys and argues that recent immigrant children are less likely to participate in sports which may be attributed to some challenges such as their parents trying to establish economic stability in the host country which makes it challenging for the parents to enroll their children in physical activities given the limited finances. Numerically, the authors show that children of recent immigrants (those who had been in Canada for less than 10 years) are 32% less likely to participate in sports compared to children of Canadian-born parents or those who have been in Canada for over 10 years.

Belonging to a Minority

Belonging to a minority is another factor that was shown in the literature to affect children's physical activity. In this context, Mahmood et al. (2019) analyze the physical activity of Asian and South Asian immigrants to Canada by employing a

multinomial logistic regression based on a cross-section data set of 9683 immigrants arguing that these individuals show low levels of physical activity. Empirically, the authors show that 60% of new immigrants were found to be physically inactive compared to more established immigrants among whom 53% were shown to be inactive.

Along the same lines, Heidinger & Cotter (2020) present a descriptive analysis using crowdsourcing cross-sectional data by Statistics Canada that complements the previous study by highlighting possible factors behind recent immigrants being less active. The authors argue that individuals belonging to a minority group are more likely to report feelings of insecurity when engaging in certain physical activities such as walking alone at night and using public transit due to the occurrence of discriminating incidences based on their skin color, ethnic background, or religious affiliation. However, one limitation of this study is that using crowdsourcing data has multiple limitations such as the self-selection problem.

Given the above studies, we hypothesize the following:

Hypothesis 1 (H1): parents' attributes influence parents' concern about their children's physical health during COVID-19.

Category #2: Child Activities and Eating Impacts

The literature on the effects of COVID-19 has shown great disturbances in the daily lives of children. Among the main disturbances are changes in screen time and eating behaviors.

Daily Screen and Daily Video Games

The impact of the COVID-19 virus outbreak on the movement and play behaviors of Canadian children and teens was investigated by different papers. In this regard, Moore et al. (2020) examine the latter effect using cross-sectional survey data covering 1568 parents collected using online surveys. The study reveals that only 4.8% of children and 0.6% of adolescents adhered to the recommended combined movement behavior during COVID-19 closures. The authors also point to decreased

physical activity and outdoor engagement along with increased sedentary behaviors such as extended screen time.

In a similar study by Guerrero et al. (2020), the authors employ a Decision Tree model to analyze survey data from 1472 Canadian parents. They highlight a significant nonadherence to the 24-hour movement guidelines amidst the COVID-19 pandemic. Moreover, the authors emphasize that adherence may be influenced by certain parental attributes such as the parental ability to limit screen time. Yet, some of the limitations of the study include self-selection bias in data collection using surveys and that the study relies only on Decision Tree which can identify patterns but cannot be used to derive causal relationships.

Eating junk

Another major change in children's behavior during COVID-19 is their eating behaviors. In this regard, Maximova et al. (2022) estimate a multivariable logistic regression using cross-sectional data from 1095 students aged between 9 and 12 years old to investigate the effects of the COVID-19 lockdown on the behaviors and well-being of elementary school children residing in socioeconomically disadvantaged communities in northern Canada. The authors show a significant drop in the physical activity of these children along with an increased consumption of snacks and junk food. They also emphasized that such significant changes in the behaviors of children as well as the presence of pre-existing unhealthy behaviors may increase the chances of developing chronic diseases in the future. Yet, possible limitations of the study include being focused on socioeconomically disadvantaged communities and participants being self-selected. On a similar note, Burkart et al. (2022) examined the dietary patterns and eating behaviors of children during the COVID-19 pandemic using a time series data on 231 children aged between 7 to 12 years and argue that there was a significant increase in children's dietary consumption of both nutritious and less nutritious meals during the period of the pandemic.

Reading

Another child activity that was shown to have a positive spillover effect on physical health is reading. In this regard, Mak and Fancourt (2020) study the relationship between increased reading at the age of 11 and its effect on health-related practices at age 14, which is the age at which children may start substance use. Using data on

11,180 children in the United Kingdom, the authors show that frequent reading is associated with lower chances of cigarette and alcohol consumption and higher chances of healthy eating habits like eating more fruits all of which contribute positively to the physical health of children. Other studies have found a positive effect of reading on other physical-related attributes such as reducing depression. For example, in their paper, Dowrick et al. (2012) analyze the effect of reading on a group of people diagnosed with depression and note that getting into reading was associated with a drop in depression levels for the examined group, which was proved to be a key factor in negatively affecting physical health by increasing chances of sedentary behavior and decreasing levels of physical exercise (Roshanaei-Moghaddam et al., 2009).

Given the above studies, we hypothesized the following:

Hypothesis 2 (H2): changes in children's activities and eating habits influence parents' concern about their children's physical health during COVID-19

Category #3: Labor Market Impacts

COVID-19 induced major changes to the labor market. On one hand, many parents were asked to work remotely from their homes to limit the spread of the virus. With children also remaining at home during lockdowns, parents had to juggle tasks between taking care of their children and performing their job tasks. On the other hand, many of those parents have also either experienced loss of their jobs or a reduction in the number of hours worked.

Balance Childcare and Work

During the COVID-19 pandemic, many parents found themselves working from home. While this was perceived as an advantage in some respects, such as reducing the number of hours commuting to work each day, it introduced new challenges to the parents. One of these challenges is having to balance between taking care of children at home and performing their work-related tasks. In a descriptive analysis study by Carroll et al. (2020), the authors use data from 254 Canadian parents and argue that a significant number of parents encountered difficulties in helping their children at home while also juggling tasks for their professional jobs. As such, many found themselves

forced to work for extended hours which often added to their overall stress levels. On a similar note, a literature review study by Como et al. (2021) on remote work and work-life wellness shows that the difficulty parents face in managing between childcare and work during COVID-19 was mainly due to the lack of a clear separation between work time and other tasks at home. However, possible limitations of this study are the socioeconomic bias whereby more than half of the surveyed parents had an income above $100,000 which is well above the Canadian GDP per capita of around $60,000[9] hence limiting the ability to generalize the findings to parents of lower socioeconomic groups, and the geographic bias being limited to families in Wellington Ontario.

Lost Job or Job Hours

As COVID-19 emerged, many parents found themselves either unemployed or working fewer hours as the market started to contract. In this regard, a descriptive study by Lemieux et al. (2020) using Canadian Labor Force Survey data shows that there has been a significant decrease in the total work hours and employment levels specifically among individuals aged 20 to 64 years. Also, they show that most of the job losses were for those in the lowest earnings quartile.

While losing a job or working fewer hours is likely to affect the ability of a parent to enroll their children in physical-related activities, other studies argue for a different point of view. For example, a study conducted by Cost et al. (2021), who employs a multinomial logistic regression that utilizes cross-sectional data collected from an online survey targeting Canadian parent's data to examine the consequences of parental unemployment during the COVID-19 pandemic, shows that parents who experienced job loss may have seen a decrease in the challenge of balancing work and childcare which in turn results in reduced stress levels. This in turn allowed parents to better monitor their children's physical activity. Furthermore, the authors argue that the emergency financial benefits provided in Canada during the pandemic have mitigated the immediate financial risks parents faced. However, the study had a demographic bias as low-income parents were underrepresented in the used data.

[9] See: https://tradingeconomics.com/canada/gdp-per-capita#:~:text=GDP%20per%20Capita%20in%20Canada%20is%20expected%20to%20reach%2045719.00,macro%20models%20and%20analysts%20expectations.

Parents' being Less Patient

Due to working from home and taking care of the children, many parents became less patient with their children during COVID-19. In our dataset, around 87% of the surveyed parents reported concerns about being less patient with their children during COVID-19. For that, we investigate the existing literature to observe whether parental stress levels impact children's physical activity.

In a paper by Walton et al. (2014), the authors employ a logistic regression model using Cross-sectional data from 110 parents in the United States to examine the link between parental stress and children's physical activity and highlight that parental stress does affect the physical health of children. Specifically, the authors argue that children with stressed parents were less likely to follow the recommended levels of physical activity. However, the study has some limitations such as data being socioeconomically biased to low-income parents in addition to the self-selection problem.

Another study confirming the linkage between parental stress and health attributes of the child was that by Stenhammar et al. (2010) who estimated a logistic regression model using cross-sectional survey data from 873 Swedish parents. The authors argue for an association between parental stress and children's body mass index (BMI) but show that the latter effect can either lead to a high BMI above the healthy level, leading to obesity, or a low BMI below, leading to being underweight whereby both BMI's negatively affect children's physical health. Yet limitations of the study include being focused on parents from Uppsala County in Sweden in addition to the self-selection of participating parents.

Given the above studies, we hypothesize the following:

Hypothesis 3 (H3): the labor market effects of the pandemic influence parents' concern about their children's physical health during COVID-19.

Category #4: Social Impacts

Socializing with Friends and Family

Research has shown that COVID-19 lockdowns limited the children's ability to socialize with friends and family, which negatively affected their physical activity. In this context, Ellis et al. (2020) who examine the effect of physical isolation during COVID-19 using hierarchical regression analyses and data from 1,054 Canadian high school students, note a considerable drop in levels of physical activity among teens falling significantly below the recommended daily threshold of 60 minutes of moderate to vigorous exercise. Similar findings were derived by Szpunar et al. (2021) who applied thematic analysis on qualitative data from interviews with parents and noted that the closures of sports facilities resulted in a decrease in children's participation in moderate to vigorous physical activity. The authors also show that this was in line with the parent's concern about the lack of social interaction among children and the limited availability of structured extracurricular activities during COVID-19. The study, however, had two limitations namely being focused on Ontario and using a small sample size of 12 interviews with parents and 9 interviews with children.

Hypothesis 4 (H4): not being able to socialize with friends and family because of COVID-19 influences parents' concern about their children's physical health during COVID-19.

Category #5: Childcare Impacts

Besides the closure of sports facilities and academic institutions, childcare services have also been closed as the number of COVID-19 cases started to increase. In this line, several papers examined the effect of closures of childcare centers on children's physical activity. For example, Carroll et al. (2020) provide a thematic analysis using online survey data from 254 families to show that closures had a significant impact on access to childcare facilities which led to fewer options for physical activity, further intensifying children's sedentary behavior. The study, however, had two main limitations, the first being focused on middle to high-income families and the second being limited to families in Wellington Ontario.

In the same vein, Lafave et al. (2021), who also apply thematic analysis on data from interviews with 17 educators working in early childhood care centers during the pandemic, highlight the different adverse effects of COVID-19 on nutrition and physical activity of children amid the reopening of early childhood education and care centers. One limitation of this study is that it is limited to the province of Alberta.

Given the above studies, we hypothesized the following:

Hypothesis 5 (H5): usage of childcare services influences parents' concern about their children's physical health during COVID-19.

Chapter 3: Data

3.1 Data Description

To answer our research question, we rely on a cross-sectional dataset published by Statistics Canada titled "Impacts of COVID-19 on Canadians' Parenting During the Pandemic[10]", which was collected between June 9, 2022, and June 22, 2022. This is approximately 2.5 years after the World Health Organization (WHO) declared the COVID-19 outbreak as a pandemic on March 11, 2020. The target sample for the questionnaire was Canadian parents with a child less than 15 years old residing in one of the ten Canadian provinces or any of the three territories during the data collection period[11].

Participation in the study was done voluntarily and data was collected directly from participants via the self-administered questionnaire that was designed to take around 5 minutes. It is worth noting that no data imputation was carried out to replace missing values in the dataset as less than 1% of the participants only did not provide answers to all questions in the survey and those responses were eliminated from the data. Moreover, the questionnaire was designed with techniques to avoid any inconsistencies and illogical responses by respondents. One of the incorporated techniques was the *Automatic Control of Flows* which ensures adapts the next survey based on the answers of the participants to the previous questions. For example, if the respondent's answer to "Are you married?" is yes, then the next question would be about the spouse otherwise the question on the spouse won't show up. This feature ensures that no questions are left empty, which is evident in the low number of missing values, in addition to ensuring a logical flow of questions. Another interesting feature that was employed was the *Use of Edits* which eliminates logical inconsistencies. For example, if a respondent is asked "Do you have children?" and his/her answer is yes, then the computer system will make sure that the answer to the next question about

[10] Link to access all details about the data: https://www150.statcan.gc.ca/n1/pub/45-25-0006/452500062020001-eng.htm

[11] Canada is divided into 10 provinces and 3 territories as follows:

- Provinces: Alberta, British Columbia, Manitoba, New Brunswick, Newfoundland and Labrador, Nova Scotia, Ontario, Prince Edward Island, Quebec, and Saskatchewan.
- Territories: Northwest Territories, Nunavut, and Yukon.

the number of children is a number equal to or greater than 1. This feature ensures that collected responses are coherent.

In addition to being highly consistent and with minimal missing values, the dataset has other features that make it suitable to answer our research question. First, with around 30,000 observations from a diverse group of families of different socioeconomic characteristics living across Canada, the data offers a comprehensive picture about the diversity in the Canadian parent population. This enhances the generalizability of the findings of this study to parents coming from different provinces and territories. Second, the data contains several factors that are similar to those found in previous research and can therefore be used to predict parental concern regarding their children's health.

The survey's main goal was to examine the concerns and experiences of parents regarding their children's mental and physical health, social life, childcare, and schooling activities in the context of COVID-19, from March 15, 2020, to June 9, 2022. Additionally, the survey intended to create a Public Use Microdata File (PUMF) which ensures data confidentiality while making it accessible for public use in examining the effects of COVID-19 on Canadian families.

This data collection through crowdsourcing techniques marks an innovative departure from traditional survey methods. In this context, crowdsourcing refers to gathering information from participants who are invited to take an online questionnaire. This technique was chosen for its timeliness, cost-effectiveness, and safety in the context of the COVID-19 pandemic. However, one limitation is that the data was not collected using a non-probabilistic approach, meaning participants self-selected, potentially leading to results that may not be easily generalized to the larger population of Canadian parents.

3.2 Dependent Variable

Our dependent variable is a categorical variable representing how concerned parents are about their children’s physical health during the COVID-19 period starting from March 15, 2020, to June 9, 2022. The survey question related to the dependent variable reads as follows:

"Due to the COVID-19 Pandemic, how concerned are you about the physical health of your child or children aged 0 to 14 years?"

The variable then takes four values from 1 to 4 depending on the level of parents' concern with 1 being "*Not at All Concerned*", 2 being "*Somewhat Concerned*", 3 being "*Very Concerned*", and lastly 4 being "*Extremely Concerned*". The distribution of the dependent variable is shown in Table 1. It is worth mentioning that over 70% of surveyed parents had a response of *Somewhat Concerned*, *Very Concerned*, or *Extremely Concerned* which highlights the motivation to examine the factors predicting parent's concern about their children's physical health during COVID-19.

Table 1: Distribution of the Dependent Variable

Parents' Concern About Their Children's Physical Health	Frequency	Percent	Cumulative Percentage
Not at all Concerned	7,172	28.74	28.74
Somewhat Concerned	12,188	48.84	77.58
Very Concerned	3,740	14.99	92.56
Extremely Concerned	1,856	7.44	100.00
Total	24,956	100.00	

Note: this table presents the distribution of the Dependent Variable which is categorical and ordinal representing how concerned parents are about their children's physical health during the COVID-19 period in Canada.

3.3 Explanatory Variables

Based on our literature review, we identified five primary predictor variable groups. We then analyzed the dataset to extract comparable variables for our model. Below is the list of explanatory variables included in our model sorted by group type. We note that ordinal categorical predictor variables have been transformed to binary for the simplicity of interpreting the respective odds ratio. For a binary predictor variable, the respective odds ratio is analyzed as the effect of the binary variable taking the value of 1 as opposed to 0 on the chances of moving to a higher category in the response variable. For categorical and ordinal predictor variables originally, the odds ratio would represent a more complex effect related to moving to a higher category in both the predictor and the dependent variable.

Parental Attributes

This category captures six variables related to the parents' individual characteristics namely: parents' gender, age, educational attainment level, immigrant status, and if they belong to a certain minority group or an indigenous community in Canada.

All variables are dummies and configured as follows:

- Female: this variable captures parents' gender and takes the value of 1 if the surveyed parent is a female and 0 if male.
- Old: this variable captures parents' age and takes the value of 1 if the surveyed parent is above 45 years old and 0 otherwise.
- University Degree: this variable captures the educational attainment of the parent and takes the value of 1 if the surveyed parent attended university and 0 otherwise.
- Immigrant: this variable captures the immigration status and takes the value of 1 if the surveyed parent is an immigrant and 0 otherwise.
- Minority: this variable captures if parents belong to a visible minority group and takes the value of 1 if the surveyed parent belongs to a visible minority in Canada and 0 otherwise.
- Indigenous[12]: this variable captures whether the surveyed parent belongs to an indigenous community in Canada and takes the value of 1 if they do and 0 otherwise.

Child Activities and Eating Impacts

As discussed in the above literature, COVID-19 has altered children's activity and eating habits in several ways. As such, our second group of variables includes three different predictors about children's activities and eating habits during COVID-19 namely: (1) how often children spent time in front of a screen for studying and watching TV, (2) how often children play video games, and lastly (3) how concerned parents are about their children's consumption of junk food during COVID-19.

All three variables are dummies and configured as follows:

[12] There are three recognized indigenous communities in Canada: First Nations, Métis, and Inuit.

- Daily Screen: this variable tracks the frequency at which children are in front of a screen and takes the value of 1 if children are in front of a screen daily for studying and watching TV, and 0 otherwise.
- Daily Video Games: this variable tracks the frequency at which children are playing video games and takes the value of 1 if they play video games daily and 0 otherwise.
- Eat Junk: this variable tracks how concerned parents are about their children's consumption of junk food taking a value of 1 if parents are somewhat, very, or extremely concerned and 0 otherwise.

Labor Market Impacts

COVID-19 caused a major disruption to the labor market from March 15, 2020, to June 9, 2022, as per our survey. This category includes three variables that track the impact on parents' work: whether they lost their jobs, experienced a significant reduction in their work hours, and how concerned they were about balancing childcare and work tasks at home during this period. Lastly, the category also includes a variable on parents' concerns about being less patient with their children during COVID-19.

All three variables are dummies and configured as follows:

- Lost Job or Hours: this variable takes the value of 1 if the parent lost his/her job or had their working hours reduced during COVID-19 and 0 otherwise.
- The Balance Between Childcare and Work: this variable takes the value of 1 if the parent is somewhat, very, or extremely concerned about balancing between childcare and work and 0 otherwise.
- Less Patient with Child: this variable takes the value of 1 if the parent is somewhat, very, or extremely concerned about being less patient with his/her children during COVID-19 and 0 otherwise.

Social Impacts

COVID-19 had a significant social impact on children, causing them to miss out on important opportunities for socialization within their communities. For that, we include a measure that captures the social impact of COVID-19 on Children.

The variable is a dummy and configured as follows:

- Connect with Friends and Family: this variable tracks the concern of parents about their children remaining connected to friends and family during COVID-19 and takes the value of 1 if the parent is somewhat, very, or extremely concerned and 0 otherwise.

Childcare Impacts

Many studies have documented a decline in work-life balance among parents during COVID-19 due to increased childcare demands. Therefore, we incorporate childcare-related factors into our model to examine their effect on parents' concerns about their children's physical health.

Both variables included are dummies and are configured as follows:

- Used Childcare: this variable tracks whether parents have used childcare services before COVID-19 and takes the value of 1 if the parent did use childcare services and 0 otherwise.
- Will Use Childcare: this variable tracks whether parents will use childcare services in the future and takes the value of 1 if the parents plan to do so and 0 otherwise.

3.4 Summary Statistics

Table 2 presents the summary statistics of the created binary categorization variables included in our model. First, we note that the average of the dependent variable is around 2 indicating that parents on average were somewhat concerned about their children's physical health during COVID-19. Moreover, we note that 91% of surveyed parents were females. Another notable average was that around 90% of surveyed parents said that their children are spending time daily in front of a screen watching TV or using their phones which agrees with literature showing an increase in sedentary

behavior during COVID-19. Moreover, around 94% of parents reported being somewhat, very, or extremely concerned about balancing their work and childcare duties. This is best explained by parents having to work mostly from home during COVID-19 with their children at home too, which adds another duty to their work as compared to the period before COVID-19 were children used to go to school. Another alarming figure is the average of parents' concern for children remaining connected with family & friends during COVID-19 being above 90% showing that almost all parents were somewhat, very, or extremely concerned about the social effect of COVID-19 on their children which comes in agreement with the literature discussing the increased isolation during periods of COVID-19. Lastly, around 88% of parents expressed concern about decreased patience with their children during COVID-19. This may be linked to work-life balance challenges, stress from COVID-19 worries, and potential job losses/reduced work hours reported by 40% of respondents.

Table 2: Summary Statistics

Variables	Mean	Standard Deviation	Minimum	Maximum
Dependent Variable				
Parents' Concern for Children's Physical Health	2.011	0.857	1	4
Explanatory Variables				
Female Parent	0.912	0.284	0	1
Old Parent	0.155	0.362	0	1
Parent Holds a University Degree	0.746	0.435	0	1
Immigrant Parent	0.119	0.324	0	1
Parents belong to a Minority Group	0.111	0.314	0	1
Parents Identify as Indigenous	0.028	0.165	0	1
Child Spends Time Daily in Front of a Screen (Multiple Activities)	0.898	0.303	0	1
Child Plays Video Games Daily	0.228	0.419	0	1
Concern for Child Eating Junk Food	0.645	0.478	0	1
Child Reads Daily	0.653	0.476	0	1
Parents' Concern for Balancing between Childcare and Work	0.937	0.244	0	1
Parents' Lost Job or Experienced a drop in Their Working Hours	0.393	0.488	0	1
Parents' Concerns About Being Less Patient With Their Children	0.872	0.334	0	1
Parents' Concern for Children Connecting with Family & Friends	0.914	0.280	0	1
Parent Will Use Childcare Services	0.354	0.478	0	1
Parents Did Not Stop Using Childcare Services	0.078	0.268	0	1

Note: this table presents the summary statistics of all variables included in the model.

3.5 Multicollinearity Checks

We conduct two diagnostic checks for multicollinearity: the *Correlation Matrix* and the *Variance Inflation Factors (VIF)*[13]. For the former check, we observe all correlation coefficients to be well below 0.8. As for the latter check, we observe a mean VIF of 1.09 with the highest individual VIF of 1.24, well below the threshold of

[13] We present a more detailed discussion on both checks in the appendix.

5. Hence, both checks show a low risk of multicollinearity among the chosen set of explanatory variables.

Chapter 4: Model Selection and Machine Learning

4.1 Model Selection

This section discusses the models used to examine factors affecting parents' perceptions of their children's physical health during COVID-19. Since our dependent variable is categorical and ordinal (measuring the degree of concern), several models can be used, both parametric and non-parametric. As such, we compare the performance of five different models:

1. Ordered Logistic Model (parametric)
2. Gaussian Naive Bayes Model (parametric)
3. Random Forest (non-parametric)
4. Decision Tree (non-parametric)
5. K-nearest-neighbor (non-parametric).

Endogeneity Considerations

One concern in model selection is endogeneity, which limits causal inference due to Reverse Causality and Omitted Variable Bias (OVB). A possible solution to the Reverse Causality problem would be employing a Two-Stage Model which is used to address the endogeneity problem using Instrumental Variables. However, implementing a Two-Stage model poses some challenges such as identifying valid instruments and ensuring that the exogeneity condition applies to them. As for the second source of Endogeneity, one could include as many explanatory variables as possible in the model to lower the bias from omitted variables. However, this approach suffers the *Curse of Dimensionality* limitation whereby the more variables we add, the more the model will use these features to fit the noise in the training dataset leading to overfitting. Consequently, overfitting the model to the training set would likely result in excellent marks for that data yet disappointing scores for out-of-sample inputs, diminishing the predictive power of the technique on novel cases. As such, we opt to add multiple explanatory variables while avoiding having too many of them so that we don’t weaken the predictive power of our model.

This study prioritizes identifying factors correlated with parental concern, focusing on the model's ability to predict, rather than establish causal relationships. In other words,

our analysis aims to forecast changes in the dependent variable's labels, emphasizing the model's predictive accuracy over unbiased causal effect estimation.

Model Evaluation

To do our comparison, we compute four different model performance metrics:

1. Cross Validation Accuracy Score (CV Accuracy Score)
2. Weighted Precision
3. Weighted Sensitivity (also known as Weighted Recall)
4. Weighted Specificity

To comprehensively evaluate the models, we compute the average of four performance metrics: Cross-Validation Accuracy Score, Weighted Precision, Weighted Sensitivity (Recall), and Weighted Specificity. This equally weighted average serves as our model selection metric, which we term the Overall Performance Score. We select the models with the highest Overall Performance Scores for further analysis in Chapter 5, one for parametric and one for non-parametric approaches.

Cross Validation Accuracy Score (CV Accuracy Score)

In machine learning, accuracy is a common metric for classification models. It reflects the proportion of correctly predicted instances relative to the total number of predictions. Mathematically, accuracy is defined as:

Equation 1: Accuracy

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}$$

Whereby:

- *Number of Correct Predictions* refers to the correctly predicted instances or the True Positives (TP) and True Negatives (TN).
- *Total Number of Predictions* refers to the total number of predicted instances which is the sum of the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

In this paper, we leverage a machine learning technique known as *Cross-Validation* to compute the *Cross Validation Accuracy Score* which is the first performance metric

we use to evaluate the performance of the different models. Cross-validation is a machine learning technique that works by splitting the splitting dataset into k random different subsamples; also referred to as folds. The model is then trained on the first fold and tested on the remaining $k-1$ folds to evaluate its performance. This process is repeated k times whereby the performance metrics obtained from each iteration are aggregated to obtain an overall average estimate of the model's performance. This estimate is the *Cross Validation Accuracy Score* which is mathematically represented as follows:

Equation 2: Cross Validation Accuracy Score

$$Cross\ Validation\ Accuracy\ Score = \frac{1}{k}\sum_{i=1}^{k} Accuracy_{\ i}$$

Whereby:

- k is the number of folds in the cross-validation process
- $Accuracy_{\ i}$ is the accuracy score obtained from the i^{th} fold.

For our example, we pick $k = 10$. As such, the data will then be shuffled and split into 10 different subsamples[14]. Out of these 10 subsamples, 9 are chosen to form the *Training Dataset,* and the remaining dataset is assigned to be the *Testing Dataset*. The model is then trained on the *Training Dataset* and is tested on the *Testing Dataset*.

Training the Model is a term used to refer to the process whereby the model learns the relationship between different *Features* and *Labels* in the training dataset. *Features* refer to the set of Independent Variables and *Labels* refer to the set of Dependent Variables. The learning process starts with the model learning the relationship between the different features and labels in the training set and iteratively repeats this process multiple times while adjusting the model's parameters, which consist of *Weights* and *Biases,* throughout the training process. *Weights* refer to the coefficients that the model's features are multiplied with while *Biases* represent the constants added to the model features. In this process of repeated adjustments of *Weights* and *Biases*, the model is solving a Minimization Problem whereby the function to be minimized

[14] While splitting the subsamples, it is essential to ensure that they are split equally and that the number of folds chosen is less than the number of observations.

represents the error difference[15] between the actual labels in the training dataset and the predicted labels in the testing dataset. When this error is minimized, we can say that the model has been trained. This machine learning process is referred to as *Gradient Descent* whereby every time the Weights and Biases are adjusted, the trained model is evaluated on the Testing Dataset by assessing how accurate is the model in predicting the correct Labels using the given Features. Once this process is finished, the accuracy score is kept, and the model is deleted. The whole practice is then repeated k times by choosing a different testing dataset while keeping the remaining $k-1$ subsamples as the training dataset. Thus, in our case, we end up with 10 accuracy scores whereby the accuracy score of the whole model will be their simple average which we refer to as the *Cross Validation Accuracy Score.*

Weighted Precision

The second performance metric is Precision. This metric, primarily used for binary classification problems, measures the proportion of correctly predicted positive instances out of all positive instances. In simpler terms, it reflects the model's ability to avoid false positives. Mathematically, precision is calculated as:

Equation 3: Precision

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

For example, if a spam detection software labels 100 emails as spam and only 70 are actually spam, the precision would be 70% (True Positives: 70, False Positives: 30).

Since our dependent variable is an ordered categorical variable, the standard Precision metric doesn't directly apply. Therefore, we compute an adjusted metric called Weighted Precision. This metric acknowledges the categorical nature of the variable by calculating precision for each level of concern ("Not Concerned," "Somewhat Concerned," etc…). Additionally, incorporated weights address potential imbalances in the class distribution of the dependent variable. The Weighted Precision metric is computed as follows:

Equation 4: Weighted Precision

[15] The equivalent of minimizing the *Error Difference* is maximizing the *Prediction Accuracy.*

$$\text{Weighted Precision} = \sum_{i=1}^{N} \text{Precision}_i \times \text{Weight}_i$$

Whereby:

- Precision_i is the ratio of the correctly predicted positive instances out of all positive instances for class i
- Weight_i is the weight assigned to class i

Weighted Sensitivity/Recall

The second performance metric is Recall. This metric, primarily used for binary classification problems, measures the correctly predicted positive instances out of all actually positive instances. As such it measures the ability of a model to identify false negatives. Mathematically, Recall is computed as follows:

Equation 5: Recall

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Going back to the spam email example, assume there are 120 actual spam emails, and the model captures 80 of those as spam. Then the Recall would be 67% as the True Positives are 80 while the False Negatives are 40.

Given that our dependent variable is an ordered categorical variable, then the standard Recall metric does not directly apply and hence we compute an adjusted precision metric known as the Weighted Recall. This metric acknowledges the categorical nature of the dependent variable by computing the Recall for each level of concern. Moreover, the incorporated weights cater to the imbalances in the class distribution of the dependent variable. Hence, the Weighted Recall metric is computed as follows:

Equation 6: Weighted Recall

$$\text{Weighted Recall} = \sum_{i=1}^{N} \text{Recall}_i \times \text{Weight}_i$$

Whereby:

- Recall_i is the ratio of the correctly predicted positive instances out of all actual positive instances for class i
- Weight_i is the weight assigned to class i

Weighted Specificity

The third performance metric is Specificity. This metric, primarily used for binary classification problems, measures the ratio of correctly predicted negative instances to the total actual negative instances. As such, it measures the model's ability to avoid false positives among the actual negative instances. Mathematically, Specificity is computed as follows:

Equation 7: Specificity

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Going back to the spam email example, assume 100 emails are non-spam and the model identifies 70 of those as non-spam. Then the Specificity would be 70% as the True Negatives are 70 while the False Positives are 30.

Given that our dependent variable is an ordered categorical variable, then the standard Specificity metric does not directly apply and hence we compute an adjusted Specificity metric known as the Weighted Specificity. This metric acknowledges the categorical nature of the dependent variable by computing the Specificity for each level of concern. Moreover, the incorporated weights cater to the imbalances in the class distribution of the dependent variable. Hence, the Weighted Specificity metric is computed as follows:

Equation 8: Weighted Specificity

$$\text{Weighted Specificity} = \sum_{i=1}^{N} \text{Specificity}_i \times \text{Weight}_i$$

Whereby:

- Specificity_i the ratio of correctly predicted negative instances to the total actual negative instances for class i
- Weight_i is the weight assigned to class i

Overall Performance Score

To comprehensively evaluate the models, we compute the simple average of all four scores discussed above (CV Accuracy Score, Weighted Precision, Weighted Recall, and Weighted Specificity). This average is termed the *Overall Performance Score* and is mathematically computed as follows:

Equation 9: Overall Performance Score

$$Overall\ Performance\ Score = \frac{CV\ Accuracy + W\_Precision + W\_Recall + W\_Specificity}{4}$$

Whereby:

- *CV Accuracy*: represents the Cross Validation Accuracy Score
- *W_Precision*: represents the Weighted Precision
- *W_Recall*: represents the Weighted Recall
- *W_Specificity*: represents the Weighted Specificity

Table 3 presents the results for all metrics. The Ordered Logistic Model achieved the highest Overall Performance Score (47.82%), followed by Random Forest (47.55%) and Decision Tree (46.83%). These scores show a substantial improvement over the random guess probability (25% in this case with four categories). We also note that the CV Accuracy Score ranking aligns with the ranking based on the Overall Performance Score which is an advantage for deciding on the best model as the CV accuracy score provides a general assessment of the model's overall correctness. Moreover, Precision, Recall, and Specificity are not inherently suited for categorical classification as they are used for binary classifications which require us to make the above-mentioned adjustment. While Weighted Precision, Weighted Recall, and Weighted Specificity provide a more comprehensive evaluation of the models' performance, having consistent results when taking them into account as compared to looking for CV Accuracy Scores only strengthens the model selection procedure.

Table 3: Models' Performance Metrics

Model	Type of Model	CV Accuracy Score	Weighted Precision	Weighted Recall	Weighted Specificity	Overall Performance Score
Ordered Logistic Model	Parametric	49.92%	39.79%	50.90%	50.65%	47.82%
Random Forest Classifier	Non-Parametric	47.52 %	41.84%	48.48%	52.34%	47.55%
Decision Tree Classifier	Non-Parametric	46.88 %	40.26%	47.82%	52.34%	46.83%
Gaussian Naive Bayes Model	Parametric	44.85%	41.56%	45.07%	50.88%	45.59%
K Nearest Neighbors Classifier	Non-Parametric	42.80%	39.74%	43.35%	50.07%	43.99%

Notes: this table presents the *Cross-Validation Accuracy Score, Weighted Precision, Weighted Recall, Weighted Specificity, and Overall Performance Score* of the five different models suited for the dependent variable which is categorical and ordinal.

4.2 Parametric Model: Generalized Ordered Logit with Partial Proportional Odds

As observed in Table 3, the Ordered Logistic Model was shown to have the highest overall performance among the five different models with a score of 47.82%. The Ordered Logistic Model can be written as shown in Equation 10 whereby j represents the categories of the dependent variable starting from category number 1 to the last category M, X_i represents a vector of predictor variables for the i^{th} observation, and $P(Y_i > j)$ represents the probability that the category of the dependent variable for the i^{th} observation is greater than category j. The exp in the numerator is an exponential function to ensure that the calculated probability is positive whereby the whole numerator is then normalized by dividing with $1 + numerator$ to make sure the estimated probability $P(Y_i > j)$ is between 0 and 1.

Equation 10: Ordered Logistic Model (Probability)

$$P(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta)}{1 + \{\exp(\alpha_j + X_i\beta)\}}, j = 1,2, \dots, M-1$$

In practice, the estimated coefficients of the Ordered Logit Model represent the Log-Odds which are calculated by taking the logarithmic of the Odds whereby the Odds is the division of $P(Y_i > j)$ by $1 - P(Y_i > j)$ as shown below in Equation 11.

Equation 11: Ordered Logistic Model (Log-Odds)

$$\log\left(\frac{P(Y_i > j)}{1 - P(Y_i > j)}\right) = \alpha_j + X_i\beta_j$$

While one may just report log odds as estimated by the model, a common practice is to present the output in terms of the Odds Ratio which is calculated by exponentiating the estimated regression coefficient and is used as a measure of the effect of the predictor variable on the response variable. As such, an odds ratio that is significant and above 1 indicates a positive effect of the predictor variable on the response variable i.e., a higher odds of moving to a higher category of the dependent variable. Similarly, an odds ratio that is significant and between 0 and 1 indicates a negative effect of the predictor variable on the response variable i.e., a higher odds of moving to a lower category of the dependent variable. The formula of the Odds Ratio is presented in Equation 12. The analysis of the odds ratio can be done in percentage terms, whereby we subtract one from the calculated odds ratio and multiply the result by 100 to see the percentage effect of the predictor variable on the response variable[16].

Equation 12: Ordered Logistic Model (Odds Ratio)

$$\text{Odds Ratio } = \exp\left(\beta_j\right)$$

One of the key assumptions of the Ordered Logistic Model is the Proportional Odds Assumption – also known as the Parallel Lines Assumption and the Parallel Regression Assumption – which assumes that the estimated β*'s* do not differ across the different categories of j. In other words, the Proportional Odds Assumption assumes that the effect of a predictor variable X on the movement from category 1 to a higher category 2, depicted by β_1, is the same as moving from category 2 to category 3 which is depicted by β_2 and so on. As argued by Williams (2006), the Ordered Logistic Model is too restrictive whereby the latter assumption of proportional odds often being violated. In this context, we test for the Proportional Odds Assumption before proceeding to the Ordered Logit Model using the Brant test[17] and report the results in Table 4. The null hypothesis or H_0 states that all predictor variables in the model have the same effect across all categories of the dependent variable (that is the estimated coefficients are all equal: $\beta_1 = \beta_2 = \cdots = \beta_{j-1}$). This is why we don't see

[16] For example, an odds ratio of a dummy variable of 1.25 is analyzed as follows: $(1.25 - 1) * 100 = 25\%$ thus when the dummy variable of interest takes the value of 1, there is 25% odds of moving up to a higher category of the dependent variable. Similarly, if the odds ratio was 0.75 then $(0.75 - 1) * 100 = -25\%$ indicating that when the dummy variable of interest takes the value of 1, there is a 25% odds of moving down to a lower category of the dependent variable.

[17] Details on the Brant test appear in the original journal article by Rollin Brant published in 1990 in the Biometrics journal (Brant, 1990).

a subscript j for the β in the formula of the Ordered Logistic Model. In other words, the null hypothesis assumes that the parallel lines assumption holds for every feature in the model. As such, a $p-value$ lower than 0.05 for the whole model makes us reject the null hypothesis and conclude that at least one predictor variable is violating the parallel/proportional odds assumption. Similarly, an individual $p-value$ less than 0.05 for any of the predictor variables shows that this specific variable violates the proportional odds assumption.

Table 4: Brant Test Results – Testing for the Parallel Regression Assumption

All Variables	chi2	p>chi2	df	Violates Proportional Odds
	184.32	0.000	32	Yes
Female Parent	1.05	0.592	2	No
Old Parent	13.07	0.001	2	Yes
Parent Holds a University Degree	21.89	0.000	2	Yes
Immigrant Parent	23.92	0.000	2	Yes
Parents belong to a Minority Group	10.90	0.004	2	Yes
Parents Identify as Indigenous	5.62	0.060	2	No
Child Spends Time Daily in Front of a Screen (Multiple Activities)	8.20	0.017	2	Yes
Child Plays Video Games Daily	19.55	0.000	2	Yes
Concern for Child Eating Junk Food	9.41	0.009	2	Yes
Child Reads Daily	2.13	0.344	2	No
Parents' Concern for Balancing Childcare, Schooling & Work	5.53	0.063	2	No
Parents' Lost Job or Experienced a drop in Their Working Hours	1.10	0.578	2	No
Parents' Concerns About Being Less Patient With Their Children	8.37	0.015	2	Yes
Parents' Concern for Children Connecting with Family & Friends	1.21	0.546	2	No
Parent Will Use Childcare Services	2.74	0.254	2	No
Parents Did Not Stop Using Childcare Services	1.94	0.378	2	No

Note: this table presents the results of the Brant test (Brant, 1990) used to check for violation of the "Proportional Odds Assumption"; a key assumption for the Ordered Logistic Model whereby a $p-value$ lower than 0.05 for the whole model indicates a violation of the assumption by at least on predictor variable and a $p-value$ lower than 0.05 for any specific variable indicates a violation of the assumption by this variable.

From Table 4, we observe a zero $p-value$ for the whole model which is less than 0.05 indicating that at least one predictor variable is violating the proportional odds assumption in the model. In that regard, Williams (2016) argues how researchers faced with this violation tend to either continue using the Ordered Logit Model while acknowledging the violation of the proportional odds assumption or switch to a Multinomial Logit Model both of which carry respective issues. The former shows a clear violation of the assumption of the Ordered Logistic Model which raises concerns about the validity and the interpretation of the estimated coefficients which tend to be

either biased or misleading. As for the latter, the output of the Multinomial Logit Model tends to be more difficult to interpret because of the many parameters it produces in addition to making no benefit from the ordering nature of the dependent variable as Multinomial Logit Models work by comparing each category on its own to a reference category. Given the problems associated with ignoring the violation of the proportional odds assumption or using the multinomial logit model, the author proposes the usage of the Generalized Ordered Logit Model with Partial Proportional Odds[18], which is a special case of the Generalized Ordered Logit Model.

The Generalized Ordered Logit Model does not impose the proportional odds assumption of the Ordered Logit Model. As such, this makes it suitable for cases where the proportional odds assumption is violated like ours. The Generalized Ordered Logistic Model is written as shown in Equation 13 whereby j represents the categories of the dependent variable starting from the first category $j = 1$ to the last category $j = M$.

Equation 13: Generalized Ordered Logistic Model

$$P(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta_j)}{1 + [\exp(\alpha_j + X_i\beta_j)]}, \mathrm{j} = 1,2, \dots, \mathrm{M} - 1$$

Compared to the Ordered Logistic Model, we see that the two formulas are almost the same with the main difference being with the j subscripts for the β whereby the Ordered Logit Model imposing the proportional odds leads to the absence of subscripts j as all β'_s for the same predictor variable X_i will be equal. On the contrary, the generalized ordered logit - which relaxes this assumption for all variables – necessitates the presence of subscripts j to show that the effect of all predictor variables on the different levels of j are not equal. As such, the Ordered Logit Model is just a special case of the less restrictive specification referred to as the Generalized Ordered Logit Model. Specifically, both models are used when the dependent has more than two categories (i.e. when $M > 2$) but one enforces proportional odds while the other does not[19].

[18] In his paper, Williams (2006) introduces the *gologit2* Stata package used to estimate the Generalized Ordered Logit Model and discusses its major strengths.
[19] When $M = 2$, the Generalized Ordered Logit Model collapses to the standard Logistic Regression Model.

In practice, the Generalized Ordered Logit Model estimates a series of binary logistic regressions by combining categories of the dependent variable and comparing them to the rest, unlike the multinomial Logit Model which compares each category to a reference group thus making no use of the ordered nature of the dependent variable. For example, in our case, $M = 4$ as our dependent variable has 4 categories: $j = 1 = not\ at\ all\ concerned$, $j = 2 = somewhat\ concerned$, $j = 3 = very\ concerned$, and $j = 4 = extremely\ concerned$. As such, the Generalized Ordered Logit Model will then produce three different β'_s for each predictor variable coming from three different binary regression comparisons. Namely, when $j = 1$, the model compares the first category (i.e., being *Not at All Concerned*) to the remaining three categories combined (i.e., being *Somewhat Concerned, Very Concerned, or Extremely Concerned)*. Similarly, when $j = 2$, the model compares the first and second categories combined (i.e., *Not at All Concerned and Somewhat Concerned*) to the remaining third and fourth categories combined (i.e*., Very Concerned and Extremely Concerned)*. Lastly, when $j = 3$, the model compares the first, second, and third categories combined (i.e. *Not At All Concerned, Somewhat Concerned, and Very Concerned*) to the fourth category (i.e*., Extremely Concerned)*. That's why for $M = 4$ we will only have $M - 1$ betas because for $j = 4$ we will be comparing all categories combined (i.e., *Not at All Concerned, Somewhat Concerned, Very Concerned, and Extremely Concerned*) to nothing.

Going back to the Brant test for the proportional odds assumption, we notice that not all variables violate this assumption as most of them have p-values greater than 0.05. In fact, only half of them violate the assumption while the other half respects it. As such one might think that relaxing the proportional odds assumption to all the variables – including the ones that do not violate it – will just create unnecessary estimates as the three β'_s of a variable that doesn't violate the proportional odds assumption are indeed the same. From here comes the intuition of not just using the "Generalized Ordered Logit Model" but rather a "Generalized Ordered Logit Model with Partial Proportional Odds" which works by relaxing the proportional odds assumption only for those variables that violate the proportional odds assumption effectively reducing the number unnecessary estimated coefficients. This is done by running a built-in behind-the-scenes Wald Test for every single predictor variable that decides for each variable at a time whether to relax the assumption or not. For example, if we have

three predictor variables, $X1$, $X2$, and $X3$, whereby $X1$ and $X2$ follow the proportional odds assumption while $X3$ violates it, then Generalized Ordered Logistic Model with Partial Proportional Odds can be written as shown in Equation 14 were we see that the β'_s are the same for the first two predictors for all j categories while the β'_s for the third predictor varies for each j.

Equation 14: Generalized Ordered Logit Model with Partial Proportional Odds

$$P(Y_i > j) = \frac{\exp\left(\alpha_j X1_i\beta1 + X2_i\beta2 + X3_i\beta3_j\right)}{1 + \left\{\exp\left(\alpha_j + X1_i\beta1 + X2_i\beta2 + X3_i\beta3_j\right)\right\}}, j = 1,2,\dots,M-1$$

As discussed above, the advantage of this specification is that we reduce the estimation of unnecessary coefficients for predictor variables with equal effect across all categories j. In the context of our model, leveraging the Partial Proportional Odds option in the Generalized Ordered Logistic model helps us reduce the number of $\beta's$ in our model from a total of 48 beta to 32 beta[20].

Another way to test whether the Ordered Logit Model is too restrictive as compared to the Generalized Ordered Logit for modeling the aforementioned data is to conduct a Likelihood Ratio Test (Table 5). The null hypothesis in this case is that there is no difference between the two models. In other words, the null hypothesis is testing whether the more restrictive model which is easier to estimate and analyze (i.e., the Ordered Logit Model in our case) is sufficient to explain the data and hence there is no need for the less restrictive model which adds complexity and is thus harder to estimate and analyze (i.e., the Generalized Ordered Logit in our case)[21]. Given that the $p - value$ from the Likelihood Ratio Test is less than 0.05, we reject the null hypothesis suggesting a statistically significant difference between the two models. In other words, the test shows that the more complex model (i.e., the Generalized

[20] Our model has 16 predicator variables which will result in 16*3=48 $\beta's$ shall we estimate a standard Generalized Ordered Logistic Model. However, given that the proportional odds apply to 8 of these predicators for which we will have just one β for each and doesn't apply for the remaining 8 predictors which we will have three β'_s for each, the new total number of $\beta's$ under the Partial Proportional Odds will be equal to (8*1) + (8*3) = 32 β'_s.

[21] This is similar to testing whether the simpler model (i.e., the Ordered Logit Model) is nested within the more complex model (i.e., the Generalized Ordered Logit Model).

Ordered Logit Model) provides a significantly better fit to the data than the simpler model (i.e., the Ordered Logit Model).

Table 5: Likelihood Ratio Test Comparing the Ordered Logit Model to the Generalized Ordered Logit Model

LR chi2(32)	179.35
p-value	0.0000

Notes: This table provides the result of the Likelihood Ratio Test used to compare the Ordered Logit Model and the Generalized Ordered Logit Model to see if the former is nested within the latter.

Similarly, we also compare the Generalized Ordered Logit with Partial Odds (i.e., the more restrictive and simpler model) to the standard Generalized Ordered Logit Model (i.e., i.e., the less restrictive and complex model) to confirm whether our choice of the less restrictive model is valid (Table 6).

Table 6: Likelihood Ratio Test Comparing the Generalized Ordered Logit Model with Proportional Odds to the standard Generalized Ordered Logit Model

LR chi2(14)	16.31
p-value	0.2949

Notes: This table provides the result of the Likelihood Ratio Test used to compare the Generalized Ordered Logit Model with Partial Proportional Odds and the standard Generalized Ordered Logit Model to see if the former is nested within the latter.

4.3 Non-Parametric Model: Decision Tree

Going back to Table 3 on the performance metrics of the five different tested models, we notice that while the Ordered Logistic Model had the highest overall performance score of 47.82%, other non-parametric techniques like Random Forest and Decision Tree were not far with overall performance scores of 47.55% and 46.83% respectively. Given that, one may not want to ignore these models for the various strengths of non-parametric techniques like Random Forests and Decision Trees.

Decision Trees are supervised non-parametric machine-learning techniques for regression and classification. They are typically trained to predict the value of a target

variable through simple decision rules inferred from features in the dataset, by splitting observations into different subsets that share a common set of characteristics. Decision trees do not require a lot of data preparation, such as normalizing features or creating dummy variables, and handle both numerical and categorical data with ease, which makes them an efficient tool in answering the research question at hand. Moreover, they are easy to interpret as they closely resemble human reasoning making them easy to visualize and understand for both technical and non-technical audiences, unlike parametric models, like the Ordered Logistic Models, which require technical knowledge about the model and how to analyze the estimated log-odds.

The Decision Trees algorithm relies on specific criteria to determine the optimal method for classifying data. These criteria, which include Gini Impurity, Information Gain, and Gain Ratio, evaluate how well a feature separates the classes, intending to produce well-separated results. The default criteria used for splitting the nodes of a Decision Tree is the Gini Impurity.

Gini Impurity is a measure of how pure each of the Decision Tree leaves are. In simple terms, a pure leaf has no misclassified instances whereas a non-pure one does contain a mix of different classes. As such impurity refers to placing an observation in the wrong class. Hence, the algorithm chooses the optimal feature for splitting the data at each node by assessing the ability of the feature to reduce the ambiguity in classifying observations. This is then repeated iteratively whereby Gini Impurity is reduced at each splitting node following a tree-like structure. This process continues until a certain maximum tree depth or a minimum leaf impurity level is reached.

Mathematically, the Gini Impurity for a particular node is computed as follows:

Equation 15: Gini Impurity

$$\text{Gini}_{\text{node}} = 1 - \sum_{i=1}^{k} p_i^2$$

Whereby

- k is the number of classes or categories.
- p_i is the probability of an observation belonging to class i being in the leaf node.

Decision Tree predicts the class of a new observation by carrying succussive splits across the tree branches until it reaches a terminal node that places it in a respective class. The predicted class of the new observation is given by Equation 16, whereby $\hat{y}$ represents the predicted label for the leaf node, $count(y_i = y)$ represents the number of samples that are classified to class y in the leaf node, and lastly the $Total\ samples\ in\ the\ leaf$ refers to the total number of samples that have been classified into the leaf node.

Equation 16: Leaf Prediction Formula Based on Decision Tree

$$\hat{y} = \text{argmax}\left(\frac{\text{count}(y_i = y)}{\text{Total samples in the leaf}}\right)$$

In a similar vein, Random Forest is another machine-learning classification technique that resembles a Forest formed by the combination of multiple individual Decision Trees. Random Forest predicts the class of a new observation by computing the mode of predictions made by multiple Decision Trees (Breiman, 2001). Mathematically, the predicted class of the new observation is given by Equation 17 whereby $\hat{y}_n$ is the class assigned to the new observation by the nth Decision Tree and $mode\ (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$ is a simple mode function that shows the most frequently predicted class by n Decision Trees.

Equation 17: Ensemble Prediction Formula in Random Forest

$$\hat{y} = \text{mode}\ (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$$

It is evident that a Random Forest requires a higher level of complexity to predict the class of an observation as opposed to a Decision Tree. While this complexity is intended to enhance the prediction accuracy of a Random Forest, it is crucial to assess whether the incremental accuracy gain outweighs the trade-off in terms of a Decision Tree's intuitive comprehension and interpretability. Hence, we compare the Overall

Performance Score of both models to assess the gain in overall predictive accuracy from employing the more complex model. We observe from Table 3 that the accuracy gains from employing a Random Forest compared to a Decision Tree is less than 1%. Consequently, we opt to employ a Decision Tree for the nonparametric part of the paper to leverage the ease of visualization and interpretability, especially in the absence of a significant improvement in predictive accuracy from a Random Forest Classifier.

For this study, the Decision Tree is utilized in two distinct approaches. Firstly, we estimate a Decision Tree with a specific depth to uncover the key determining factors employed in the parametric model. Secondly, we compute the *Gini Impurity-Based Feature Importance* for all the chosen features to evaluate their ability to reduce the Gini Impurity when incorporated into the model thus providing a comprehensive view of their significance. Hence, this approach presents a case where researchers can leverage machine learning techniques to mine for key features from a pool of predictors which can be visualized by a simple Decision Tree. The effect of these predictors is then computed and analyzed using econometric techniques. Such integration between Machine Learning techniques and Econometric Analysis has recently attracted significant interest in both academic research and real-world applications as discussed by Yang et al. (2022).

Chapter 5: Results

5.1 Parametric Evidence: Generalized Ordered Logit Model with Partial Proportional Odds

In this section, we discuss the parametric findings from the Generalized Ordered Logit Model with Partial Proportional Odds. As shown earlier in section 4.2, some variables meet the proportional odds assumption while others do not. Therefore, we divide this section into two parts. The first analyzes the findings of the variables that satisfy the Proportional Odds Assumption, while the second analyzes the findings of the variables that do not. The results of the estimation are presented in Table 7 below.

Table 7: Results of the Ordered Logit Model and the Generalized Ordered Logit Model with Partial Proportional Odds (Odds Ratio)

Explanatory Variables	Ordered Logit Model (Proportional Odds)				Generalized Ordered Logit Model with Partial Proportional Odds								
	Same Odd Ratios			Proportional Odds Assumption	SC, VC, and EC versus NC			VC and EC versus NC and SC			EC versus NC, SC, and VC		
	Odds Ratio	Standard Error	Significance		Odds Ratio	Standard Error	Significance	Odds Ratio	Standard Error	Significance	Odds Ratio	Standard Error	Significance
Parental Attributes													
Female	(0.966)	0.041		Yes	(0.969)	0.041		(0.969)	0.041		(0.969)	0.041	
Old	(1.139)	0.039	***	No	(1.241)	0.054	***	(1.050)	0.045		(1.118)	0.074	
Holds a University Degree	(0.764)	0.022	***	No	(0.853)	0.030	***	(0.691)	0.025	***	(0.669)	0.036	***
Immigrant Parent	(1.170)	0.048	***	No	(1.009)	0.050		(1.315)	0.066	***	(1.395)	0.105	***
Belong to a Minority Group	(1.670)	0.071	***	No	(1.525)	0.082	***	(1.687)	0.085	***	(1.985)	0.146	***
Identify as Indigenous	(1.343)	0.099	***	Yes	(1.344)	0.099	***	(1.344)	0.099	***	(1.344)	0.099	***
Child Activity & Eating													
Child Spends Time Daily in Front of a Screen (Multiple Activities)	(1.116)	0.046	***	No	(1.161)	0.053	***	(1.038)	0.058		(0.910)	0.079	
Child Plays Video Games Daily	(1.110)	0.032	***	No	(1.019)	0.035		(1.220)	0.045	***	(1.262)	0.072	***
Concern for Child Eating Junk Food	(2.039)	0.055	***	No	(2.126)	0.065	***	(1.879)	0.069	***	(1.885)	0.114	***
Child Reads Daily	(0.830)	0.022	***	Yes	(0.827)	0.022	***	(0.827)	0.022	***	(0.827)	0.022	***
Labor Market Impacts													
Concern for Balancing Childcare & Work	(1.366)	0.071	***	No	(1.442)	0.082	***	(1.222)	0.085	***	(1.100)	0.118	
Lost Job/Experienced a drop in Work Hrs.	(1.050)	0.026	**	Yes	(1.049)	0.026	*	(1.049)	0.026	*	(1.049)	0.026	*
Being Less Patient with Their Children	(1.205)	0.046	***	No	(1.269)	0.054	***	(1.124)	0.057	**	(0.971)	0.075	
Social Impacts													
Concern for Connecting with Family and Friends	(2.105)	0.094	***	Yes	(2.086)	0.093	***	(2.086)	0.093	***	(2.086)	0.093	***
Childcare Impacts													
Will Use Childcare Services	(0.900)	0.024	***	Yes	(0.899)	0.024	***	(0.899)	0.024	***	(0.899)	0.024	***
Did Not Stop Using Childcare Services	(0.879)	0.040	***	Yes	(0.877)	0.040	***	(0.877)	0.040	***	(0.877)	0.040	***
Number of Observations	24,956				24,956								
Pseudo R2	0.0330				0.0357								
Log-likelihood	-28,623				-28,542								
Akaike Information Criterion (AIC)	57,285				57,158								
Bayesian Information Criterion (BIC)	57,439				57,458								

Notes: This table presents the Odds Ratios (in parentheses) from the Ordered Logit Model and the Generalized Ordered Logit Model. Significance levels are represented by asterisks as follows: *** p<.01, ** p<.05, * p<.1

5.1.1 Findings for Variables Complying with the Proportional Odds Assumption

We first start by analyzing the effect of variables that comply with the proportional odds assumption imposed by the Ordered Logit Model. Notice that when estimated using the Generalized Ordered Logit with Proportional Odds, variables complying with the Proportional Odds Assumption will have the same effect across the different categories of the dependent variable similar to that depicted by the standard Ordered Logit Model. From the findings of the Brant Test in Table 4 and the results of the models in Table 7, we see that seven of the chosen predictor variables were shown to respect the Proportional Odds Assumption and are discussed below.

- Parents' Gender: Complying with the proportional odds assumption, results showed that female parents were 3.4% less likely to be concerned about their children's physical health during COVID-19, although this effect was not statistically significant ($p > 0.05$). This finding contradicts Waters et al. (2000) and Van der Vegt & Kleinberg (2020), who predicted an effect of parents' gender.

- Indigenous Identity: Parents with an indigenous identity were 34.3% more likely to be concerned about their children's health during the pandemic, with a statistically significant effect ($p < 0.01$). This supports studies showing higher obesity and chronic disease rates among Indigenous youth in Canada compared to non-Indigenous youth (Kriska et al., 2001; Katzmarzyk 2008).

- Daily Reading: Having children read daily during COVID-19 emerged as one of the most influential factors in reducing parental concern about their children's physical health. Specifically, results showed that parents of children who read daily were 17% less likely to be concerned, with a statistically significant effect ($p < 0.01$). This aligns with research on the positive effects of reading on children's health (Roshanaei-Moghaddam et al., 2009; Dowrick et al., 2012; Mak and Fancourt, 2020).

- Job Loss/Reduced Work Hours: parents who lost their jobs or experienced reduced work hours were 5% more likely to be concerned about their children's health during COVID-19, with a statistically significant effect ($p < 0.05$). This relatively small effect, as depicted by an odds of 5%, might be explained by the discussion made by Cost et al. (2021) who argue that although parents losing their jobs or working fewer hours is likely to affect their ability to enroll their children in

physical-related activities, the emergency financial benefits provided in Canada during the pandemic may have lessened that effect.

- Concern for Social Connection: A key factor in our study was parents' concern about their children's ability to socialize with family and friends during COVID-19. This concern increased their worry about their children's physical health by over 100%, with a statistically significant effect ($p < 0.01$). Our findings are consistent with those reported by Ellis et al. (2020) and Szpunar et al. (2021) who showed that limited social interaction during COVID-19 increased parents' concern about their children's physical health. Furthermore, our study highlights that parental concern about social connection is the second most important factor influencing concern about children's physical health during COVID-19.

- Childcare Use: Both childcare-related variables - planning to use childcare and not stopping childcare use - were shown to reduce parental concern about children's physical health during COVID-19. Planning to use childcare reduced concern by 10%, while not stopping childcare use reduced it by 12%. Both effects were statistically significant ($p < 0.01$). This aligns with research by Carroll et al. (2020) who argued that childcare closures during COVID-19 significantly impacted access to childcare facilities, leading to fewer options for physical activity and increased sedentary behavior in children.

5.1.2 Findings for Variables Violating the Proportional Odds Assumption

Unlike variables discussed earlier, those violating the proportional odds assumption offer richer insights. The Generalized Ordered Logit model allows us to explore the heterogeneous effect of these variables across different concern levels for children's physical health during COVID-19.

- Parents' Age: the first variable violating the proportional odds assumption is the Parents' Age. Looking at the Generalized Ordered Logit Model, we observe all three odd ratios to be greater than one. This indicates that older parents are more likely to have higher concern about their children's physical health during COVID-19. However, closer examination shows us that only the first odds ratio is significant. Hence, our findings show that older parents are more likely to be *Somewhat Concerned, Very Concerned*, and *Extremely Concerned* as opposed to *Not at All Concerned*.

  Moving to the second and third odd ratios - measuring the odds of being *Very* or *Extremely Concerned* versus *Not at All* or *Somewhat Concerned* and being *Extremely Concerned* versus *Not at All, Somewhat*, or *Very Concerned* – we observe them to be insignificant. This shows that while being an old parent increases the level of concern about children's physical health, this feature is not likely to push the parents' concern to high levels thus making them *Very* or *Extremely Concerned*.

  Had we followed the findings of the Ordered Logit Model, we would have missed out on key insights about the dynamic effect of parents' age. Specifically, we would have concluded that being an old parent equally increases the chance of moving up the ladder of concern even reaching *Very Concerned* and *Extremely Concerned,* which proved to not be the case by the Generalized Ordered Logit Model. Hence, employing the Generalized Ordered Logit doesn't only solve the violation of the proportional odds assumption, but also helps us uncover deeper insights that have been masked by the Ordered Logit Model. Such differences between the findings of the two models can be observed also in the other variables that violate the proportional odds assumption.

  Our findings on the effect of parental age contribute to the literature showing an absence of consensus on the effect of parents' age on the physical health of children whereby authors like de Buhr, E., & Tannen, A. (2020) and Petersen et al. (2020) argue for a strong effect of parental age on children's physical activity while authors like Davids & Roman (2014) show that parental age does not significantly affect a child's physical health. In this regard, we show that both views may be relevant as our findings show that parental age may increase the level of concern

and is not likely to push the parents' concern to high categories thus making them very concerned.

- University Degree: all three odd ratios of the educational attainment variable are significant and less than 1. This shows that parents with a university degree are less likely to be concerned about their children's physical health during COVID-19. We also note that this effect differs a lot across the odds ratios as it increases from the first odds ratio to the second and from the second to the third. Hence, holding a university degree is likely to make parents less worried.

  Numerically, the first odds ratio is 0.853. This means that parents with a university degree are 14.7% less likely to be *somewhat*, *very,* or *extremely concerned* compared to being *not at all concerned.* The second odds of 0.691 indicate a higher effect than the first odds showing that parents with a university degree are 30.9% less likely to be *very* or *extremely concerned* compared to being *not at all or somewhat concerned*. Lastly, the third odds show an even higher effect with an odds ratio of 0.669. This shows parents with a university degree are 33.1% less likely to be *extremely concerned* compared to being *not at all, somewhat*, or *very concerned.*

  Hence, findings from our study align with those made by de Buhr & Tannen (2020), showing that parents with higher educational attainment tend to be less concerned about their children's physical health.

- Immigrant Status: our findings show that immigrant parents in Canada are more likely to be concerned about their children's physical health during COVID-19. This is evident in all three odds being above one although only the second and third odds were shown to be significant.

  Empirically, we observe from our second odds ratio that immigrant parents are 31.5% more likely to be *very* or *extremely concerned* compared to being *not at all or somewhat concerned*. Moreover, the third odds show that immigrant parents are 39.5% more likely to be *extremely concerned* compared to being *not at all, somewhat*, or *very concerned*. Hence, our Generalized Ordered Logit model does not only highlight how being an immigrant adversely affects parents' concerns but also shows that immigrant parents are more likely to be at higher levels of concern (i.e. *Very* or *Extremely Concerned*). Also, findings from the third odds of the

Generalized Ordered Logit show that the true effect of being an immigrant parent is twice as big as that depicted by the Ordered Logit model showing an increased odds of concern by only 17%.

- Visible Minority: as for the effect of parents belonging to a visible minority on their level of concern, findings of the Generalized Ordered Logit Model show a positive effect. Specifically, we observe all three odds to be significant and greater than 1 with a major increase in the effect across the odds.

  Empirically, the first odds ratio shows that parents belonging to a visible minority in Canada are 52.5% more likely to be *somewhat*, *very,* or *extremely concerned* compared to being *not at all concerned.* As for the second odds, the effect becomes higher with parents belonging to a minority being 68.7% more likely to be *not at all* or *somewhat concerned* compared to being *very,* or *extremely concerned*. Moreover, the effect increases even higher with the third odds showing that parents belonging to a minority are almost twice as likely to be *extremely concerned* compared to being *not at all, somewhat*, or *very concerned*.

  Thus, results from our model show that belonging to a visible minority strongly affects the chances of parents being at the high-end level of concern. Hence, findings from our study are in line with those in existing literature such as those of Mahmood et al. (2019) and Heidinger & Cotter (2020), showing that parents belonging to a minority or indigenous groups tend to be more concerned about their children's physical health.

- Daily Screen Time: regarding the effect of children spending time daily in front of a screen for various activities on the parent's concern about their physical health, our findings show a positive relation. Empirically, parents with children spending time daily in front of a screen are shown to be 16% more likely to be *somewhat*, *very,* or *extremely concerned,* as depicted by the first odds ratio.

  As for the second and third odds ratios, our findings show that both of them are insignificant. Hence while children spending time daily in front of a screen for various activities does increase parents' concern about their physical health, it is not the case that this will likely make parents *Very* or *Extremely Concerned*.

Compared to the literature, findings from our study are consistent with those of Moore et al. (2020) and Guerrero et al. (2020) showing that the increase in screen time during the pandemic contributes to heightened levels of concern for parents about their children's physical health. Moreover, while the literature presented argues broadly about the negative effect of screen time, our study differentiates between screen time spent studying or watching TV and that spent playing video games. As such, our next variable, *Daily Video Games*, analyzes the effect of children playing video games daily on their parents' perception of their physical health. Thus, our paper offers a broader perspective on the effect of screen time that is not shown by other authors in the literature.

- Daily Video Games: our findings on the effect of children playing video games on their parents' concerns is different than that of the children being in front of a screen daily for various activities. In this regard, we show that parents of children playing video games daily are likely to be very or extremely worried as depicted by the second and third odds being significant and greater than one. This was not the case for the effect of the previous variable which only had the first odds ratio significant and above one. Thus, our paper highlights the importance of what children are doing in front of the screen, something that the literature has overlooked.

  Empirically, we note that parents of children playing video games daily are 22% more likely to be *very* or *extremely concerned* compared to being *not at all* or *somewhat concerned.* Moreover, parents were shown to be 26.2% more likely to be *extremely concerned* compared to being *not at all*, *somewhat, or very concerned.* Similar to the effect of being an immigrant, generalized ordered logit shows again an effect twice as much as that depicted by the ordered logit model.

  One possible reason behind the differences between the two variables about screen time is that the first tracks whether children are in front of the screen daily for multiple tasks. Such tasks include studying remotely, which was the new normal during COVID-19 lockdowns. So, while this did increase parents' concern about their children's physical health, it didn't make them *very* and *extremely concerned* as this was part of the new learning routine. On the contrary, the second variable (i.e., playing video games daily) is more specific to children being in front of the

screen daily to play video games which was shown to push parents' concern to the higher categories of *very* and *extremely concerned*.

- Eating Habits: parents of children consuming junk food during COVID-19 are shown to be more likely concerned about their physical health as depicted by all three odds being significant and greater than one. For the first odds, we note that parents of children consuming junk food during COVID-19 are 126% more likely to be somewhat, *very,* or *extremely concerned.* Similarly, the second odds show that these parents are 87.9% more likely to be *very* or *extremely concerned* and 88.5% more likely to be *extremely concerned*. Hence our findings show that eating habits are one of the most influential factors increasing parents' concerns about their children's physical health during COVID-19.

  While results from our study agree with those of Maximova et al. (2022) and Burkart et al. (2022) showing that children's eating habits play a significant role in increasing parents' concern about their physical health, our paper strongly highlights this feature as a key one as depicted by it having the highest odds ratio of all predictor variables. Moreover, we show in the next section of the non-parametric findings that eating junk food was chosen by the Decision Tree to be the first splitting node. This points out to it as a key determinant among the pool of explanatory variables.

- Work-Life Balance: another factor showing a positive effect on parent's concern about their children's physical health during COVID-19 is being concerned about balancing between childcare and work. In this regard, our model shows that parents struggling to balance childcare and work-related tasks are 44.2% more likely to be *somewhat*, *very*, or *extremely concerned. Results also show that these parents were* 22.2% more likely to be *very* or *extremely concerned*. As for the last odds ratio measuring the odds of being *extremely concerned*, we see that it is positive but not significant.

  These findings indicate that while parents struggling to balance their work and taking care of their children are more likely to be concerned about their children's physical health, it is unlikely that this will make the parents *extremely concerned.* Compared to the literature such as the work of Carroll et al. (2020), our model does not only highlight the effect of parental stress as key but also attempts to

quantify this concern by computing the odds of this effect. On the contrary, the discussed literature often relies on thematic analysis which has several limitations. One major limitation is the subjectivity in choosing the themes by the researcher. This can hence lead to biased or misleading findings. Another source of bias generated by the researchers is when they only select the data that matches their desired findings from the study.

- Parental Patience with Children: the effect of parents becoming less patient during COVID-19 on their concern about their children's physical health shows a similar effect to that of balancing between work and childcare. In this regard, our findings show that parents who are becoming less patient with their children during COVID-19 are 26.9% more likely to be *somewhat*, *very*, or *extremely concerned.* Moreover, we also show that less patient parents are 12.4% more likely to be *very* or *extremely concerned*. As for the last odds ratio measuring the odds of being *extremely concerned*, we see that it is not significant. Hence, our findings are in line with those presented by Walton et al. (2014) and Stenhammar et al. (2010) arguing for an association between parental stress and children's physical health measures such as physical activity and body mass index (BMI).

5.2 Non-Parametric Evidence: Decision Tree and Feature Importance

5.2.1 Decision Tree

We first estimate a Decision Tree (presented in Figure 1) with a maximum depth of 3 that examines the key predictor variables in our model following a non-parametric way. From the Decision Tree figure, we see 5 different predictor variables originating from the five different predictor variable categories that were chosen according to the existing literature on the subject matter. Namely, the variables are:

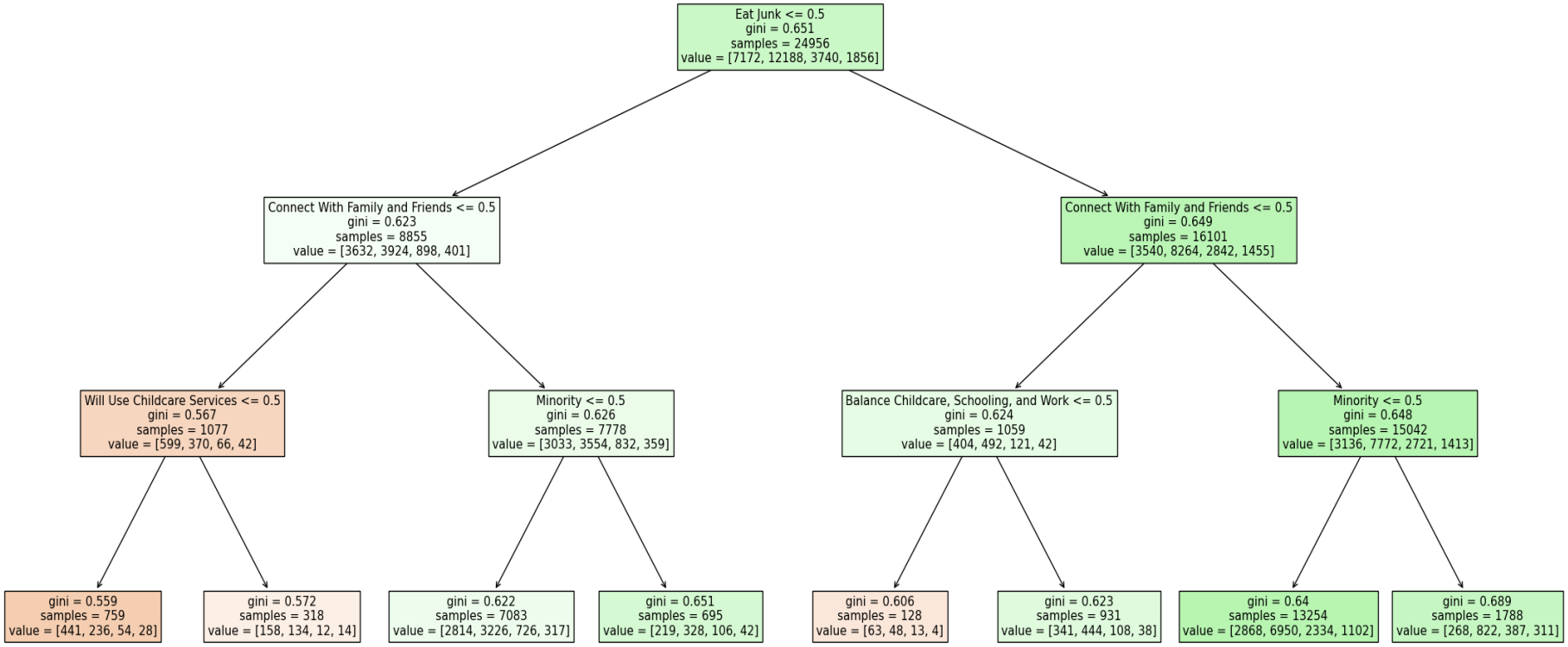
- Parents Concern About Children Eating Junk Food During COVID-19 (Originates from the Predicator Variables Category of "Child Activity and Eating Habits")
- Parents Concern for Children Remaining Connected with Friends and Family (Originates from the Predicator Variables Category of "Social Impacts")
- Parents Planning to Use Childcare Services After COVID-19 (Originates from the Predicator Variables Category of "Childcare Impacts")

- Parents Belonging to a Visible Minority (Originates from the Predicator Variables Category of "Parental Attributes")
- Parents Concern about Balancing Between Childcare and Work During Covid-19 (Originates from the Predicator Variables Category of "Labor Market Impacts")

Starting with the root node (i.e., Eat Junk), the labels $Eat\ Junk \leq 0.5$, $samples$, $Gini\ impurity$, and $Value$ refers to the following:

- $Eat\ Junk \leq 0.5$: refers to the splitting criteria with the left branch indicating a true value (i.e., $Eat\ Junk \leq 0.5$) and the right branch indicating a false value (i.e., $Eat\ Junk \geq 0.5$). Given that Eat Junk is a binary variable that tracks how concerned parents are about their children's consumption of junk food taking a value of 1 if parents are *Somewhat*, *Very*, or *Extremely Concerned* and 0 otherwise, then the left branch is for when the variable takes the value of zero (i.e., *Not at All Concerned*) and the right branch is when the variable takes the value of one (i.e. *Somewhat*, *Very*, or *Extremely Concerned*).
- $samples$: refers to the total number of observations in the leaf node. Given that this is the first leaf node, the number of observations in it will resemble the total number of observations of $24{,}956$; a number that we can see in the table of the distribution of the dependent variable (i.e., Table 1) and the regression results (i.e., Table 7).
- $gini$: refers to the *Gini Impurity* which is a measure of how impure the leaf node is. In other words, it calculates the probability that a randomly chosen observation is misclassified (i.e., assigned to the wrong category of the dependent variable). Applying the formula of the Gini impurity in Equation 15 for the first node (i.e. for the whole dataset before any splitting), we get a $Gini\ Impurity = 1 - \sum(28.74\%)^2 + (48.84\%)^2 + (14.99\%)^2 + (7.44\%)^2 = 0.65086131,$ hence the Gini impurity of 0.651 of the first node.
- $value$: refers to the distribution of observations in the leaf node among the different categories of the dependent variable. As such for the first leaf node with no splitting yet, we see a similar distribution to that of *Table 1: Distribution of the Dependent Variable.* Hence, $Value = [7172{,}12188{,}3740{,}1856]$.

Figure 1: Decision Tree with a Maximum Depth of Three



Notes: This figure presents a Decision Tree with a Maximum Depth of Three.

5.2.2 Feature Importance Based on Gini Impurity

For the second part of the non-parametric estimation, we check for key predictor variables following another technique called "Feature Importance Based on Gini Impurity" whereby results are presented in the bar graph in Figures 2 and 3. Unlike the Decision Tree which chooses the best predictor variable to split the data in a way that achieves the highest immediate drop in Gini Impurity at each node, the Feature Importance Based on Gini Impurity checks the features that contribute to lowering the Gini Impurity across the entire tree not only at a specific leaf node like in a Decision Tree. As such, the top five predictor features following the Gini Impurity are:

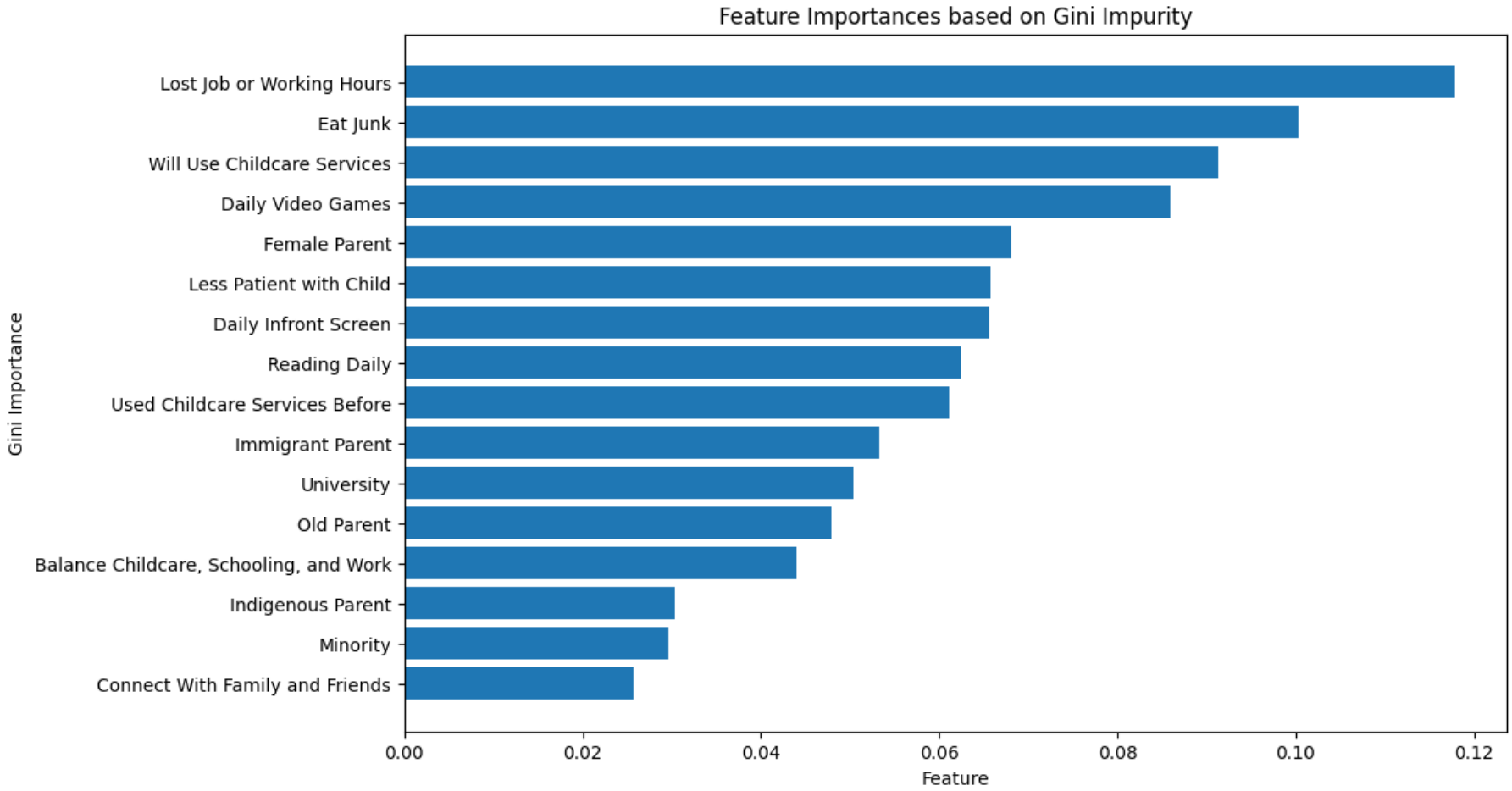
- Parents Losing Job or Having Experienced a Drop in Working Hours (Originates from the Predicator Variables Category of "Labor Market Impacts")
- Parents Concern About Children Eating Junk Food During COVID-19 (Originates from the Predicator Variables Category of "Child Activity and Eating Habits")
- Parents Planning to Use Childcare Services After COVID-19 (Originates from the Predicator Variables Category of "Childcare Impacts")
- Children Playing Video Games Daily During COVID-19 (Originates from the Predicator Variables Category of "Child Activity and Eating Habits")
- Female Parent (Originates from the Predicator Variables Category of "Parental Attributes")

In conclusion, we observe that following the *Decision Tree Model* and the *Feature Importance Based on Gini Impurity*, researchers can pinpoint the key predictor features from among a large pool of predictor variables with a non-parametric setup. While such techniques offer an easily visualized and interpretable output compared to the parametric approach, the non-parametric approach still falls behind in quantifying the effect of the chosen key predictor variables. Additionally, the non-parametric approach fails to predict the direction of the effect of the chosen key predictor variables on the response variable (i.e., does this key predictor variable increase or decrease parents' concern about their children's physical health during COVID-19?).

Hence, our approach presents a case where researchers can benefit from combining non-parametric machine learning techniques with parametric econometric modeling in several ways. First, it allows researchers to examine the relationship between the set of predictor variables and response variables in different ways whereby non-

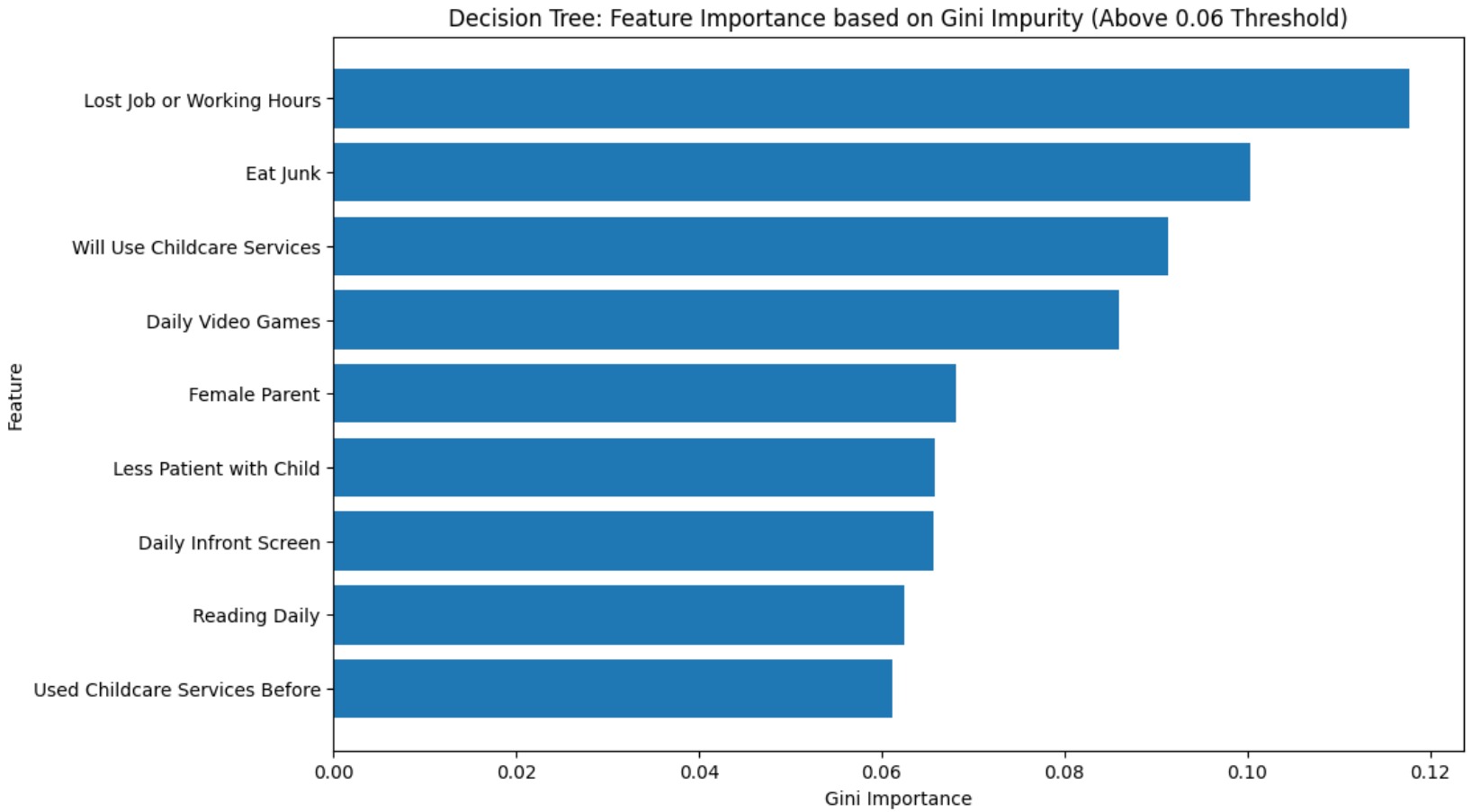
parametric works well for identifying possible effects without prior assumptions that require testing and validation (as we had in our Ordered Logit Model with the Proportional Odds Assumption). As for the parametric model, this will then be used to quantify such patterns from the exploratory analysis done in the non-parametric model along with testing for the established hypothesis. Another advantage of combining parametric and non-parametric approaches is that one can be the robustness check for the other, specifically the non-parametric can act as a robustness check for the parametric as it doesn't force any assumptions. For example, our parametric approach (i.e., the Generalized Ordered Logit Model) clearly shows us how factors like Eating Junk Food and Connecting with Friends and Family had the highest odds ratio among all predicator variables which was confirmed by our non-parametric model (i.e., the Decision Tree) which used the *Eat Junk* variable as the main splitting node and the *Connect with Friends and Family* as the second main splitting variable. Lastly, combining parametric and non-parametric approaches widens the audience of the paper as it targets both technical audience, who can delve deeper into the analysis of the odds ratio at different categories of the response variable, and non-technical audience, who benefit from the ease of visualization and interpretability of the non-parametric model.

Figure 2: Gini Impurity-Based Feature Importance from Decision Tree



Notes: this figure presents Gini Impurity-Based Feature Importance showing the effect of predictor variables in lowering the Gini Impurity of a Decision Tree.

Figure 3: Gini Impurity-Based Feature Importance from Decision Tree with a Threshold Above 0.06



Notes: this figure presents Gini Impurity-Based Feature Importance showing the effect of predictor variables in lowering the Gini Impurity of a Decision Tree by at least 0.06.

Chapter 6: Limitations

While our study sheds light on key factors influencing parental concern, it's important to acknowledge some limitations of the applied techniques.

Reverse Causality: One potential problem in our model is the possibility of having reverse causality, which is a source of endogeneity, which in turn limits the ability to derive causal inferences. For example, eating junk food and children's physical health may exhibit reverse causality as although eating junk food normally leads to bad physical health, one can also postulate that people with bad physical health tend to care less about their eating habits which may promote bad eating behaviors such as eating junk food. However, given that this study aims to find key predictors associated with parents' concern about their children's physical health, we are only interested in the predictive accuracy of the model instead of deriving causal relationships. In other words, our predictive analysis aims to forecast variations in the labels of the dependent variable and hence the focus is more on the accurate predictions of the model as opposed to the unbiased estimation of causal effects. Given that the cross-validation accuracy score of our Generalized Ordered Logistic Model is around 50%, the model demonstrates a relatively strong predictive power on out-of-sample data, especially considering the ordinal and categorical nature of the response variable with 4 labels.

Omitted Variable Bias: Another possible source of endogeneity is the Omitted Variable Bias (OVB) whereby a relevant predictor of the concern of parents about their children's physical health might have been omitted from the model. As such, the estimates of the model may have been biased because the effect of the omitted variable is wrongly attributed to the variables that are included in the model. To avoid OVB, one could include as many explanatory variables as possible in the model. However, this approach suffers the *Curse of Dimensionality* limitation whereby the more variables we add, the more the model will use these features to fit the noise in the training dataset leading to overfitting. This will in turn cause the model to perform very well on the training data but poorly on the out-of-sample data, hence lowering the predicative ability of the model. As such, we opt to add multiple explanatory variables while avoiding having too many of them so that we don't weaken the predictive power of our model.

Chapter 7: Conclusion

This paper aims to examine the factors predicting Canadian parents' concern about their children's physical health during COVID-19. Our research question is important and contributes to the literature in several ways. First, it recognizes the key role of parents in influencing their children's physical activity by examining the factors that affect parents' perceptions of their children's physical health. By analyzing these factors, we are better able to know what makes parents more concerned about their children's physical health during a pandemic which is important in case we face a similar pandemic in the future, especially in the context of a virus with high mutation ability.

Second, our research question is important because no paper has yet looked into the topic we are examining. There are only a small number of publications, namely those by McCormack et al. (2020), Ostermeier et al. (2022), and Szpunar et al. (2022), that fall within the scope of our work; each of these has several limitations that our study addresses. For instance, common limitations to all three studies limited data's sample size as well as being focused on a single geographic location. Such limitations are critical in research as they raise concerns as to whether the findings can be generalized to the bigger population which the sample was taken from and whether such findings can be generalized to other regions of the country.

For example, McCormack et al., (2020) aim at examining the relationship between parents' anxiety from COVID-19 and children's physical activity and sedentary practices. To do so, the authors rely on data from a survey covering 345 parents of children aged between 5 and 17 years old in Calgary, Alberta. Similarly, Ostermeier et al. (2022) rely on data from 27 parents of children enrolled in Grade 5 in London, Ontario, who were interviewed to study the effect of COVID-19 on their children's physical activity. By the same token, the paper of Szpunar et al. (2022) relies on data from 382 parents in Ontario, Canada to explain the perspectives of parents with children aged between 0 and 12 years old regarding their physical activity during the COVID-19 pandemic. Hence, none of the latter papers rely on data with a significant number of observations. This is key to ensure that the data is representative of the population it is taken from. Moreover, the observations were all concentrated in specific geographical locations which may render the results ungeneralizable for the

parents in other provinces. On the contrary, our paper utilizes a much bigger dataset with 24,956 observations spanned across the different Canadian provinces and territories. By using such data in our paper, we acknowledge the regional differences among Canadian parents. As such, insights derived from the study are likely to be more representative of Canadian parents in different provinces compared to the latter papers that focus on parents in specific provinces and cities.

The third contribution of our paper is the methodology used to answer the research question that comes at the intersection of econometric and machine learning techniques. As such, our paper employs a Generalized Ordered Logit Regression with Partial Proportional Odds, which is a parametric econometric technique along with a Decision Tree, which is a non-parametric machine learning technique. Hence, our approach presents a case where researchers can benefit from combining non-parametric machine learning techniques with parametric econometric modeling (Yang et al., 2022) to examine the relationship between the set of predictor variables and the response variables in different ways whereby non-parametric models work well for identifying possible patterns without prior assumptions that require testing and validation while parametric model can then be used to quantify and test the patterns observed from the exploratory analysis of the non-parametric models.

Findings from the Decision Tree show five key features affecting parents' concern about their children's physical health during COVID-19 after the first two splits, namely: (1) Parents' Concern About Their Children's Consumption of Junk Food, (2) Parents' Concern about Limitation for their Childrens to remain connected with Friends and Family, (3) Parents Concern for Balancing between Childcare and Work Tasks at Home, (4) Parents identifying with a visible Minority group in Canada, and lastly (5) Parents willing to make use of childcare services when they open after the pandemic.

As for findings from the Generalized Ordered Logistic Regression with Proportional Odds quantifying the effect of the latter variables shown by the first two splits of the Decision Tree, we note that Parents Concerned About Their Children Consuming Junk Food during COVID-19 and those Concerned for Their Children Remaining Connected with Family & Friends were both shown to be twice as likely to be more concerned about their children’s physical health. Moreover, findings show that Parents

Belonging to a Visible Minority and those Concerned About Balancing Between Childcare and Work tasks are 67% and 36.6% more concerned about their children's physical health, respectively. Lastly, parents willing to make use of childcare services when they open after the pandemic were shown to be 12% less concerned about their children's physical health.

Given the ongoing uncertainties surrounding potential future pandemics, understanding the factors that impacted parents' concern about their children's physical health during COVID-19 becomes even more critical. This knowledge equips policymakers to develop effective strategies that can lessen the adverse effects on parents and children should a similar situation arise. At the time of writing this paper, public concerns were mainly about two emerging Omicron subvariants and one bacterial infection. For the former, experts, such as Health Canada's Chief Medical Advisor Dr. Supriya Sharma, remain uncertain as to whether the current omicron variants, namely the *EG.5* and *BA.2.86*, will evolve in a way similar to COVID-19 ultimately causing a pandemic.

Regarding bacterial infection, several countries across the world are currently witnessing a surge in the cases of *Mycoplasma Pneumoniae* among children after the outbreak of the bacteria in northern China in May 2023. In this regard, the rapid increase in cases in countries like France led to the classification of this bacterial infection as an epidemic. In Quebec, experts such as Dr. Donald Vinh, who is a specialist in infectious diseases at McGill University Health Centre, argue that although there have not been many reported cases of the bacteria, we cannot ensure that we are safe from a possible outbreak. Dr. Vinh further notes that the low number of reported cases may be attributed to limited testing for the bacterial spread in Quebec.

Given the possibility of the emergence of a COVID-19-like scenario in the future, our research aims to guide policymakers to factors that influence parents' concern about their children's physical health during a pandemic. Such factors are key to directing policymakers toward certain policies aimed at lessening the adverse effects of a possible future pandemic. In this regard, our findings show that certain parental attributes, such as being old, immigrant, and belonging to a minority or indigenous group, are all factors contributing to higher concern about children's physical health

during COVID-19. Accordingly, policymakers are encouraged to develop targeted support programs for these groups of parents who are more adversely affected by the COVID-19 pandemic than their respective counterparts. These policies may include special financial aid programs and counseling services for these groups of parents.

Our study also shows that children spending a lot of time in front of a screen and eating junk food both adversely affect parents' concerns about their physical health during COVID-19. As such, policymakers are encouraged to hold campaigns to raise awareness among parents and children about the health risks of frequent screentime and junk food consumption. Moreover, policies that keep children active and entertained are key to encouraging children to reduce sedentary behavior, such as prolonged screen time playing video games, and improve their eating habits. These policies may include online home sports and art classes that keep children physically active, especially during lockdowns. Moreover, having stricter regulation on the advertisement of unhealthy food targeted to children can play a key role in shifting their consumption towards healthy alternatives.

Additionally, our paper shows that parents facing a challenge to balance childcare tasks and work and those who lost their jobs or experienced a drop in working hours are more concerned about their children's physical health during COVID-19. As such, policies to alleviate these effects may include support for flexible working hours so that parents can better coordinate between taking care of their children and work. Also, policymakers may want to offer workshops to parents about parenting during a pandemic and ways for better stress management during such trying times.

Lastly, our findings show that challenges for children to remain connected with family and friends during COVID-19 make parents more concerned about their physical health. Hence, it is recommended that policymakers take the necessary measures to ensure that children remain connected with their community during a pandemic. Examples of such measures to keep children connected could be encouraging virtual meetups. Furthermore, measures to keep children active and socializing during a pandemic include encouraging socially distanced activities with friends and family such as sports that are played at a distance like tennis, frisbee, kite flying, group bicycle rides, and fitness classes in the open air.

While the main observation may seem straightforward – that parents who exhibit greater overall concern for their children also tend to prioritize their physical health – the significance of the study lies in the nuanced exploration of the factors influencing parental perceptions. By examining the specific determinants shaping parental concerns about children's physical health during the COVID-19 pandemic, our research fills a fundamental gap in the existing literature on the subject matter. In this regard, we do not only identify key factors but also clarify their implications for public health policies and interventions. Thus, while the central finding may appear unsurprising, the depth of analysis and implications drawn from our study contribute valuable insights to the field.

Bibliography

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 1171-1178.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Breiman, L. (2017). *Classification and Regression Trees*. Routledge.

Burkart, S., Parker, H., Weaver, R. G., Beets, M. W., Jones, A., Adams, E. L., ... & Armstrong, B. (2022). Impact of the COVID-19 pandemic on elementary schoolers' physical activity, sleep, screen time and diet: a quasi-experimental interrupted time series study. *Pediatric Obesity*, *17*(1), e12846.

Canadian 24-hour Movement Guidelines: 24Hour Movement Guidelines. (2016, June). Retrieved from: http://csepguidelines.ca/

Carroll, N., Sadowski, A., Laila, A., Hruska, V., Nixon, M., Ma, D. W., ... & Guelph Family Health Study. (2020). The impact of COVID-19 on health behavior, stress, financial and food security among middle to high income Canadian families with young children. *Nutrients*, *12*(8), 2352.

Chaput, J. P., Colley, R. C., Aubert, S., Carson, V., Janssen, I., Roberts, K. C., & Tremblay, M. S. (2017). Proportion of preschool-aged children meeting the Canadian 24-Hour Movement Guidelines and associations with adiposity: results from the Canadian Health Measures Survey. *BMC Public Health*, *17*(5), 147-154.

Clark, W. (2008). Kids' sports. *Canadian Social Trends*, *85*, 54-61.

Como, R., Hambley, L., & Domene, J. (2021). An exploration of work-life wellness and remote work during and beyond COVID-19. *Canadian Journal of Career Development*, *20*(1), 46-56.

Cost, K. T., Crosbie, J., Anagnostou, E., Birken, C. S., Charach, A., Monga, S., ... & Korczak, D. J. (2021). Mostly worse, occasionally better: impact of COVID-19 pandemic on the mental health of Canadian children and adolescents. *European Child & Adolescent Psychiatry*, 1-14.

Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, 14(3), 391-403.

Davids, E. L., & Roman, N. V. (2014). A systematic review of the relationship between parenting styles and children's physical activity. *African Journal for Physical Health Education, Recreation and Dance*, *20*(sup-2), 228-246.

Dawson, T. (2020). As the COVID-19 pandemic hit, provinces declared states of emergency. Now many are up for renewal. *National Post*, *15*.

de Buhr, E., & Tannen, A. (2020). Parental health literacy and health knowledge, behaviours and outcomes in children: a cross-sectional survey. *BMC Public Health*, *20*(1), 1-9.

Dowrick, C., Billington, J., Robinson, J., Hamer, A., & Williams, C. (2012). Get into Reading as an intervention for common mental health problems: exploring catalysts for change. *Medical Humanities*, *38*(1), 15-20.

Ellis, W. E., Dumas, T. M., & Forbes, L. M. (2020). Physically isolated but socially connected: Psychological adjustment and stress among adolescents during the initial COVID-19 crisis. *Canadian Journal of Behavioural Science*, *52*(3), 177.

Guerrero, M. D., Vanderloo, L. M., Rhodes, R. E., Faulkner, G., Moore, S. A., & Tremblay, M. S. (2020). Canadian children's and youth's adherence to the 24-h movement guidelines during the COVID-19 pandemic: a decision tree analysis. *Journal of Sport and Health Science*, *9*(4), 313-321.

Hayes, G., Dowd, K. P., MacDonncha, C., & Donnelly, A. E. (2019). Tracking of physical activity and sedentary behavior from adolescence to young adulthood: a systematic literature review. *Journal of Adolescent Health*, *65*(4), 446-454.

Heidinger, L., & Cotter, A. (2020). Perceptions of personal safety among population groups designated as visible minorities in Canada during the COVID-19 pandemic.

Janssen, Ian, and Allana G. LeBlanc. "Systematic review of the health benefits of physical activity and fitness in school-aged children and youth." *International Journal of Behavioral Nutrition and Physical Activity* 7, no. 1 (2010): 1-16.

Katzmarzyk, P. T. (2008). Obesity and physical activity among Aboriginal Canadians. *Obesity*, *16*(1), 184-190.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*, *72*(6), 558-569.

Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*, 261-283.

Kriska, A. M., Hanley, A. J., Harris, S. B., & Zinman, B. (2001). Physical activity, physical fitness, and insulin and glucose concentrations in an isolated Native Canadian population experiencing rapid lifestyle change. *Diabetes Care*, *24*(10), 1787-1792.

Lacoste, Y., Dancause, K. N., Gosselin-Gagne, J., & Gadais, T. (2020). Physical activity among immigrant children: a systematic review. *Journal of Physical Activity and Health*, *17*(10), 1047-1058.

Lafave, L., Webster, A. D., & McConnell, C. (2021). Impact of COVID-19 on early childhood educator's perspectives and practices in nutrition and physical activity: A qualitative study. *Early Childhood Education Journal*, *49*(5), 935-945.

Lemieux, T., Milligan, K., Schirle, T., & Skuterud, M. (2020). Initial impacts of the COVID-19 pandemic on the Canadian labour market. *Canadian Public Policy*, *46*(S1), S55-S65.

Mahmood, B., Bhatti, J. A., Leon, A., & Gotay, C. (2019). Leisure time physical activity levels in immigrants by ethnicity and time since immigration to Canada: Findings from the 2011–2012 Canadian Community Health Survey. *Journal of Immigrant and Minority Health*, *21*, 801-810.

Mak, H. W., & Fancourt, D. (2020). Reading for pleasure in childhood and adolescent healthy behaviours: Longitudinal associations using the Millennium Cohort Study. *Preventive Medicine*, *130*, 105889.

Maximova, K., Khan, M. K., Dabravolskaj, J., Maunula, L., Ohinmaa, A., & Veugelers, P. J. (2022). Perceived changes in lifestyle behaviours and in mental health and wellbeing of elementary school children during the first COVID-19 lockdown in Canada. *Public Health*, *202*, 35-42.

McCormack, G. R., Doyle-Baker, P. K., Petersen, J. A., & Ghoneim, D. (2020). Parent anxiety and perceptions of their child's physical activity and sedentary behaviour during the COVID-19 pandemic in Canada. *Preventive Medicine Reports*, *20*, 101275.

Moore, L. L., Lombardi, D. A., White, M. J., Campbell, J. L., Oliveria, S. A., & Ellison, R. C. (1991). Influence of parents' physical activity levels on activity levels of young children. *Journal of Pediatrics*, *118*(2), 215-219.

Moore, S. A., Faulkner, G., Rhodes, R. E., Brussoni, M., Chulak-Bozzer, T., Ferguson, L. J., ... & Tremblay, M. S. (2020). Impact of the COVID-19 virus outbreak on movement and play behaviours of Canadian children and youth: a national survey. *International Journal of Behavioral Nutrition and Physical Activity*, *17*(1), 1-11.

Moore, S. A., Faulkner, G., Rhodes, R. E., Vanderloo, L. M., Ferguson, L. J., Guerrero, M. D., ... & Tremblay, M. S. (2021). Few Canadian children and youth were meeting the 24-hour movement behaviour guidelines 6-months into the COVID-19 pandemic: Follow-up from a national study. *Applied Physiology, Nutrition, and Metabolism*, *46*(10), 1225-1240.

Ostermeier, E., Tucker, P., Clark, A., Seabrook, J. A., & Gilliland, J. (2021). Parents' report of canadian elementary school children's physical activity and screen time during the COVID-19 pandemic: A longitudinal study. *International Journal of Environmental Research and Public Health*, *18*(23), 12352.

Ostermeier, E., Tucker, P., Tobin, D., Clark, A., & Gilliland, J. (2022). Parents' perceptions of their children's physical activity during the COVID-19 pandemic. *BMC Public Health*, *22*(1), 1459.

Petersen, T. L., Møller, L. B., Brønd, J. C., Jepsen, R., & Grøntved, A. (2020). Association between parent and child physical activity: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, *17*(1), 1-16.

Rhodes, R. E., Perdew, M., & Malli, S. (2020). Correlates of parental support of child and youth physical activity: a systematic review. *International Journal of Behavioral Medicine*, *27*, 636-646.

Roshanaei-Moghaddam, B., Katon, W. J., & Russo, J. (2009). The longitudinal effects of depression on physical activity. *General Hospital Psychiatry*, *31*(4), 306-315.

Sigmundová, D., Sigmund, E., Badura, P., & Hollein, T. (2020). Parent-child physical activity association in families with 4-to 16-year-old children. *International Journal of Environmental Research and Public Health*, *17*(11), 4015.

Simpson, R. F., Hesketh, K. R., Ellis, K., & van Sluijs, E. M. (2022). What research evidence exists about physical activity in parents? A systematic scoping review. *BMJ Open*, *12*(4), e054429.

Stenhammar, C., Olsson, G. M., Bahmanyar, S., Hulting, A. L., Wettergren, B., Edlund, B., & Montgomery, S. M. (2010). Family stress and BMI in young children. *Acta Paediatrica*, *99*(8), 1205-1212.

Szpunar, M., Saravanamuttoo, K., Vanderloo, L. M., Bruijns, B. A., Truelove, S., Burke, S. M., ... & Tucker, P. (2022). Children's Physical Activity during COVID-19 in Ontario, Canada: Parents' Perspectives. *International Journal of Environmental Research and Public Health*, *19*(22), 15061.

Szpunar, M., Vanderloo, L. M., Bruijns, B. A., Truelove, S., Burke, S. M., Gilliland, J., ... & Tucker, P. (2021). Children and parents' perspectives of the impact of the COVID-19 pandemic on Ontario children's physical activity, play, and sport behaviours. *BMC Public Health*, *21*(1), 1-17.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Trost, S. G., & Loprinzi, P. D. (2011). Parental influences on physical activity behavior in children and adolescents: a brief review. *American Journal of Lifestyle Medicine*, *5*(2), 171-181.

Van der Vegt, I., & Kleinberg, B. (2020). Women worry about family, men about the economy: Gender differences in emotional responses to COVID-19. In *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12* (pp. 397-409). Springer International Publishing.

Walton, K., Simpson, J. R., Darlington, G., & Haines, J. (2014). Parenting stress: a cross-sectional analysis of associations with childhood obesity, physical activity, and TV viewing. *BMC Pediatrics*, *14*, 1-7.

Waters, E., Doyle, J., Wolfe, R., Wright, M., Wake, M., & Salmon, L. (2000). Influence of parental gender and self-reported health and illness on parent-reported child health. *Pediatrics*, *106*(6), 1422-1428.

Williams, R. (2006). Generalized Ordered Logit/Partial Proportional Odds Models for Ordinal dependent variables. S*tata Journal*, *6*(1), 58-82.

Williams, R. (2016). Understanding and Interpreting Generalized Ordered Logit Models. *Journal of Mathematical Sociology*, *40*(1), 7-20.

Williamson, A. A., Mindell, J. A., Hiscock, H., & Quach, J. (2019). Sleep problem trajectories and cumulative socio-ecological risks: birth to school-age. *Journal of Pediatrics*, *215*, 229-237.

World Health Organization. (2020). COVID 19 Public Health Emergency of International Concern (PHEIC). Global research and innovation forum: towards a research roadmap.

Yang, M., McFowland III, E., Burtch, G., & Adomavicius, G. (2022). Achieving reliable causal inference with data-mined variables: A random forest approach to the measurement error problem. *INFORMS Journal on Data Science*, *1*(2), 138-155.

Zecevic, C. A., Tremblay, L., Lovsin, T., & Michel, L. (2010). Parental influence on young children's physical activity. *International Journal of Pediatrics*, *2010*.

Appendix

Diagnostic Check for Multicollinearity Using the Correlation Matrix

We first proceed with a preliminary check for multicollinearity by estimating the correlation between the set of explanatory variables which are presented in the Correlation Matrix of Table 8. While there is no set threshold to what is considered a benchmark for multicollinearity (Craney and Surles, 2002), some scholars such as Kim (2019) argue that a correlation coefficient of 0.8 or more between any two variables may indicate a severe multicollinearity problem. In our correlation matrix, we observe that all correlation coefficients are well below 0.8 which indicates a low risk of multicollinearity among our set of features.

Diagnostic Check for Multicollinearity Using the Variance Inflation Factors

After estimating the correlation matrix, we proceed into a more comprehensive testing for multicollinearity namely estimating the Variance Inflation Factors (VIFs) which is another diagnostic measure that shows how much the change in a variable is amplified by the presence of a related variable. The formula for the VIF estimation as discussed by Craney and Surles (2002) is shown in Equation 19 whereby VIF_i is the VIF for the i^{th} predictor variable for total predictor variables of p and r_i^2 is the R-squared value (often referred to also as the Coefficient of Determination) from regressing the i^{th} predictor variable on the remaining predictor variables. As such, a VIF value close to 1 indicates that the i^{th} predictor variable is not highly correlated with the remaining predictor variables (equivalent to having a near zero value for the R-squared meaning that the remaining predictor variable predicts almost nothing in the variation of the i^{th} predictor variable). The results of the VIF test are shown in Table 9.

Equation 18: Formula for Computing the Variance Inflation Factors (VIF)

$$\text{VIF}_i = \frac{1}{1 - r_i^2}$$

While there is no well-established consensus on the cutoff value for the VIF to indicate that we have multicollinearity, some scholars like Craney and Surles (2002) argue that a VIF value greater than 5 indicates a high risk of multicollinearity while others like Kim (2019) argue that a value greater than 10 is what indicates a high risk of multicollinearity. In light of the mean VIF being 1.09 in our model with the highest individual VIF being 1.24, this indicates a low risk of multicollinearity among our set of features.

Table 8: Correlation Matrix (Diagnostic Check for Multicollinearity)

Variables	Female parent	Old Parent	Parent Holds a University Degree	Immigrant Parent	Parents belong to a Minority	Parents Identify as Indigenous	Child Spends Time Daily in Front of a	Child Plays Video Games Daily	Concern for Child Eating Junk Food	Child Reads Daily	Concern for Balancing between	Parents' Lost Job or had a drop in Their	Concern About being Less Patient	Concern for Connecting with Family	Parent Will Use Childcare Services	Parents Did Not Stop Using
Female Parent	1.000															
Old Parent	-0.050	1.000														
Parent Holds a University Degree	-0.041	0.053	1.000													
Immigrant Parent	-0.043	0.045	0.101	1.000												
Parents belong to a Minority Group	-0.030	-0.015	0.094	0.424	1.000											
Parents Identify as Indigenous	0.024	-0.001	-0.100	-0.062	-0.060	1.000										
Child Spends Time Daily in Front of a Screen	0.019	0.078	-0.030	-0.022	-0.015	0.007	1.000									
Child Plays Video Games Daily	-0.019	-0.082	0.015	0.044	0.034	-0.010	-0.032	1.000								
Concern for Child Eating Junk Food	0.027	0.066	-0.067	0.009	0.040	0.018	0.205	-0.089	1.000							
Child Reads Daily	0.009	-0.186	0.182	-0.014	-0.021	-0.057	-0.125	0.157	-0.199	1.000						
Concern for Balancing between Childcare & Work	-0.020	-0.037	0.117	0.018	0.019	-0.021	0.035	-0.014	0.080	0.022	1.000					
Parents' Lost Job or had a drop in Their Work Hours	0.040	0.009	-0.174	0.018	-0.003	0.026	-0.020	0.013	0.024	-0.040	-0.032	1.000				
Concern About being Less Patient With Children	0.015	-0.104	0.031	-0.013	0.013	-0.009	0.054	-0.034	0.138	0.026	0.172	-0.009	1.000			
Concern for Connecting with Family and Friends	0.001	-0.015	0.025	-0.016	-0.020	-0.007	-0.005	-0.012	0.096	0.023	0.114	0.004	0.130	1.000		
Parent Will Use Childcare Services	-0.034	-0.166	0.088	0.015	0.004	-0.011	-0.047	0.034	-0.019	0.144	0.106	-0.052	0.113	0.031	1.000	
Parents Did Not Stop Using Childcare Services	0.026	-0.084	0.028	-0.020	-0.005	0.012	-0.037	0.005	-0.006	0.059	0.038	-0.042	0.031	0.011	0.224	1.000

Note: this presents the correlation matrix used as a preliminary diagnostic check for the severity of multicollinearity. As argued by Kim (2019), a correlation coefficient of 0.8 or more between any two variables may indicate a severe multicollinearity problem.

Table 9: Variance Inflation Factors (Diagnostic Check for Multicollinearity)

Variable	VIF	SQRT VIF	Tolerance	R-Squared
Female Parent	1.01	1.01	0.9884	0.0116
Old Parent	1.09	1.05	0.9136	0.0864
Parent Holds a University Degree	1.12	1.06	0.8937	0.1063
Immigrant Parent	1.24	1.11	0.8092	0.1908
Parents belong to a Minority Group	1.23	1.11	0.8116	0.1884
Parents Identify as Indigenous	1.02	1.01	0.9837	0.0163
Child Spends Time Daily in Front of a Screen (Multiple Activities)	1.06	1.03	0.9423	0.0577
Child Plays Video Games Daily	1.04	1.02	0.9639	0.0361
Concern for Child Eating Junk Food	1.12	1.06	0.8921	0.1079
Child Reads Daily	1.16	1.08	0.8647	0.1353
Parents' Concern for Balancing Childcare, Schooling & Work	1.07	1.03	0.9386	0.0614
Parents' Lost Job or Experienced a drop in Their Working Hours	1.04	1.02	0.9634	0.0366
Parents' Concerns About Being Less Patient with Their Children	1.08	1.04	0.9236	0.0764
Parents' Concern for Children Connecting with Family & Friends	1.03	1.02	0.9662	0.0338
Parent Will Use Childcare Services	1.12	1.06	0.8955	0.1045
Parents Did Not Stop Using Childcare Services	1.06	1.03	0.9437	0.0563
Mean VIF	1.09			

Notes: This table presents the output of the VIF. The first column is the individual values of the VIF for each of the predictor variables computed following Equation 19 discussed earlier. The second, third, and fourth columns are all built on the VIF values whereby the second column is the square root of the VIF, the third column is for a measure called Tolerance and computed as follows: $Tolerance = 1/VIF$ and lastly the fourth column is for the R-Squared which is computed as follows: $R^2 = 1 - Tolerance$

Alternative Modeling Technique: Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) is another machine learning classification technique that was first introduced by Tibshirani (1996). It is a regularization technique similar to Ridge Regression that extends the linear regression model by adding a penalty term to the loss function. This term may eliminate certain irrelevant features from the model by zeroing their coefficients so that the model becomes less sensitive to the noise and random fluctuations in the training data which in turn reduces the risk of overfitting. By accounting for overfitting, Lasso essentially performs feature selection which helps to prevent other problems such as multicollinearity which may arise from having similar features that can be highly correlated. Moreover, feature selection is relevant when the number of features is high that it approaches the number of observations which results in very high variance thus causing poor predictions from the model.

Overfitting refers to the case when the model performs well on the training data but poorly on the testing data. Mathematically, the Lasso regression aims at minimizing the following loss function presented in the following equation:

Equation 19: Lasso Regression Loss Function

$$L_{\text{lasso}}(\hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

Whereby:

- $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$: is the sum of squared residuals which is the squared difference between the actual value of the dependent variable for the i^{th} observation y_i and the predicted value of the dependent variable for the i^{th} observation $\hat{y}_i$ whereby n is the number of observations.
- $\lambda \sum_{j=1}^{m} |\hat{\beta}_j|$: is the penalty term – also known as L1 penalty – which is responsible for the regularization whereby λ is the penalty or regularization parameter that controls the strength of the regularization, m is the number of features or predictor variables in the model, and $\hat{\beta}_j$ is the estimated coefficient of the j^{th} feature.

To check for overfitting and see if Lasso is necessary, we perform cross-validation on the Generalized Ordered Logistic Regression Model to test for overfitting. This is done to compare the average accuracy between the training set and the testing. If the difference in accuracies between the two sets is high, then we conclude that the model is overfitting as it performs well on the training set but poorly on the testing set. By looking at the average accuracy scores of the training and testing datasets in Table 10, we observe that the difference between the two is minimal which indicates that the model is not overfitting the data.

Table 10: Average Training and Testing Accuracy Scores

Data set	**Average Accuracy Score**
Training	50.08%
Testing	49.89%

Note: this table presents the *Average Cross Validation Accuracy Score* of the training dataset and the testing dataset. The *Average Accuracy Score* refers to the mean accuracy of 10 different accuracy scores (i.e., $k = 10$).

Other performance metrics used to check for possible overfitting are the Mean Squared Error (MSE) and the Standard Deviation of the MSE (std MSE). The MSE is the average of the squared differences between the predicted values of the i^{th} observation of the dependent variable and the actual observed value of it. A low MSE indicates that the model's predictions are close to the actual values. Moreover, a low standard deviation of the MSE suggests that the model's performance is consistent across different subsets of the data. Hence, a low standard deviation of the MSE indicates low variability in the model's prediction which signifies that the model is not overfitting. Mathematically, the MSE and the standard deviation of the MSE (std MSE) are calculated as shown in Equations 21 and 22 and the results are presented in Table #11. Given that the MSE value is low, one can conclude that the model's predictions are close to the actual values suggesting that the model is performing well in terms of prediction accuracy. Moreover, the low std MSE suggests that the model's performance is consistent across the different subsets. Hence, with low MSE and std MSE, there is no clear evidence of overfitting as the model appears to be robust across the different subsets of the data.

Equation 20: Mean Squared Error Formula (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Whereby:

- y_i : is the actual value of the dependent variable for the i^{th} observation.
- $\hat{y}_i$: the predicted value of the dependent variable for the i^{th} observation.
- n : is the total number of subsets or folds

Equation 21: Mean Squared Error Standard Deviation (std MSE)

$$\text{std MSE} = \sqrt{\frac{\sum_{i=1}^{n} (MSE_i - \text{ mean MSE})^2}{n}}$$

Whereby:

- MSE_i : is the mean squared error for the i^{th} subset.
- mean MSE: is the mean of all *MSE* values across the subsets.
- $\hat{y}_i$: the predicted value of the dependent variable for the i^{th} observation.
- n : is the total number of subsets or folds.

Table 11: Evaluation Metrics on Testing Set: MSE and std MSE

Measure	Value
Mean Squared Error (MSE)	0.028
MSE Standard Deviation (std MSE)	0.015

Note: this table presents the Mean Squared Error (MSE) and the MSE Standard Deviation (std MSE) of the Testing dataset used as evaluation metrics to check if the model's predictions are close to the actual values and if the model's performance is consistent across the different subsets of data.